

Jian-Huang Lai · Cheng-Lin Liu
Xilin Chen · Jie Zhou · Tieniu Tan
Nanning Zheng · Hongbin Zha (Eds.)

LNCS 11258

Pattern Recognition and Computer Vision

First Chinese Conference, PRCV 2018
Guangzhou, China, November 23–26, 2018
Proceedings, Part III

3
Part III



 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, Lancaster, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Zurich, Switzerland

John C. Mitchell

Stanford University, Stanford, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

C. Pandu Rangan

Indian Institute of Technology Madras, Chennai, India

Bernhard Steffen

TU Dortmund University, Dortmund, Germany

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbrücken, Germany

More information about this series at <http://www.springer.com/series/7412>

Jian-Huang Lai · Cheng-Lin Liu
Xilin Chen · Jie Zhou · Tieniu Tan
Nanning Zheng · Hongbin Zha (Eds.)

Pattern Recognition and Computer Vision

First Chinese Conference, PRCV 2018
Guangzhou, China, November 23–26, 2018
Proceedings, Part III

Editors

Jian-Huang Lai
Sun Yat-sen University
Guangzhou, China

Cheng-Lin Liu
Institute of Automation
Chinese Academy of Sciences
Beijing, China

Xilin Chen
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China

Jie Zhou
Tsinghua University
Beijing, China

Tieniu Tan
Institute of Automation
Chinese Academy of Sciences
Beijing, China

Nanning Zheng
Xi'an Jiaotong University
Xi'an, China

Hongbin Zha
Peking University
Beijing, China

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-030-03337-8

ISBN 978-3-030-03338-5 (eBook)

<https://doi.org/10.1007/978-3-030-03338-5>

Library of Congress Control Number: 2018959435

LNCS Sublibrary: SL6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

© Springer Nature Switzerland AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Welcome to the proceedings of the First Chinese Conference on Pattern Recognition and Computer Vision (PRCV 2018) held in Guangzhou, China!

PRCV emerged from CCPR (Chinese Conference on Pattern Recognition) and CCCV (Chinese Conference on Computer Vision), which are both the most influential Chinese conferences on pattern recognition and computer vision, respectively. Pattern recognition and computer vision are closely inter-related and the two communities are largely overlapping. The goal of merging CCPR and CCCV into PRCV is to further boost the impact of the Chinese community in these two core areas of artificial intelligence and further improve the quality of academic communication. Accordingly, PRCV is co-sponsored by four major academic societies of China: the Chinese Association for Artificial Intelligence (CAAI), the China Computer Federation (CCF), the Chinese Association of Automation (CAA), and the China Society of Image and Graphics (CSIG).

PRCV aims at providing an interactive communication platform for researchers from academia and from industry. It promotes not only academic exchange, but also communication between academia and industry. In order to keep track of the frontier of academic trends and share the latest research achievements, innovative ideas, and scientific methods in the fields of pattern recognition and computer vision, international and local leading experts and professors are invited to deliver keynote speeches, introducing the latest advances in theories and methods in the fields of pattern recognition and computer vision.

PRCV 2018 was hosted by Sun Yat-sen University. We received 397 full submissions. Each submission was reviewed by at least two reviewers selected from the Program Committee and other qualified researchers. Based on the reviewers' reports, 178 papers were finally accepted for presentation at the conference, including 24 oral and 154 posters. The acceptance rate is 45%. The proceedings of the PRCV 2018 are published by Springer.

We are grateful to the keynote speakers, Prof. David Forsyth from University of Illinois at Urbana-Champaign, Dr. Zhengyou Zhang from Tencent, Prof. Tamara Berg from University of North Carolina Chapel Hill, and Prof. Michael S. Brown from York University.

We give sincere thanks to the authors of all submitted papers, the Program Committee members and the reviewers, and the Organizing Committee. Without their contributions, this conference would not be a success. Special thanks also go to all of the sponsors and the organizers of the special forums; their support made the conference a success. We are also grateful to Springer for publishing the proceedings and especially to Ms. Celine (Lanlan) Chang of Springer Asia for her efforts in coordinating the publication.

We hope you find the proceedings enjoyable and fruitful reading.

September 2018

Tieniu Tan
Nanning Zheng
Hongbin Zha
Jian-Huang Lai
Cheng-Lin Liu
Xilin Chen
Jie Zhou

Organization

Steering Chairs

Tieniu Tan	Institute of Automation, Chinese Academy of Sciences, China
Hongbin Zha	Peking University, China
Jie Zhou	Tsinghua University, China
Xilin Chen	Institute of Computing Technology, Chinese Academy of Sciences, China
Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China
Long Quan	Hong Kong University of Science and Technology, SAR China
Yong Rui	Lenovo Group

General Chairs

Tieniu Tan	Institute of Automation, Chinese Academy of Sciences, China
Nanning Zheng	Xi'an Jiaotong University, China
Hongbin Zha	Peking University, China

Program Chairs

Jian-Huang Lai	Sun Yat-sen University, China
Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China
Xilin Chen	Institute of Computing Technology, Chinese Academy of Sciences, China
Jie Zhou	Tsinghua University, China

Organizing Chairs

Liang Wang	Institute of Automation, Chinese Academy of Sciences, China
Wei-Shi Zheng	Sun Yat-sen University, China

Publicity Chairs

Huimin Ma	Tsinghua University, China
Jian Yu	Beijing Jiaotong University, China
Xin Geng	Southeast University, China

International Liaison Chairs

Jingyi Yu	ShanghaiTech University, China
Pong C. Yuen	Hong Kong Baptist University, SAR China

Publication Chairs

Zhouchen Lin Peking University, China
Zhenhua Guo Tsinghua University, China

Tutorial Chairs

Huchuan Lu Dalian University of Technology, China
Zhaoxiang Zhang Institute of Automation, Chinese Academy of Sciences, China

Workshop Chairs

Yao Zhao Beijing Jiaotong University, China
Yanning Zhang Northwestern Polytechnical University, China

Sponsorship Chairs

Tao Wang iQIYI Company, China
Jinfeng Yang Civil Aviation University of China, China
Liang Lin Sun Yat-sen University, China

Demo Chairs

Yunhong Wang Beihang University, China
Junyong Zhu Sun Yat-sen University, China

Competition Chairs

Xiaohua Xie Sun Yat-sen University, China
Jiwen Lu Tsinghua University, China

Website Chairs

Ming-Ming Cheng Nankai University, China
Changdong Wang Sun Yat-sen University, China

Finance Chairs

Huicheng Zheng Sun Yat-sen University, China
Ruiping Wang Institute of Computing Technology, Chinese Academy
of Sciences, China

Program Committee

Haizhou Ai Tsinghua University, China
Xiang Bai Huazhong University of Science and Technology, China

Xiaochun Cao	Institute of Information Engineering, Chinese Academy of Sciences, China
Hong Chang	Institute of Computing Technology, China
Songcan Chen	Chinese Academy of Sciences, China
Xilin Chen	Institute of Computing Technology, China
Hong Cheng	University of Electronic Science and Technology of China, China
Jian Cheng	Chinese Academy of Sciences, China
Ming-Ming Cheng	Nankai University, China
Yang Cong	Chinese Academy of Science, China
Dao-Qing Dai	Sun Yat-sen University, China
Junyu Dong	Ocean University of China, China
Yuchun Fang	Shanghai University, China
Jianjiang Feng	Tsinghua University, China
Shenghua Gao	ShanghaiTech University, China
Xinbo Gao	Xidian University, China
Xin Geng	Southeast University, China
Ping Guo	Beijing Normal University, China
Zhenhua Guo	Tsinghua University, China
Huiguang He	Institute of Automation, Chinese Academy of Sciences, China
Ran He	National Laboratory of Pattern Recognition, China
Richang Hong	Hefei University of Technology, China
Baogang Hu	Institute of Automation, Chinese Academy of Sciences, China
Hua Huang	Beijing Institute of Technology, China
Kaizhu Huang	Xi'an Jiaotong-Liverpool University, China
Rongrong Ji	Xiamen University, China
Wei Jia	Hefei University of Technology, China
Yunde Jia	Beijing Institute of Technology, China
Feng Jiang	Harbin Institute of Technology, China
Zhiguo Jiang	Beihguo University, China
Lianwen Jin	South China University of Technology, China
Xiao-Yuan Jing	Wuhan University, China
Xiangwei Kong	Dalian University of Technology, China
Jian-Huang Lai	Sun Yat-sen University, China
Hua Li	Institute of Computing Technology, Chinese Academy of Sciences, China
Peihua Li	Dalian University of Technology, China
Shutao Li	Hunan University, China
Wu-Jun Li	Nanjing University, China
Xiu Li	Tsinghua University, China
Xuelong Li	Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, China
Yongjie Li	University of Electronic Science and Technology of China, China
Ronghua Liang	Zhejiang University of Technology, China
Zhouchen Lin	Peking University, China

Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China
Huafeng Liu	Zhejiang University, China
Huaping Liu	Tsinghua University, China
Qingshan Liu	Nanjing University of Information Science and Technology, China
Wenyin Liu	Guangdong University of Technology, China
Wenyu Liu	Huazhong University of Science and Technology, China
Yiguang Liu	Sichuan University, China
Yue Liu	Beijing Institute of Technology, China
Guoliang Lu	Shandong University, China
Jiwen Lu	Tsinghua University, China
Yue Lu	East China Normal University, China
Bin Luo	Anhui University, China
Ke Lv	Chinese Academy of Sciences, China
Huimin Ma	Tsinghua University, China
Zhanyu Ma	Beijing University of Posts and Telecommunications, China
Deyu Meng	Xi'an Jiaotong University, China
Qiguang Miao	Xidian University, China
Zhenjiang Miao	Beijing Jiaotong University, China
Weidong Min	Nanchang University, China
Bingbing Ni	Shanghai Jiaotong University, China
Gang Pan	Zhejiang University, China
Yuxin Peng	Peking University, China
Jun Sang	Chongqing University, China
Nong Sang	Huazhong University of Science and Technology, China
Shiguang Shan	Institute of Computing Technology, Chinese Academy of Sciences, China
Linlin Shen	Shenzhen University, China
Wei Shen	Shanghai University, China
Guangming Shi	Xidian University, China
Fei Su	Beijing University of Posts and Telecommunications, China
Jian Sun	Xi'an Jiaotong University, China
Jun Sun	Fujitsu R&D Center Co., Ltd., China
Zhengxing Sun	Nanjing University, China
Xiaoyang Tan	Nanjing University of Aeronautics and Astronautics, China
Jinhui Tang	Nanjing University of Science and Technology, China
Jin Tang	Anhui University, China
Yandong Tang	Shenyang Institute of Automation, Chinese Academy of Sciences, China
Chang-Dong Wang	Sun Yat-sen University, China
Liang Wang	National Laboratory of Pattern Recognition, China
Ruiping Wang	Institute of Computing Technology, Chinese Academy of Sciences, China
Shengjin Wang	Tsinghua University, China
Shuhui Wang	Institute of Computing Technology, Chinese Academy of Sciences, China

Tao Wang	iQIYI Company, China
Yuanquan Wang	Hebei University of Technology, China
Zengfu Wang	University of Science and Technology of China, China
Shikui Wei	Beijing Jiaotong University, China
Wei Wei	Northwestern Polytechnical University, China
Jianxin Wu	Nanjing University, China
Yihong Wu	Institute of Automation, Chinese Academy of Sciences, China
Gui-Song Xia	Wuhan University, China
Shiming Xiang	Institute of Automation, Chinese Academy of Sciences, China
Xiaohua Xie	Sun Yat-sen University, China
Yong Xu	South China University of Technology, China
Zenglin Xu	University of Electronic and Technology of China, China
Jianru Xue	Xi'an Jiaotong University, China
Xiangyang Xue	Fudan University, China
Gongping Yang	Shandong University, China
Jie Yang	Shanghai JiaoTong University, China
Jinfeng Yang	Civil Aviation University of China, China
Jufeng Yang	Nankai University, China
Qixiang Ye	Chinese Academy of Sciences, China
Xinge You	Huazhong University of Science and Technology, China
Jian Yin	Sun Yat-sen University, China
Xu-Cheng Yin	University of Science and Technology Beijing, China
Xianghua Ying	Peking University, China
Jian Yu	Beijing Jiaotong University, China
Shiqi Yu	Shenzhen University, China
Bo Yuan	Tsinghua University, China
Pong C. Yuen	Hong Kong Baptist University, SAR China
Zheng-Jun Zha	University of Science and Technology of China, China
Daoqiang Zhang	Nanjing University of Aeronautics and Astronautics, China
Guofeng Zhang	Zhejiang University, China
Junping Zhang	Fudan University, China
Min-Ling Zhang	Southeast University, China
Wei Zhang	Shandong University, China
Yanning Zhang	Northwestern Polytechnical University, China
Zhaoxiang Zhang	Institute of Automation, Chinese Academy of Sciences, China
Qijun Zhao	Sichuan University, China
Huicheng Zheng	Sun Yat-sen University, China
Wei-Shi Zheng	Sun Yat-sen University, China
Wenming Zheng	Southeast University, China
Jie Zhou	Tsinghua University, China
Wangmeng Zuo	Harbin Institute of Technology, China

Contents – Part III

Document Analysis

Chinese Painting Rendering by Adaptive Style Transfer.	3
<i>Wanxin Zou, Xutao Li, and Sengping Li</i>	
The Accurate Guidance for Image Caption Generation.	15
<i>Xinyuan Qi, Zhiguo Cao, Yang Xiao, Jian Wang, and Chao Zhang</i>	
Large-Scale Visible Watermark Detection and Removal with Deep Convolutional Networks.	27
<i>Danni Cheng, Xiang Li, Wei-Hong Li, Chan Lu, Fake Li, Hua Zhao, and Wei-Shi Zheng</i>	
Learning to Generate Realistic Scene Chinese Character Images by Multitask Coupled GAN	41
<i>Qingxiang Lin, Lingyu Liang, Yaoxiong Huang, and Lianwen Jin</i>	
A Recognition Method of the Similarity Character for Uchen Script Tibetan Historical Document Based on DNN	52
<i>Xiaojuan Wang, Weilan Wang, Zhenjiang Li, Yiqun Wang, Yuehui Han, and Zhanjun Hao</i>	
Research on the Method of Tibetan Recognition Based on Component Location Information	63
<i>Yuehui Han, Weilan Wang, Yiqun Wang, and Xiaojuan Wang</i>	
Research on Text Line Segmentation of Historical Tibetan Documents Based on the Connected Component Analysis	74
<i>Yiqun Wang, Weilan Wang, Zhenjiang Li, Yuehui Han, and Xiaojuan Wang</i>	
Online Handwriting Tibetan Character Recognition Based on Two-Dimensional Discriminant Locality Alignment.	88
<i>Zhengqi Cai and Weilan Wang</i>	
Complex Printed Uyghur Document Image Retrieval Based on Modified SURF Features	99
<i>Aliya Batur, Patigul Mamat, Wenjie Zhou, Yali Zhu, and Kurban Ubul</i>	
Deep Word Association: A Flexible Chinese Word Association Method with Iterative Attention Mechanism	112
<i>Yaoxiong Huang, Zecheng Xie, Manfei Liu, Shuaitao Zhang, and Lianwen Jin</i>	

Face Recognition and Analysis

Face Recognition Based on Multi-view: Ensemble Learning	127
<i>Wenhui Shi and Mingyan Jiang</i>	
Conditional Face Synthesis for Data Augmentation	137
<i>Rui Huang, Xiaohua Xie, Jianhuang Lai, and Zhanxiang Feng</i>	
Face Image Illumination Processing Based on GAN with Dual Triplet Loss	150
<i>Wei Ma, Xiaohua Xie, Jianhuang Lai, and Junyong Zhu</i>	
Face Detection and Encryption for Privacy Preserving in Surveillance Video	162
<i>Suolan Liu, Lizhi Kong, and Hongyuan Wang</i>	
Content-Aware Face Blending by Label Propagation	173
<i>Lingyu Liang and Xinglin Zhang</i>	
Facial Expression Recognition Based on Region-Wise Attention and Geometry Difference	183
<i>Heran Du, Huicheng Zheng, and Mingjing Yu</i>	
Score-Guided Face Alignment Network Under Occlusions	195
<i>Xiang Yan, Huabin Wang, Qi Wang, Jinjie Song, and Liang Tao</i>	
Robust Face Detector with Fully Convolutional Networks	207
<i>Yingcheng Su, Xiaopei Wan, and Zhenhua Guo</i>	
Nuclear Norm Based Superposed Collaborative Representation Classifier for Robust Face Recognition	219
<i>Yongbo Wu and Haifeng Hu</i>	
Face Image Set Recognition Based on Bilinear Regression	233
<i>Wen-Wen Hua and Chuan-Xian Ren</i>	
Semi-supervised Learning of Deep Difference Features for Facial Expression Recognition	245
<i>Can Xu, Ruyi Xu, Jingying Chen, and Leyuan Liu</i>	

Feature Extraction and Selection

Noise Level Estimation for Overcomplete Dictionary Learning Based on Tight Asymptotic Bounds	257
<i>Rui Chen and Changshui Yang</i>	
Perceptual Compressive Sensing	268
<i>Jiang Du, Xuemei Xie, Chenye Wang, and Guangming Shi</i>	

Differential and Integral Invariants Under Möbius Transformation	280
<i>He Zhang, Hanlin Mo, You Hao, Qi Li, and Hua Li</i>	
Automatic Classifier Selection Based on Classification Complexity	292
<i>Liping Deng, Wen-Sheng Chen, and Binbin Pan</i>	
Gradient-Based Representational Similarity Analysis with Searchlight for Analyzing fMRI Data	304
<i>Xiaoliang Sheng, Muhammad Yousefnezhad, Tonglin Xu, Ning Yuan, and Daoqiang Zhang</i>	
Feature Aggregation Tree: Capture Temporal Motion Information for Action Recognition in Videos	316
<i>Bing Zhu</i>	
Adaptive Ensemble Probabilistic Matrix Approximation for Recommendation	328
<i>Xingxing Li, Liping Jing, and Huafeng Liu</i>	
A Deep Structure-Enforced Nonnegative Matrix Factorization for Data Representation	340
<i>Yijia Zhou and Lijun Xu</i>	
An Embedded Method for Feature Selection Using Kernel Parameter Descent Support Vector Machine	351
<i>Haiqing Zhu, Ning Bi, Jun Tan, and Dongjie Fan</i>	
Multimodal Joint Representation for User Interest Analysis on Content Curation Social Networks.	363
<i>Lifang Wu, Dai Zhang, Meng Jian, Bowen Yang, and Haiying Liu</i>	
LTSG: Latent Topical Skip-Gram for Mutually Improving Topic Model and Vector Representations	375
<i>Jarvan Law, Hankz Hankui Zhuo, JunHua He, and Erhu Rong</i>	
Improve the Spoofing Resistance of Multimodal Verification with Representation-Based Measures.	388
<i>Zengxi Huang, Zhen-Hua Feng, Josef Kittler, and Yiguang Liu</i>	
Machine Learning	
Function-Guided Energy-Precision Optimization with Precision-Rate-Complexity Bivariate Models	403
<i>Hao Liu, Rong Huang, and Zhihai He</i>	
Point Cloud Noise and Outlier Removal with Locally Adaptive Scale	415
<i>Zhenxing Mi and Wenbing Tao</i>	

Robust Multi-view Subspace Learning Through Structured Low-Rank Matrix Recovery 427
Jiamiao Xu, Xinge You, Qi Zheng, Fangzhao Wang, and Peng Zhang

An Online Learning Approach for Robust Motion Tracking in Liver Ultrasound Sequence 440
Chunxu Shen, Huabei Shi, Tao Sun, Yibin Huang, and Jian Wu

Set-to-Set Distance Metric Learning on SPD Manifolds 452
Zhi Gao, Yuwei Wu, and Yunde Jia

Structure Fusion and Propagation for Zero-Shot Learning 465
Guangfeng Lin, Yajun Chen, and Fan Zhao

A Hierarchical Cluster Validity Based Visual Tree Learning for Hierarchical Classification 478
Yu Zheng, Jianping Fan, Ji Zhang, and Xinbo Gao

Robust Shapelets Learning: Transform-Invariant Prototypes 491
Huiqi Deng, Weifu Chen, Andy J. Ma, Qi Shen, Pong C. Yuen, and Guocan Feng

A Co-training Approach for Multi-view Density Peak Clustering 503
Yu Ling, Jinrong He, Silin Ren, Heng Pan, and Guoliang He

Boosting Sparsity-Induced Autoencoder: A Novel Sparse Feature Ensemble Learning for Image Classification 514
Rui Shi, Jian Ji, Chunhui Zhang, and Qiguang Miao

Matrix-Instance-Based One-Pass AUC Optimization 527
Changming Zhu, Chengjiu Mei, Hui Jiang, and Rigui Zhou

Piecewise Harmonic Image Restoration with High Order Variational Model 539
Bibo Lu, Zhenzhen Huangfu, and Rui Huang

Dynamic Delay Based Cyclic Gradient Update Method for Distributed Training 550
Wenhui Hu, Peng Wang, Qigang Wang, Zhengdong Zhou, Hui Xiang, Mei Li, and Zhongchao Shi

Semi-supervised Dictionary Active Learning for Pattern Classification 560
Qin Zhong, Meng Yang, and Tiancheng Zhang





Multi-feature Shared and Specific Representation for Pattern Classification. . . 573
Kangyin Ke and Meng Yang

Distillation of Random Projection Filter Bank for Time Series Classification	586
<i>Yufei Lin, Sen Li, and Qianli Ma</i>	
Jointly Sparse Reconstructed Regression Learning	597
<i>Dongmei Mo, Zhihui Lai, and Heng Kong</i>	
Multi-scale Attributed Graph Kernel for Image Categorization	610
<i>Duo Hu, Qin Xu, Jin Tang, and Bin Luo</i>	
Author Index	623

Document Analysis



Chinese Painting Rendering by Adaptive Style Transfer

Wanxin Zou¹ , Xutao Li¹  , and Sengping Li² 

¹ Department of Electronic Engineering, Shantou University, Shantou 515063, China
{wxzou, lixt}@stu.edu.cn

² Department of Mechanical and Electronic Engineering, Shantou University,
Shantou 515063, China
spli@stu.edu.cn

Abstract. Chinese painting is distinct from other art in that the painting elements are exhibited by complex water-and-ink diffusion and shows gray, white and black visual effect. Rendering such a water-and-ink painting with polychrome style is a challenging problem. In this paper, we propose a novel style transfer method for Chinese painting. We firstly decompose the Chinese painting with adaptive patches based on its structure, and locally colorize the painting. Then, the colorized image is used for guiding the process of texture transfer that is modeled in Markov Random Field (MRF). More precisely, we improve the classic texture transfer algorithm by modifying the compatibility functions for searching the optimal matching, with the chromatism information. The experiment results show that proposed adaptive patches can well preserve the original content while match the example style. Moreover, we present the transfer results with our method and recent style transfer algorithms, in order to make a comparison.

Keywords: Chinese painting rendering · Style transfer
Adaptive patch-based texture transfer · Markov Random Field

1 Introduction

As a traditional art in China, Chinese painting differs from other art in its expressive brush strokes and ink diffusion. To ideally render water-and-ink painting, many researchers attempted to use computer simulation for such complicated texture generation [13, 15]. In this paper, we aim to render Chinese painting with other artistic style, which is regarded as a style transfer problem.

Style transfer is to synthesize an image that combines the structure of a original image with the artistic style of the example image. In this work, it

Supported by NSFC No. 61471229/61573233. and Department of Education of Guangdong Province (2015KCXTD018/2017KCXTD015).

is a process of migrating a style from an example image to Chinese painting, which can be generally regarded as transferring two different painting style. In animation production and video post-production fields, style transfer and related approaches are highly interested as they facilitate generating different scenes [9, 12]. Although various methods have been proposed for this issue, style transfer task has not been well-defined. The core difficulty is how to distinguish style feature from semantic content in an image, including all visual attributes such as texture, strokes, color and shading.

Previous study offers two distinct methods for style transfer: One is generalization of classic texture synthesis approaches, such as the works in [2, 3], in which optimal patches of a single image are expected to be found based on local similarity. An alternative technique for style transfer problem emerged in recent years, defining content and style representation of two images and using Convolutional Neural Networks (CNN) to merge the corresponding content and style [7].

Our work is motivated by patch-based texture synthesis approaches. In spite of traditional patch-based texture synthesis methods made an impressive success for style transfer, the limitations should be overcome. For example, the local texture synthesis is accomplished in the same and fixed size patches throughout the whole image, where the size of the patch is a tradeoff between the style and the content to be preserved in the output image. The size of patch should be large enough to exhibit the patterns that characterize the example style, yet small enough to reconstruct the realistic content of original image. Another limitation is that traditional constrains in transferring consider only luminance and local neighboring similarity of target image, without color information. Hence, we propose a style transfer method for Chinese painting which is able to overcome the limitations. The main contributions of this work are summarized as follows:

- We adaptively divide target Chinese painting into patches based on its local similarity for texture synthesis, instead of using patches of constant size, so as to achieve a realistic reconstruction of the original image while present most noticeable example style;
- Constraints are modified in the process of texture synthesis, where color is considered as a relevant factor guiding local texture transfer. It may guarantee the validity in transfer process, where the futile texture is prevented.

2 Related Works on Style Transfer

Style transfer can be considered as a special case of texture synthesis, where the content image influences the regular synthesis process. In the literatures of traditional texture synthesis and transfer, example-based methods are to generate a texture image by computing non-parametric sampling from a given example style image based on Markov Random Field (MRF). One of the earliest works in [2] by Efros and Leung takes a pixel to be synthesized by random sampling from a set of candidate pixels that are selected from an example texture image. This process is repeated for every output pixel by growing from the initial region until

all the output pixels have been already synthesized. Intuitively, the neighborhood size should be equal to the texture element sizes. Otherwise, the output texture may be too random or regular pattern may be reduced. The quality and speed of these pixel-based approaches [2, 14] were improved by path-based one. In [3], a patch-quilting procedure for texture synthesis is proposed, and then extended it to texture transfer. Patch-based texture transfer is similar to pixel-based one, except that instead of synthesizing pixels, it copies patches.

The work in [8] suggested texture optimization as texture synthesis method beyond pixel-based and patch-based algorithms. The algorithm synthesizes an output texture in the units of pixels, but unlike previous pixel-based methods that synthesize pixels one by one in a greedy fashion, this technique considered all pixels together, determining pixel values by minimizing a quadratic energy function. This energy function has been modified by the latest work in [4] to match the transfer task better. In details, both content and example style image were restricted by a segmentation mask adding to the energy function, in order to determine which parts to be transferred and preserved.

Recently, an impressive work of style transfer is using Convolutional Neural Networks (CNN)[7]. Their methods adopt a pre-trained CNN to extract features from both the style and the content images, respectively.

Motivated by [5], which consider an explicit probability density modeling of the problem and computes an approximate Maximum a Posteriori(MAP) solution based on an iterative optimization of Belief Propagation or Graph cuts, we propose a novel style transfer method for Chinese painting. Unlike the traditional patch-based algorithm in [3], we propose an adaptive patch for style transfer. Especially given that our target image in this work is black-and-white Chinese painting with expressive content, we improve classic style transfer algorithms by modifying the optimal match condition to overcome such a challenge.

3 Problem Description

Traditionally, Chinese painting (water-and-ink) is presented by ink diffusion of different degree on the Xuan paper. The objects are in a wide range of scale, painted by complex and expressive brush strokes. In other words, while some scene objects are always painted with rough brushwork, the key objects are painted in detail with subtle brushwork. For example, in Fig. 1(a), the distant mountains are roughly painted by great water-and-ink diffusion but the fisherman and the texture of the mountains nearby are exhibited subtly by slight ink spreading. Moreover, ink diffusion can be also used for rendering Chinese painting as “color”, such as the representation of cloud and shading.

Our goal is to transfer other artistic styles such as impressionism and post-impressionism to Chinese painting. Consequently, we propose a style transfer method that adopts an adaptive patch for patch-based texture transfer, and colorization to guide the process of style transfer. At first, we give the problem definition of style transfer for Chinese painting.

Given a Chinese painting $C : \Omega_C \in \mathbb{R}^3$, and a style image $S : \Omega_S \in \mathbb{R}^3$ with certain style. We aim to synthesis an image C_{out} which captures the style

of S while preserves the semantic content of C . This can be considered as finding a mapping $f : \Omega_C \rightarrow \Omega_S$ which confirms each element $X \in \Omega_C$ with a corresponding element $Y = f(X) \in \Omega_S$.

Applying a similar idea for patch-based texture transfer, the correspondence mapping f should be a piecewise constant translation mapping on region $P = \{P_i\}_{i=1}^n$ of Ω_C . In order to extract the style feature of S while preserving the structure of C , the region P should be obtained based on the painting elements of C , and the texture as well as color of S should be taken into account for the optimal corresponding $f(x)$. Especially, to transfer the style elegantly, smoothness is required on the boundary between neighboring regions.

4 Style Transfer for Chinese Painting

In this section we detail the proposed style transfer algorithm. In order to meet the requirements mentioned above, our approach can be divided into three main steps:

- Adaptively decompose Ω_C into n regions P ;
- Locally render Ω_C according to the color of S ;
- Find the optimal mapping f based on MRF model;

Moreover, corresponding experiment results are presented to illustrate the performance of each step. We note that our style transfer is accomplished in YUV color space, since we consider both luminance and chrominance in the process of texture transfer.

4.1 Adaptive Decomposition for Chinese Painting

We firstly recall that in patch-based texture transfer, the original image to be rendered is decomposed into fixed size patches, and assign one node of a Markov network. Generally, if the size of patches are small (for example the size of 8×8), the content of original image can be ideally reconstructed yet the style of the example style image is nearly obvious; on the contrary, if large patches have been chosen for texture synthesis, the considerable details of original image are lost. To reconstruct the realistic content of original Chinese painting while inheriting the example style, we divide the original image into adaptive-size patches based on its structure and pixel distribution.

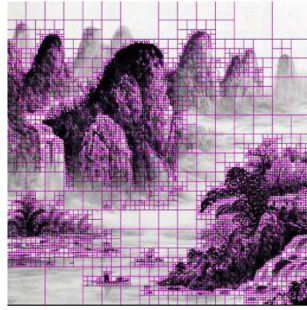
Let decomposition starts with one single region $P_i \in \Omega_C$, of size $m \times m$. Each region P_i is divided into four equal squares, with each size of $\frac{m}{2} \times \frac{m}{2}$, if pixel value $X_i = (x_1, x_2, \dots, x_{m \times m}) \in P_i$ satisfies:

$$D(X_i) > \sigma \text{ or } m > \omega \quad (1)$$

where σ is the threshold; $D(X_i) = (\max(X_i) - \min(X_i))$ presents the difference between the maximum and minimum value in region P_i , and ω is the maximum patch size allowed in the quadtree.



(a) Original image



(b) Adaptive decomposition



(c) Style image



(d) Style transfer result

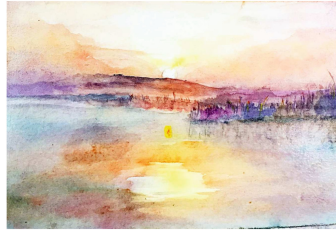
Fig. 1. Illustration of adaptive decomposition (Color figure online)

The local variance of a quadtree cell decides whether a cell is divided into four cells, which depends the details in C . As illustrated in Fig. 1(b), the more delicate elements in the original image are divided into the more smaller patches to be transferred, such as the trees and fisherman nearby. Thus, the content of original image can be perfectly preserved in texture synthesis, while the style feature can be reflected as much as possible, as showed in Fig. 1(d). Obviously, our decomposition only depends on the structure of original image, rather than the stopping criteria for quadtree splitting in [6].

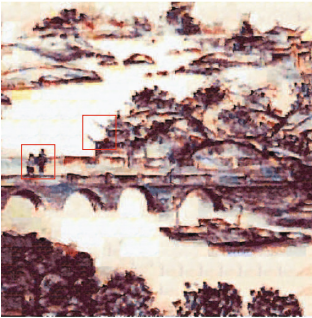
In Fig. 2, we present a comparison of between our adaptive patches and Image Quilting with fixed-size patches in [3]. To make it clear, We choose the smallest size of patch allowed in two algorithms, and highlight two specific differences in the results by red rectangles. It can be observed that two persons on the bridge and the curved roof of pavilion reconstructed by our method are more clearer than those reconstructed by Image Quilting as showed in Fig. 2(c) and Fig. 2(d). These results present that our adaptive patches preserves the original content better than fixed-size patches.



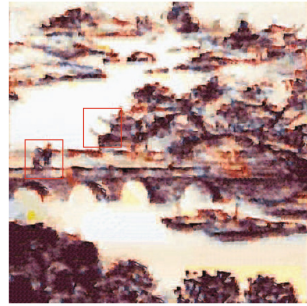
(a) Original image



(b) Style image



(c) Our method



(d) Image Quilting [3]

Fig. 2. Comparison of adaptive patches and fixed-size patches: Our method with adaptive patches and the minimal size of patches is 4×4 as showed in (c); (d) present the result of Image Quilting algorithm with patches of fixed size 8×8 . (Color figure online)

4.2 Locally Color Transfer

Color style transfer is an essential step in style transfer which has usually been done separately after texture transfer in classic approaches. Due to that the brightness and darkness in Chinese painting are exhibited by complex ink diffusion, the colors are usually gray, black and white, while the other artistic style is generally colorful. Without chrominance information, the color fidelity of example style cannot be guaranteed during reconstructing C_{out} . It is worse that the futile texture may appear which is not conform to semantic content of the original image. Thus, instead of transferring texture only in luminance, we consider the chrominance information.

Here, we preprocess colorization for original Chinese painting before texture transfer. Specific colors in S are extracted as color seeds for local rendering through colorization method suggested in [10]. Then, the rendered image \tilde{C} guides the texture transfer as one of criteria in chrominance. In detail, we search

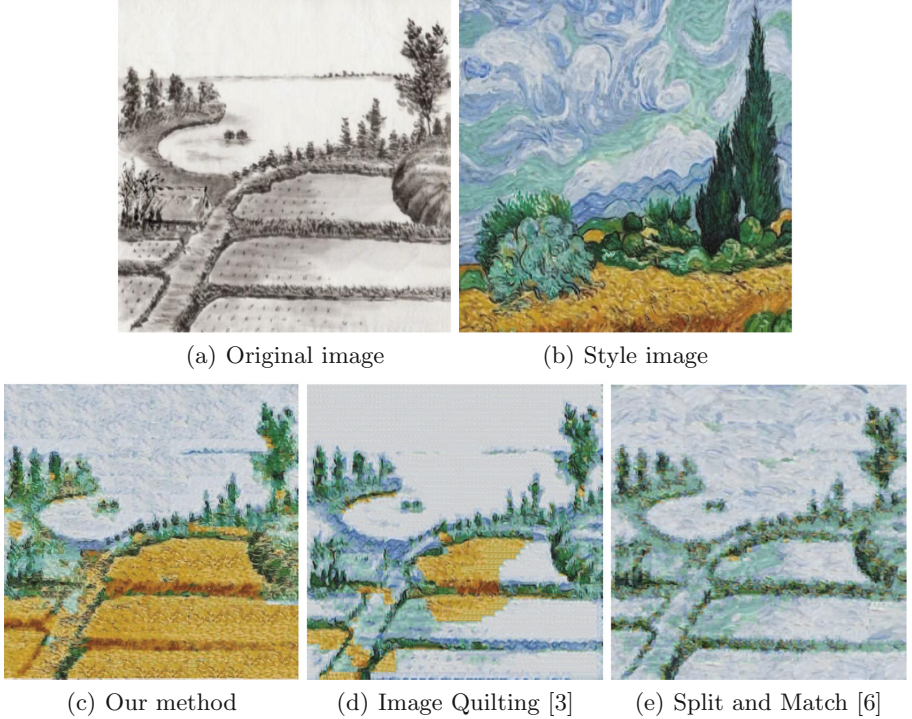


Fig. 3. Illustration of locally color transfer: Our result is more reasonable than the one of Image Quilting method, since there is futile texture on the farmland by Image Quilting [3]. And the color gamut of our result is more similar to the color gamut of style image compared to the results of Split and Match method [6]. (Color figure online)

for the optimal match for texture transfer in luminance as well as chrominance (in YUV color space), which is described in next 4.3.

Similarly, we show the transfer results in Fig. 3. It is noted that if consider luminance as the only matching condition for texture transfer, the futile texture are synthesized, as showed in Fig. 3(d). From semantic understanding, the color of farmland should be yellow or green, but Image Quilting algorithm (and other classic methods that only consider luminance) synthesizes blue and white texture. As presented in Fig. 3(c), compared with the traditional algorithms, our method can obtain a reasonable output image since the chrominance is considered. In addition, the color gamut of our result is more similar to the color gamut of style image than the results of Split and Match method shown in 3(e). It is indicated that the color style can be better extracted with chromatism information.

4.3 Optimal Match

As mentioned above, both the original image and example style image are divided into patches where each patch is one node of a Markov network. With the framework of Markov Random Field (MRF), the problem of patch-based style transfer can be solved through computing the Maximum Posteriori from a well chosen joint probability distribution on all patches [5]. Thus, the optimal mapping f can finally be found with MRF model.

The MRF model in our work is illustrated in Fig. 4, which can be found that the links on original image connect adaptive patches rather than fixed size patches. We search for the optimal match for each patch by finding maximum a posteriori (MAP), which is equally maximizing the joint probability over the X_i and Y_i , that can be written as

$$Pr(X_1, X_2, \dots, X_N, Y_1, Y_2, \dots, Y_N) = \prod_{(i,j) \in N} \Psi_{i,j}(X_i, X_j) \prod_{k \in N} \Phi_k(X_k, Y_k), \quad (2)$$

where $\Psi_{i,j}(X_i, X_j)$ are pairwise interaction potentials between neighboring nodes i and j , while $N(i, j)$ denotes the neighbors of patches. $\Psi_{i,j}(X_i, X_j)$ ensures that neighboring patches are similar in their overlapping region and it can be written as

$$\Psi_{i,j}(X_i, X_j) = \exp(-E(X_i, X_j)) \quad (3)$$

where $E(X_i, X_j) = \|X_i - X_j\|^2$ is the error term of the overlapping region between two patches. $\Phi_k(X_k, Y_k)$ are the data penalty functions given by

$$\Phi_k(X_k, Y_k) = \exp(-\theta(X_k, Y_k)). \quad (4)$$

where θ is the weighted error term between the newly chosen block and the old blocks. As discussed in 3.3, we use colored image \tilde{C} to guide the texture transfer, hence, $\theta[X_k, Y_k]$ is defined as

$$\theta(X_k, Y_k) = \alpha d(X_k, Y_k)_{Ori} + \beta d(X_k, Y_k)_{Ch} + \mu d(X_k, Y_k)_L. \quad (5)$$

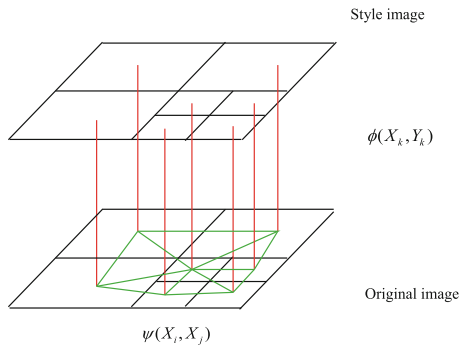


Fig. 4. Markov network for our work: Each node in the network describes a local adaptive patch of original or example image.

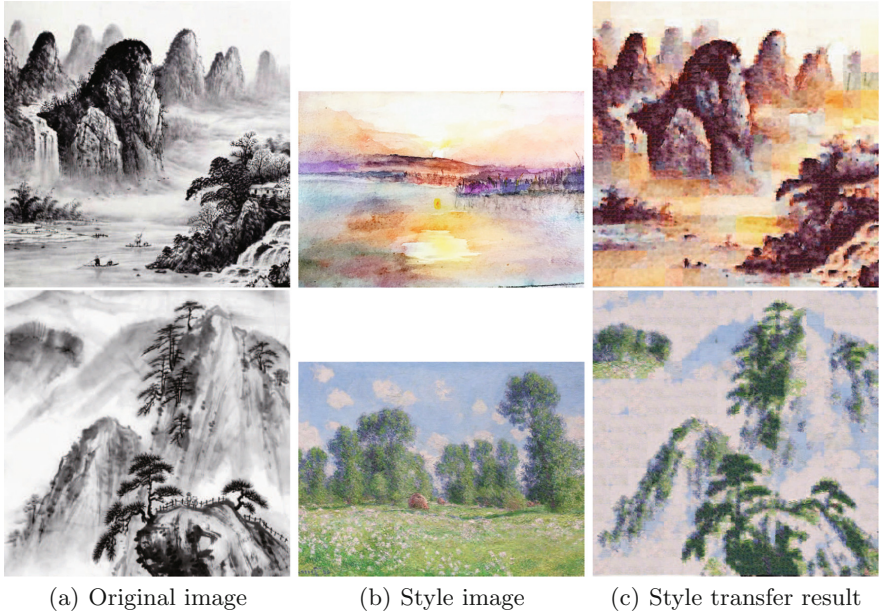


Fig. 5. Transfer results on different style examples: Original Chinese painting (left column), example style images (middle column), and the style transfer result (right column).

We modify the criterion in [3] by adding $d(X_k, Y_k)_{Ch}$, presenting the square error of patches between rendered image \tilde{C} and example style image S . $d(X_k, Y_k)_{Ori}$ is the square error of the overlapping regions in the original image C , and $d(X_k, Y_k)_L$ is the square error term of patches between original image and style image in luminance. α , β and μ are three positive weights that no bigger than 1 (respectively fixed to 0.2, 0.2 and 0.6 in all experiments).

Finally, we achieve an optimal boundary of adjacent patches to remove visibly artificial seams. This minimal cost path through the overlap region can be done with dynamic programming [1]. Other transfer results with respect to different example style are presented in Fig. 5. With different style, our algorithm is able to transfer example style while ideally reconstruct the content of the original painting.

5 Comparison of Our Method and Other Approaches

In this section, we would like to make a comparison between our method and recent style transfer approaches.

As shown in Fig. 6, we present the experimental results with our method and a popular method Convolutional Neural Network (CNN) with the parameter setting in [7]. Both our method and CNN achieve ideal reconstruction for



Fig. 6. Comparison with CNN approach: Original Chinese paintings (first column), different style images (second column), our results (third column), and results of CNN approach (last column). (Color figure online)

original content. The subtle texture feature of the style images can be captured with our method such as the wavy strokes in Van Gogh’s *Starry night*. Even the detail texture element like the yellow and white points are preserved in our result, which hardly appear in CNN transfer results. And the color gamut of our results is more closer to the color gamut of style images, compared with the results of CNN. This is due to that in the style transfer process, we choose the optimal patches in the original style image as the generated patches in stead of extracting the abstract style feature. While CNN uses deep and abstract style representation, it loses low-level pixel features of the style image. Moreover, CNN has the trade-off problem of style and content matching, which has been mentioned in [7]. Similarly, the transfer method in [11] applies MRF prior defining the loss function for CNN to control the abstract style layout yet our algorithm improves compatibility functions of MRF to generates style directly from the style image, rather than extracting the abstract style step by step.

As we mentioned in Sect. 4.1, our adaptive decomposition for the content image only depends on the local variance, while the recent work by Frigo et al., in [6] also regards the similarity between the content image and the style image as the decomposition criterion. Most importantly, compare with Split and

Match method transfers color style separately after texture transfer, we combine texture transfer and color style transfer, by guiding the texture transfer process with chromatism information. As depicted in Fig. 3, our result maintains the original color style of the style image including green, blue and yellow color. Yet the results of Split and Match method almost miss yellow color feature. The color gamut of our result is more closer to the color gamut of style image.

6 Conclusion

In this paper, we regard the rendering problem of Chinese painting as a style transfer issue and propose a new style transfer method for Chinese painting. Based on the characters of Chinese painting where the painting elements are always have obviously distinct scale, adaptive-size patches are applied for texture transfer in our approach. Additionally, we modify the constraints in texture transfer based on MRF model, considering color information of both style image and colorized original image. The local colors of style image are extracted as color seeds for rendering the black-and-white Chinese painting, which helps to guide the process of texture transfer.

The experimental results of each step are presented to clearly illustrate the improvement by our proposed algorithm. The results suggest that decomposing target Chinese painting with adaptive patches to be transferred is able to well preserve the original content while transferring example style, and the color style can be captured with chromatism information. Finally, we discuss the comparison of our method and other state-of-the-art style transfer methods, including patch-based approach and CNN framework.

References

1. Davis, J.: Mosaics of scenes with moving objects. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, p. 354 (1998)
2. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, p. 1033 (2002)
3. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of Siggraph, pp. 341–346 (2001)
4. Elad, M., Milanfar, P.: Style transfer via texture synthesis. *IEEE Trans. Image Process.* **26**(5), 2338 (2017)
5. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. *Int. J. Comput. Vis.* **40**(1), 25–47 (2000)
6. Frigo, O., Sabater, N., Delon, J., Hellier, P.: Split and match: example-based adaptive patch sampling for unsupervised style transfer. In: *Computer Vision and Pattern Recognition*, pp. 553–561 (2016)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423 (2016)
8. Kwatra, V., Essa, I., Bobick, A., Kwatra, N.: Texture optimization for example-based synthesis. *ACM Trans. Graph.* **24**(3), 795–802 (2005)

9. Kyprianidis, J.E., Collomosse, J., Wang, T., Isenberg, T.: State of the “art”: a taxonomy of artistic stylization techniques for images and video. In: Iberoamerican Optics Meeting and Latin American Meeting on Optics, Lasers, and Applications (2013)
10. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. *ACM Trans. Graph.* **23**(3), 689–694 (2004)
11. Li, C., Wand, M.: Combining Markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2479–2486 (2016)
12. Li, C., Wand, M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 702–716. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_43
13. Way, D.L., Lin, Y.R., Shih, Z.C.: The synthesis of trees in chinese landscape painting using silhouette and texture strokes. *J. WSCG* **10**, 499–506 (2002)
14. Wei, L.Y., Levoy, M.: Fast texture synthesis using tree-structured vector quantization, pp. 479–488 (2000)
15. Wong, H.T.F., Ip, H.H.S.: Virtual brush: a model-based synthesis of chinese calligraphy. *Comput. Graph.* **24**(1), 99–113 (2000)



The Accurate Guidance for Image Caption Generation

Xinyuan Qi, Zhiguo Cao^(✉), Yang Xiao, Jian Wang, and Chao Zhang

National Key Lab of Science and Technology of Multispectral
Information Processing, School of Automation,
Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
{silliam_qi, zgcao, Yang_Xiao, M201572352,
zhangC_22}@hust.edu.cn

Abstract. Image caption task has been focusing on generating a descriptive sentence for a certain image. In this work, we propose the accurate guidance for image caption generation, which guides the caption model to focus more on the principle semantic object while making human reading sentence, and generate high quality sentence in grammar. In particular, we replace the classification network with object detection network as the multi-level feature extractor to emphasize what human care about and avoid unnecessary model additions. Attention mechanism is utilized to align the feature of principle objects with words in the semantic sentence. Under these circumstances, we combine the object detection network and the text generation model together and it becomes an end-to-end model with less parameters. The experimental results on MS-COCO dataset show that our methods are on part with or even outperforms the current state-of-the-art.

Keywords: Image caption · Object detection · Attention mechanism
Deep learning

1 Introduction

Image caption task aims at automatically generating a descriptive sentence to describe the content of an image with an English sentence [1]. With the explosive increase in digital images and the rapid development in deep learning, teaching machines to understand images as humans is drawing great interests. At the outset, computer vision task aims at classifying the category of a single image (image classification). Hereafter, researchers try to locate the position of objects in more complicated scenes (object detection). After that, researchers further want to distinguish the category of per-pixel (semantic segmentation). Along with this fruitful development route, researchers owe it to comprehending the semantic information of the picture better and better. Meanwhile, another understanding of the images' semantic information is to describe an image's content with a human-like sentence (image caption). This idea is closer to human's habit when there is a scene in front of their eyes. While caption task seems obvious for human beings, it is much more difficult for machine since it requires the 'translation' model to capture several semantic information from a certain image. Such as scenes,

objects, attributes, relative position and so on. Another challenge of caption task is to generate descriptive sentence meeting the grammar rules.



Fig. 1. This is an example picture in MS-COCO dataset. The caption ground truth is “Several surfers riding a small wave into the beach”. The proportion of principal object (humans and surfboards) is well low. There are too much redundant information, such as sky, which will make it harder for attention mechanism to align the principal object with the noun composition in the descriptive sentence.

Recently, Neural network methods [2, 3] dominate the literature in image captioning. The encoder-decoder architecture in Neural Machine Translation [4] inspire these methods very much. In contrast to original Neural Machine Translation model, image caption model replace the recurrent neural network (RNNs) with convolutional neural network (CNNs) as encoder. CNNs encode the input image into a feature vector, which represents the semantic information of the image. Then a sequence modeling approach (e.g., Long Short-Term Memory (LSTM) [5]) decodes the semantic feature vector into a sequence of words. Such architecture applies to the vast majority of image caption model.

The method to combine CNNs and RNNs together directly will result that the information of the input image decreases by iterations. In this situation, researchers start to utilize image guidance [3], attributes [6] or region attention [7] as the extra input into LSTM decoder for better performance. The original intention comes from visual attention, which has been known in Psychology and Neuroscience for a long time. Attention mechanism highly relies on the quality of the input image. If there are too much redundant information in the image, it will be hard for attention mechanism to capture the principal information. As shown in Fig. 1, the proportion of principal objects (humans and surfboards) is very low. CNNs-encoder usually reduce the dimension of feature vector a lot, which will make it harder for attention mechanism to capture the information for subject, object and other noun composition. In this condition, if we insist on applying attention mechanism to the whole image like [7], caption model may not know what to describe.

In Natural Language Processing, scientists take the noun composition in a sentence as the focus, which people care more about. In image caption task, the noun composition corresponds to the principal object in an image. To help image caption model to

capture the principal object more accurate, we propose to get help from object detection task. Object detection task has been studied for a long time. CNNs framework is widely used and rapidly developed in object detection task, such as R-CNN [11], Fast-RCNN [12], Faster-RCNN [13]. These models are able to capture principal objects in the image very well. So we propose to make use of the feature of object detection methods to encode the image and generate guidance for the language generate model. We call it as accurate guidance. This advance also means to combine the higher level of semantic information in computer vision task with the semantic meaning in human-reading sentence.

We implement our model based on a single state-of-the-art object detection network Faster-RCNN [13], for accuracy and speed. Simultaneously, our model can be trained end-to-end, which will make the object detection module to adjust itself to suit for the image caption task. We take the Google NIC [7] as the baseline and compare our methods with popular attention models on the commonly used MS-COCO dataset [9] with publicly available splits of training, validation and testing sets. We evaluate methods on standard metrics. Our proposed methods outperform all of them and achieve state-of-the-art across different evaluation metrics.

The main contributions of our paper are as follows. First, we propose accurate guidance mechanism to help the caption model capture the principal object more precisely and infer their relationships from global information simultaneously. Second, the proposed method utilize a single object detection network as the multi-level feature extractor and demonstrates a less complicated way to achieve end-to-end training of attention-based captioning model, whereas state-of-the-art methods [3, 6, 19] involve LSTM hidden states or image attributes for attention, which compromises the possibility of end-to-end optimization.

2 Related Work

Recent successes of deep neural networks in machine translation catalyze the adoption of neural networks [8] in solving image caption problems. Early works of neural networks-based image caption include the multimodal RNN [10] and LSTM [5]. In these methods, neural networks are used to both image-text embedding and sentence generating.

Attention mechanism has recently attracted considerable interest in LSTM-based image captioning [3, 6]. Xu et al. [7] proposed to integrate visual attention through the hidden state of LSTM model. You et al. [6] propose to fusion visual attributes extracted from images with the input or output of LSTM. These methods achieve state-of-the-art performance but they highly rely on the quality of the pre-specified visual attributes. Our method also use attention mechanism. Different from the predecessors, we consider the object detection-dependent attention to generate high quality guidance rather than search at the whole noisy image. It is an adaptive method to obtain high quality features.

Reinforcement Learning has recently been introduced into image caption task [20] and achieved state-of-the-art performance due to optimize the evaluation metrics directly. These methods are generally applicable training approach not the

improvement for the caption model. Thus, we don't compare with them but believe that our model will gain much higher performance with Reinforcement Learning.

[19] first proposes to utilize object detection method in image caption task. However, it utilize Fast-RCNN to detect and VGG net [15] to locate. The caption model is very redundant. While generating guidance, it keep the region of its bounding box unchanged and set remaining regions to mean value of the training set for each object in image. This process will bring much interference to the caption model. Our method solves these puzzles by taking the single object detection network as the multi-level feature extractor. In this way, our method is a clean architecture for the ease of end-to-end learning.

3 Methods

Our accurate guidance model includes a multi-level feature extraction module (MFEM) and a principal object guiding LSTM (po-gLSTM). Figure 2 shows the structure of our model. We first describe how to use object detection network as MFEM to simultaneously extract the features of the whole image (fea_w) and principle objects (fea_o) in Sect. 3.1. Then, we introduce our po-gLSTM which will take advantage of the multi-level feature to guide the LSTM to describe the image more precise in Sect. 3.2.

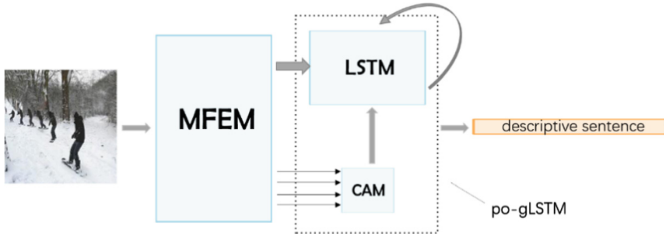


Fig. 2. The structure of our accurate guidance model

3.1 Multi-level Feature Extraction Module

Figure 3 shows the framework of multi-level feature extraction module. The MFEM consists of two parts: (1) fea_o extraction network (above the red dotted line); (2) fea_w extraction network (below the red dotted line). It is a variant of Faster-RCNN [13]. In order to capture the principle objects better, for an input image I , we suppose to utilize object detection network to find the potential objects and extract fea_o , which denoted as $fea_o = \{obj_1, \dots, obj_N\}$ and formulated as formula (1). N is the number of potential objects. RPN (Region Proposal Network) splits the principle object parts from the whole image. CNN_{θ_2} is to further extract the features after RPN.

$$fea_o = CNN_{\theta_2}\{RPN[CNN_{\theta_1}(I)]\} \quad (1)$$

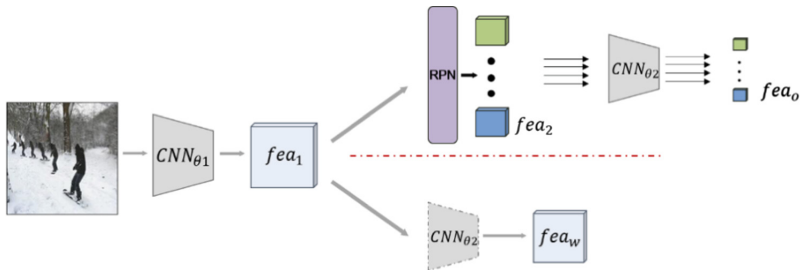


Fig. 3. The structure of the MFEM (Color figure online)

Simultaneously, we also need fea_w so that the po-gLSTM can get the information of scenes and infer the relationship between objects. In this situation, the original output of Faster-RCNN framework does not meet the conditions. Thus, we try to fix it's framework so that it can extract fea_w at the same time. As shown in the part below the red dotted line in Fig. 3, we get a copy of the feature after CNN_{θ_1} and take it into CNN_{θ_2} directly. Then we get an imitation classification network followed with fea_w , formulated as formula (2).

$$fea_w = CNN_{\theta_2}[CNN_{\theta_1}(I)] \quad (2)$$

Notice that the CNN_{θ_2} with dotted border (below the red dotted line) is the same with the CNN_{θ_2} with solid border (above the red dotted line). We do not increase the model parameters but obtain fea_w successfully. Faster-RCNN argues the size of input image should be larger than $600 \text{ pixel} \times 600 \text{ pixel}$. For reducing the model parameters, we replace its' fully connected layer with the Global Average Pooling layer to embedding fea_w and resize it to fit the size of the principle object guiding LSTM's input, formulated as follows:

$$x_0 = Pool_{ave}(fea_w) \quad (3)$$

x_0 is utilized to initialize decoder in Sect. 3.2. Here, we have already gotten the multi-level feature of the input image. The multi-level feature carries the multi-level semantic information. As later experiments will demonstrate, multi-level feature extraction module will help the model to focus more on the principle objects and achieve better performance.

3.2 Principal Object Guiding LSTM (po-gLSTM)

As shown in Fig. 4, the function of po-gLSTM is to decode the multi-level semantic information of the image and generate corresponding descriptive sentence. In this section, we will first introduce the condition attention module to obtain the principle object information for the current word. Then we will introduce how to make use of the principle object information to guide the LSTM to generate sentence. Both of above, we treat them as a whole and call it as po-gLSTM.

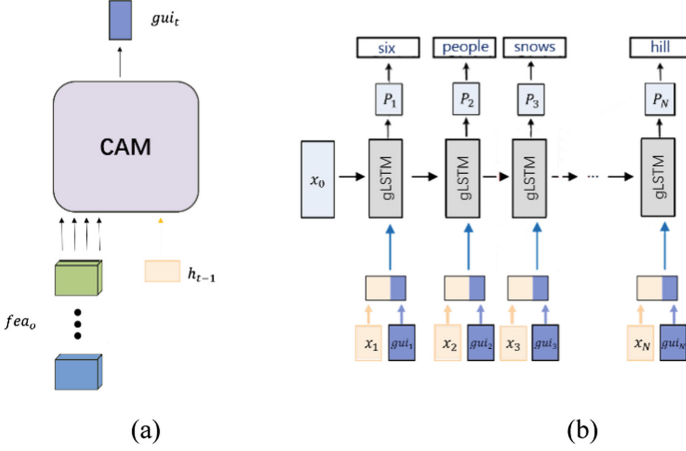


Fig. 4. (a) CAM is the condition Attention Module, which is to generate guidance information (gui_t) by principle object features (fea_o) and the information of hidden layer at previous step (h_{t-1}). (b) This sketch map shows how to utilize x_0 and gui_t to generate descriptive sentence. Both of (a) and (b) make up the po-gLSTM.

Condition Attention Module

With the multi-level feature extraction module, fea_o and fea_w of an input image will be extracted easily. Each word in caption is represented by a one-hot vector and the captioning sentence is a sequence of input vectors (x_1, \dots, x_T). Same as previous methods, we utilize fea_w to initialize the decoder (LSTM), the decoder then computes a sequence of hidden states (h_1, \dots, h_t) and a sequence of outputs (y_1, \dots, y_t). The primer decoder only accesses fea_w (encoded as x_0) once at the beginning of the learning process, which will loss most of the information of image I by iterations, and output incorrect words or stop too early. To avoid this, we proposed to utilize condition attention module (CAM) [6] to stress the role of principle objects and supply necessary information lost by iterations. CAM is formulated as followed:

$$a_t^i = W \tanh(W_{ao} obj_i + W_{ah} h_{t-1}) \quad i = 1, \dots, N \quad (4)$$

$$\alpha_t = \text{softmax}(a_t) \quad (5)$$

$$gui_t = \sum_{i=1}^N \alpha_t^i obj_i \quad (6)$$

W, W_{ao}, W_{ah} are learnable parameters. N is the number of principle object in an image. a_t^i is the relevance of obj_i and h_{t-1} . The elements of α_t is utilized to combine the guiding information (principle objects). gui_t is the guidance at iteration t .

With attention mechanism, model will know “where to see” while generating every word. We also make a visualization of attention mechanism to prove it in later experiment.

Guiding LSTM

The generated sentence by the LSTM model may lose track of the original image content since it only accesses the image content once at the beginning of the learning process, and forgets the image even after a short time. To make use of gui_t mentioned above and supplement the forgotten information if necessary, we propose to utilize an extension of the LSTM model, named the guiding LSTM (gLSTM) [3], which extracts semantic information from the input image and feeds it into the LSTM model every time step as extra information. Its' gate and memory cell can be formulated as follows:

$$i'_t = \sigma(W'_i[h'_{t-1}, x'_t, gui_t]) \quad (7)$$

$$f'_t = \sigma(W'_f[h'_{t-1}, x'_t, gui_t]) \quad (8)$$

$$o'_t = \sigma(W'_o[h'_{t-1}, x'_t, gui_t]) \quad (9)$$

$$\widetilde{C}'_t = \tanh(W'_c[h'_{t-1}, x'_t, gui_t]) \quad (10)$$

$$C'_t = f'_t C'_{t-1} + i'_t \widetilde{C}'_t \quad (11)$$

$$h'_t = o'_t * \tanh(C'_t) \quad (12)$$

$$x'_{t+1} = W'_{emb}(\log \text{softmax}(W'_h h'_t)) \quad (13)$$

Where W'_s denote learnable weighs, $*$ represent element-wise multiplication, $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent function, x'_t stands for input at t -th iteration, i'_t for the input gate, f'_t for the forget gate, o'_t for the output gate, C'_t for state of the memory cell, h'_t for the hidden state.

o'_t decides what to forget in C'_t . Its' decision is up to h'_{t-1} and x'_t . In original LSTM, when o'_t decides that forgetting some information is helpful for x'_{t+1} , it will be impossible for $x'_t (t' > t + 1)$ to utilize the forgotten information. The longer the descriptive sentence, the worse the condition like this is.

gLSTM is able to supplement the forgotten information if necessary. Condition attention module will also help to pick the most helpful principle object for x'_{t+1} . And we call our gLSTM with principle object condition attention module as op-gLSTM. Somebody may doubt weather emphasizing the principle object so much is helpful. Our experiment will verify that the model can infer the relationship better with stronger principle object information and it will cause no trouble for extracting the scene from fea_w .

One benefit of op-gLSTM is that it allows the language model to learn semantic attention automatically through the back-propagation of the training loss. While [19] only utilize objects and locations, other semantic information, such as scenes and motion relationship, is discarded.

4 Experiments

4.1 Dataset and Experiment Setup

Dataset

We use MS-COCO dataset [9] in our experiments. The dataset contains 123287 images respectively and each is annotated with 5 sentences using Amazon Mechanical Turk. There are 80 classes included in the dataset. We use 113287 images for training, 5000 images for validation and 5000 images for testing.

Experiment Setup

The inputting image is resized to 600 pixel \times 600 pixel. The training process contains three stages: (1) pre-train the object detection network (Faster RCNN) on MS-COCO dataset. (2) combine the multi-level feature extract module (a variant of the pre-trained Faster RCNN) with our po-gLSTM and train the po-gLSTM to equip it with the ability to decode. (3) train the integral model end-to-end to help our multi-level feature extract module and po-gLSTM fusion better. Four standard evaluation metrics, e.g. BLUE, METEOR, ROUGE_L, and CIDER, are used evaluate the property of the generated sentence.

4.2 Comparison Between Different CNNs Encoders

Encoder is used to extract the semantic feature of the input image. The property of the extracted feature is decisive to our caption model. To explore which encoder is more proper, we use three different CNNs in our experiments, including 50-layer and 101-layer ResNets [14] and 16-layer VGGNet [15]. Table 1 shows the experimental result.

Table 1. Results of different CNNs encoders. All values are reported as percentage (%).

CNNs encoders	MS-COCO dataset						
	B1	B2	B3	B4	M	R	C
Ours-VGG16	70.9	53.1	38.4	27.4	23.5	51.3	88.0
Ours-RESNET50	72	54.4	39.8	28.9	24.1	52.3	90.8
Ours-RESNET101	72.9	55.6	41.0	29.9	24.7	53.1	96

The experimental results show that deeper CNNs achieves higher scores on all metrics. This indicates that deeper CNNs can capture better semantic features, which contain more and better information for descriptive sentence generation. The guidance of deeper CNNs is much more accurate.

4.3 Comparison to the State-of-the-Art

Several related models have been proposed in Arxiv preprints since the original submission of this work. We also include these in Table 2 for comparison.

Table 2 shows the comparison results. Our models, both VGG16-based and RESNET101-based, outperform other models at the same scale in most metrics by a large margin, ranging from 1% to 5%. Models with attention mechanism, such as ATT [6], Det+Loc [19] achieve better score than models without attention mechanism, such as NIC [7] and LRCN [16]. Det+Loc [19] also utilize the object detection network whose scores are better than the models with classification network. Notice that, our VGG16-based model gets comparable performance with FC-2 K [20] (Resnet-101 based). Meanwhile, our RESNET101-based model outperforms FC-2 K in all metrics. it’s up to 5.1% in CIDER. Det+Loc is an object detection-based model, which utilize beam search (beam size 4) while testing. Without beam search, our VGG16-based model outperforms it in Blue_1 and CIDER and slightly inferior to it in other metrics. Det+Loc. introduce too much redundant information, which results in that its’ poorer performance.

Table 2. Results of different caption models. All values are reported as percentage (%).

Caption models	MSCOCO dataset						
	B1	B2	B3	B4	M	R	C
NICs	66.6	46.1	32.9	24.6	–	–	–
LRCN	62.8	44.2	30.4	21.0	–	–	–
m-RNN	67.0	49.0	35.0	25.0	–	–	–
Soft-Attention	70.7	49.2	34.4	24.3	23.9	–	–
Hard-Attention	71.8	50.4	35.7	25.0	23.0	–	–
g-LSTM	67.0	49.1	35.8	26.4	22.7	–	81.3
ATT	70.9	53.7	40.2	30.4	24.3	–	–
RA-SF	69.1	50.4	35.7	24.6	22.1	50.1	78.3
(RA-SF)-BEAM10	69.7	51.9	38.1	28.2	23.5	50.9	83.8
(Det.+Loc.)-BEAM4	70.4	53.1	39.2	29.0	23.8	52.1	85.0
FC-2K	–	–	–	28.6	24.1	52.3	90.9
Ours-VGG16	70.9	53.1	38.4	27.4	23.5	51.3	88.0
Ours-RESNET101	72.9	55.6	41.0	29.9	24.7	53.1	96

The results of comparison are strong evidence that (1) the object detection task does have the ability to help with image caption model and our multi-level feature extract module is better suitable for caption task. (2) Our end-to-end model can help the two modules merge to get better performance in caption task.

4.4 Comparison Between Different Beam Search Size

In this section, we introduce Beam Search (BS) to replace Maximum Probability Sampling Mechanism. BS is a heuristic algorithm, which will consider more situations to generate better sentence while testing. The larger the beam size is, the more situation will be considered. We take gLSTM as comparison and Table 3 shows the experimental results.

Table 3. Results of different Beam Size. All values are reported as percentage (%).

Beam size	Model	MS-COCO dataset						
		B1	B2	B3	B4	M	R	C
2	gLSTM	70.2	52.7	38.8	28.7	24.1	51.6	88.5
	Ours-VGG16	71.7	54.3	40.3	29.8	24.2	52.2	92.5
3	gLSTM	70.2	52.8	39.1	29.0	24.1	51.6	88.9
	Ours-VGG16	71.1	53.9	40.2	30.0	24.2	52.3	92.6
4	gLSTM	69.9	52.6	39.0	29.0	24.0	51.4	88.4
	Ours-VGG16	70.7	53.5	39.9	30.0	24.2	52.2	92.1

From Table 3, we can see that the performance of a model varies in different beam size. Simultaneously, our model always outperforms gLSTM and it surpass Det+Loc. at beam size = 4. This is another evidence that our accurate guided model is better than other methods.

4.5 Qualitative Results

Figure 5 shows qualitative captioning results. To emphasize the effectiveness of our accurate guidance model and for fair comparison, we compare our VGG16-based model with the baseline model (NIC).



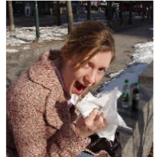

Examples		<p>NIC: a woman is eating a hot dog in a park.</p> <p>Ours: a woman is eating a slice of pizza.</p> <p>GT: There is a woman eating a slice of pizza.</p>	(a)
		<p>NIC: a bird is standing on a rock near a large body of water.</p> <p>Ours: a bird sitting on top of a pile of rocks.</p> <p>GT: A small orange bird standing on a collection of rocks.</p>	(b)
		<p>NIC: a man in a suit and tie standing in a park.</p> <p>Ours: a little girl that is holding a stuffed bear.</p> <p>GT: A girl sitting on a stone wall and eating.</p>	(c)
		<p>NIC: a man is riding a motorcycle on a dirt bike.</p> <p>Ours: a person jumping a dirt bike in the air.</p> <p>GT: A person up in the air with a motor bike.</p>	(d)

Fig. 5. Qualitative results: **NIC** is the baseline model; **Ours** means our VGG16 based model; **GT** is the ground truth.

The example images include similar colors and rare actions. Our proposed model can better capture objects in the target image, such as “a slice of pizza” in image (a) and “a little girl” in image (b). Our po-gLSTM can better capture the scenes and relationships between objects, such as “on a pile of rocks” in image (b), “in the air” in

image (d) and “holding” in image (c), “jumping” in image (d). Assuredly, our model may fail in some cases, such as “bear” in image (c). It is mainly due to there is no class named as “hamburger” while training the object detection network and the hamburger is covered with a white wrapping paper, which is hard for object detection task. If the performance of object detection task gets better, our proposed model can achieve better performance simultaneously. The qualitative result shows that object detection network does do much help to capture the principle objects. Our model does not loss the information of scenes and relationships between objects but it can even do better.

4.6 Visualization of Condition Attention Mechanism

In this section, we visualize the focus of CAM. The brighter part refers to higher attention. Taking the first row as example, our proposed model focus exactly on the bus in the image while generating the word-“bus”. When generating “parked”, the CAM focus more on where the car and ground contact. This indicates that our po-gLSTM does have the ability to focus on the effective objects all the time (Fig. 6).



Fig. 6. The visualization of condition attention mechanism on feature maps.

5 Conclusion

In this work, we propose the framework of accurate guidance for image caption. It combines a variety of object detection network (MFEM) and gLSTM with the help of attention mechanism (po-LSTM). We show in our experiments that the proposed methods significantly improve the baseline method and outperform the current state-of-the-art on MS-COCO dataset, which supports our argument of explicit consideration of getting help from object detection task.

References

1. Kulkarni, G., et al.: BabyTalk: understanding and generating simple image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1601–1608. IEEE Computer Society (2011)
2. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 677–691 (2015)
3. Jia, X., et al.: Guiding the long-short term memory model for image caption generation. In: IEEE International Conference on Computer Vision, pp. 2407–2415. IEEE Computer Society (2015)
4. Bahdanau, D., et al.: Neural machine translation by jointly learning to align and translate. *Comput. Sci.* (2014)
5. Hochreiter, S., et al.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
6. You, Q., et al.: Image Captioning with Semantic Attention. In: IEEE Computer Vision and Pattern Recognition, pp. 4651–4659. IEEE Computer Society (2016)
7. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. *Comput. Sci.* 2048–2057 (2015)
8. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Comput. Sci.* (2014)
9. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
10. Karpathy, A., et al.: Deep visual-semantic alignments for generating image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137. IEEE Computer Society (2015)
11. Ross, G., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Computer Vision and Pattern Recognition, pp. 580–587. IEEE Computer Society (2014)
12. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE Computer Society (2015)
13. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2015)
14. He, K., et al.: Deep Residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE Computer Society (2016)
15. Simonyan, K., et al.: Very deep convolutional networks for large-scale image recognition. *Comput. Sci.* (2014)
16. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 677–691 (2017)
17. Mao, J., et al.: Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint [arXiv:1412.6632](https://arxiv.org/abs/1412.6632) (2014)
18. Fu, K., et al.: Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2321–2334 (2015)
19. Yang, Z., Zhang, Y.-J., Rehman, S., Huang, Y.: Image captioning with object detection and localization. In: Zhao, Y., Kong, X., Taubman, D. (eds.) ICIG 2017. LNCS, vol. 10667, pp. 109–118. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71589-6_10
20. Rennie, S.J., et al.: Self-critical sequence training for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1179–1195. IEEE Computer Society (2017)



Large-Scale Visible Watermark Detection and Removal with Deep Convolutional Networks

Danni Cheng¹, Xiang Li¹(✉), Wei-Hong Li², Chan Lu¹, Fake Li¹, Hua Zhao¹,
and Wei-Shi Zheng²

¹ Ctrip Group, Shanghai, China

dcheng@Ctrip.com, lixiang651@gmail.com

² Sun Yat-sen University, Guangdong, China

Abstract. Visible watermark is extensively used for copyright protection with the wide spread of online image. To verify its effectiveness, there are many researches attempt to detect and remove visible watermark thus it increasingly becomes a hot research topic. Most of the existing methods require to obtain the prior knowledge from watermark, which is not applicable for images with unknown and diverse watermark patterns. Therefore, developing a data-driven algorithm that suits for various watermarks is more significant in realistic application. To address the challenging visible watermark task, we propose the first general deep learning based framework, which can precisely detect and remove a variety of watermark with convolutional networks. Specifically, general object detection methods are adopted for watermark detection and watermark removal is implemented by using image-to-image translation model. Comprehensive empirical evaluation are conducted on a new large-scale dataset, which consists of 60000 watermarked images with 80 watermark classes, the experimental results demonstrate the feasible of our introduced framework in practical. This research aims to increase copyright awareness for the spread of online images. A reminder of this paper is that visible watermark should be designed to not only be striking enough for ownership declaration, but to be more resistant for removal attacking.

Keywords: Visible watermark · Watermark detection
Watermark removal · Deep convolutional networks

1 Introduction

Image, serving as an important information carrier for E-commercial and social media, is widely employed and rapidly spreads nowadays. In modern life, many online images are embedded with visible watermarks for ownership declaration. In order to avoid the misuse of copyrighted images, it requires to perform watermark detection upon images before we use these images. Therefore, it is necessary

to develop a watermark detector that is able to automatically and accurately detect visible watermarks in images. Furthermore, as visible watermark plays an important role in copyright protection, for purpose to verify its effectiveness, a number of scientists attempt to attack it by removing watermark from images after detection. Visible watermark detection and removal increasingly becomes a hot research topic [1–6].

Developing robust visible watermark detection and removal methods remain as a challenging task due to the diversification of visible watermarks. More specifically, visible watermarks may consist of texts, symbols or graphic etc, leading to the challenge of extracting discriminative feature from unknown and diverse patterns of watermarks. In addition, the variations of the shape, location, transparency and size of the watermarks in various sorts of watermarked image makes it hard to estimate the area of watermark in practical situation.

Although researchers have extensively explored the visible watermark detection and removal problems [1–6], these works require handcraft feature from images which highly depends on the prior knowledge. Thus, developing a feasible approach that is able to tackle aforementioned challenges for watermarked images remains to be an unsolved problem. Recently, despite deep convolutional networks have shown their strong performance on feature representation for computer vision problems through taking advantage of massive image data, there is a lack of deep learning method for watermark detection as well as removal, and a lack of large-scale watermark dataset. Due to this fact, we contribute a large-scale watermark dataset and further utilize deep learning to generalize the detection and removal of unknown and diverse watermark patterns.

In this work, we propose a new visible watermark processing framework consisting of the robust large-scale watermark detection and removal components. Both of watermark detection and removal are build upon deep convolutional networks. Generally speaking, we exploit the trained watermark detector to locate the area where there is a watermark, which will be cropped out and used for the removal. To be more specific, we adopt the framework of current state-of-the-art object detectors as our watermark detection basic network, which is further implemented to be suitable for detecting and locating visible watermarks in images. In the removal procedure, we cast the watermark removal into an image-to-image translation problem, where we propose a full convolutional architecture to transfer the watermarked pixels into the original unmarked pixels effectively. Finally, both components are able to collaborate together to perform visible watermark detection and removal tasks automatically and consistently.

In summary, the main contributions of this work are: (1) It is the first work that formulate the visible watermark detection as an object detection problem and adapt existing detectors to make them suitable for automatical watermark detection. To achieve this, we contribute a new large-scale visible watermark dataset with dense annotations to facilitate the lack of large-scale image dataset for visible watermark detection task. (2) We propose an integrative deep learning based framework to fully address the visible watermark processing problem including detection and removal. Moreover, extensive comparison experiments

are conducted to evaluate our proposed framework and the experimental results demonstrate the effectiveness and efficiency of our proposed framework for complex visible watermark detection and removal tasks in real-world scenarios.

2 Related Work

Watermark Detection and Removal. In watermark detection and removal literature, existing methods can be divided into two categories: (a) single image schemes [1–3]. (b) stock images schemes (a large stock of images with same type of watermark) [4, 5]. For single image schemes, Santoyo-Garcia et al. [1] proposed to decompose a watermarked image and then distinguish the watermarked area from the structure image. Pei and Zeng [2] utilized Independent Component Analysis (ICA) for watermarked image recovery. These methods have to extract handcraft features from the whole watermarked image, which makes it very inefficient for these methods to be implemented for detecting and removing watermarks with diverse visible patterns. As for stock image schemes, Dekel et al. [4] proposed to estimate the outline sketch and alpha matte of watermarks from a batch of images. In this case, visible watermarks are regarded as foregrounds, whose attributes are required to be the same. Xu et al. [5] proposed an watermark removal technique which assumes the pending images have the same resolution and watermark region as those of training images. Despite the stock-based approaches can estimate the outline of watermark for stock images, these methods are not suitable for detecting and removing watermarks in real-world scenarios where the images are high potentially marked with unknown watermarks or the pattern of watermarks in different images might be distinct. To overcome these challenge, we proposed a new deep learning based framework which can effectively detect and remove watermark with unknown patterns.

Object Detection. Since we formulate the watermark detection as an object detection task in this paper, existing generic object detectors are related to ours. Currently the deep learning based object detection methods can be divided into two-stage approaches [8–10] and one-stage methods [11–15]. Since the one-stage methods take privilege of their high effectiveness and efficiency, they become the mainstream of object detection. For example, YOLOv2 and RetinaNet can obtain the state of art performance in accuracy with high speed (i.e. performing real-time object detection).

Image Inpainting. Related to watermark removal, image inpainting inpaints missing regions in an image, which gains huge benefit from a variety of Generative Adversarial Networks (GAN) based models [16, 17]. Different from image inpainting, in visible watermark removal, those pixels in watermarked area are not missing. They instead embedded some background information. Hence, in this work, we utilize the generator architecture to achieve the transformation between watermarked pixels and unmarked pixels, which is proved to be very effective in our work.

3 Methodology

In this work, we aim at automatically and precisely detecting unknown and diverse visible watermarks in images and exploring watermark removal in an effective and efficient way. In this section, we present our visible watermark processing framework. Firstly, a large-scale image dataset for visible watermark processing is introduced. In general, our whole pipeline can be divided into two separate modules: (1) the watermark detection module and (2) the watermark removal module. To be more specific, we illustrate our watermark detection module which is built on the existing deep learning based general object detection methods in Sect. 3.2 and the watermark removal one is detailed in Sect. 3.3. The illustration of our proposed visible watermark processing framework is shown in Fig. 1.

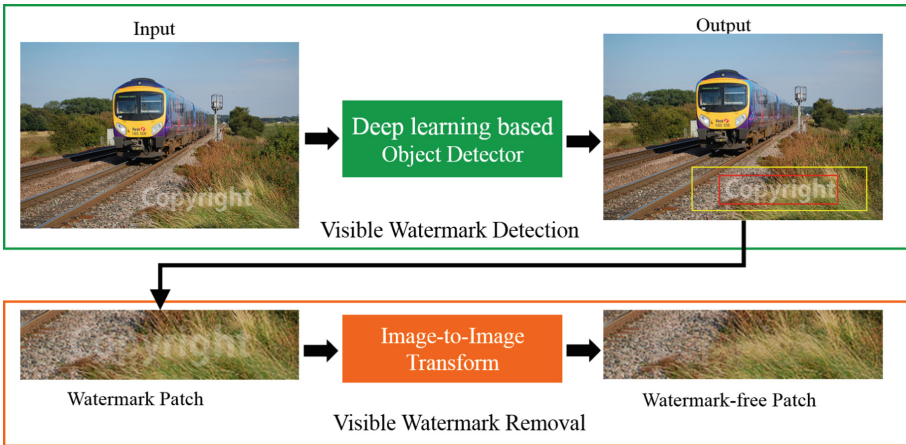


Fig. 1. The pipeline of our visible watermark processing framework. In the period of detection, the goal is to judge whether a image has watermark and locate the watermarked area (the red box). Then, we enlarge the detection boundingbox (the yellow box) and crop the watermarked patch to generate the input for watermark removal. (Color figure online)

3.1 Large-Scale Visible Watermark Dataset

At present there is no watermarked image dataset available for large-scale visible watermark detection and removal. To fill this gap, we contribute a new watermarked image dataset, containing 60000 watermarked images made of 80 watermarks, with 750 images per watermark. Specifically, the original images used in the training and test sets are randomly chosen from the train/val and test sets in PASCAL VOC2012 dataset [18] with replacement respectively. The

80 categories of watermarks cover a vast quantity of patterns, including English and Chinese, which are collected from renowned E-commercial brand, websites, organization, personal, and etc (see Fig. 2(a)). The entire watermarks are transferred into binary image with alpha channel for opacity setting. Furthermore, the size, location and transparency of each watermark in different images are distinct and set randomly. The diversity of watermarks makes our dataset more general (see Fig. 2(b)).

Another important distinction between our dataset and the conventional small-scale watermark dataset [4] is the watermarks in training set are not used for constructing images in test set. To be more particular, in existing watermark dataset, watermarks in training set and test set are exactly the same. This would lead to the situation where the watermark detector trained on such dataset can not work well on detecting unknown watermarks in images, which is impractical. Therefore, to meet the demand of watermark detection in real-world scenarios, in our dataset, watermarks in test set are different from those in training set. More specifically, train set contains 80% sorts of watermark and the test set includes the remaining.

In traditional pattern recognition tasks, object annotation is a time-consuming and tedious procedure. During generating watermarked image, we save the location size of the embedded watermark and original image at the same time. With our large-scale visible watermark dataset, it is possible to develop a significant deep learning based framework for facilitating visible watermark tasks.



Fig. 2. The diversity of our proposed large-scale watermark dataset.

3.2 Visible Watermark Detection

Visible watermark detection, one of fundamental topics in the computer vision field, is essential for various important applications, such as intellectual property protection in e-commerce, copyright declaration for business intelligence, and visual online advertising, etc. In this work, instead of directly exploiting an

existed watermark detector to detect watermarks at the beginning of our watermark processing framework, we consider to develop a new and more robust one.

From the machine learning perspective, watermark detection can be viewed as an two classification task, where the cropped image patches are classified into the watermark or background category. However, in real-world scenarios, images always contain various contents and the pattern, content, location, size, number of the watermarks in images are unknown. Developing a robust method to detect watermarks in images in the wild is inherent challenging and remains unsolved. In this paper, we formulate watermark detection as an object detection problem. Generally speaking, the recent deep learning based algorithms for generic object detection, e.g. Faster RCNN [10] YOLO [11, 12], RetinaNet [15] are appropriate for our detection task.

Figure 1 shows the proposed deep learning based framework for watermark detection in images. To be more specific, our model takes as input a watermarked image and estimates the probabilities of all candidates with different scale and ratio at all location in the image classified as the area which is tightly covered by a watermark. Considering that the efficiency of watermark method is one of most important criterions in watermark detection, we adopt the one-stage detection methods in our watermark detection framework.

Thanks to the large-scale watermark dataset proposed in this work, our proposed watermark detector can be trained effectively. More importantly, our proposed method can detect watermarks in images effectively and efficiently under unknown condition such as the unknown watermarks in images and so on.

3.3 Visible Watermark Removal

Once the watermarks in images are accurately detected, the detection results can be used for further image-based watermarks processing such as watermark removal, watermark recognition, etc. In this work, we mainly investigate the former task, the watermark removal, and develop image transformation based method for it.

Image transformation, where an image transformation model takes as input an image and generate a different image to facilitate specific tasks, is one of the popular computer vision topics. Examples like image denoising, super-resolution, image style translation, etc., have taken significant steps since convolutional neural network serves as an indispensable foundation for these works. Inspired by the success of image transformation using deep learning technique [16, 19], we propose an effective visible watermark removal system based on deep neural networks.

As shown in Fig. 3, the system consists of two components: watermark removal network and loss network. Each watermarked patch x is fed into the watermark removal network to obtain the estimated watermark free patch \tilde{y} . Then the $L1$ loss and perceptual loss are calculated based on the ground truth and the estimated patches.

The whole network is trained to minimize the loss function via the combination of the two during training. During the test procedure, merely a forward

transformation is required via passing watermarked patch through the watermark removal network.

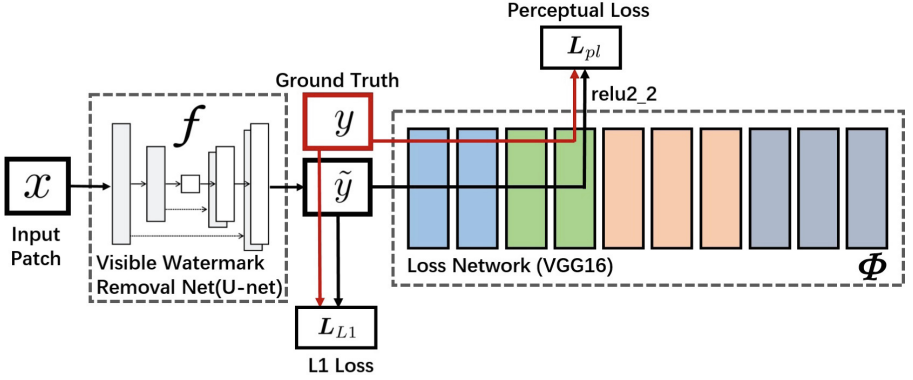


Fig. 3. The illustration of our proposed system for visible watermark removal. We leverage the U-net architecture for transferring visible watermarked patch into the watermark free one. The difference between the outputted watermark free patch and the ground truth watermark free patch is calculated using $L1$ loss, and perceptual loss is exploited for measuring the perceptual features of visible watermark. Therefore, the total loss of the watermark removal module during training process comprise the $L1$ loss and the perceptual one. The loss network for calculating perceptual loss is pre-trained on ImageNet for image classification, which remains fixed during the training process.

Network Architectures. Rather than transferring a whole image pixel-to-pixel, our work focuses on partial transformation task (i.e. transfer a specific patch of a image). More specifically, pixels inside the detected area are expected to be recovered to unmarked condition, while those in unmarked area in the watermarked image will remain unchanged. Specifically, we adapt the architecture of our removal network as that of the U-net [7]. This network is mirror symmetrical in structure, with skip connection between corresponding blocks. In this way, the shallow features near to the input get combination with those high-level features so that the low-level features such as location and texture of input image can be preserved.

Objective Function. The $L1$ loss penalizes the pixel distance between the ground truth and the output, which has been proved to have good performance in matching the pixel value of the input with those of the ground truth, and synthesizing the output [16]. Hence it is adopted in our network and is denoted as L_{L1} .

$$L_{L1}(x, y) = \|f(x) - y\|_1, \quad (1)$$

where x denoted as an input watermarked patch detected and cropped from an watermarked image, y refers to the ground truth patch without watermark. $f(x)$

is the output of U-net. As $L1$ loss is calculated based on per-pixel value in a whole image, it will be huge when each pixel has a small change and the image has little difference in visual. Besides, as the perceptual loss, which has been proved to be efficient in capturing the semantic information of the source image, depends on high level feature from convolutional layer, using perceptual loss can result in a more realistic output. Supposed the feature size of the j_{th} convolutional layer of loss network is $C_j \times H_j \times W_j$, the convolutional transformation is denoted as Φ_j and \tilde{y} is the estimated watermark free patch which is equal to $f(x)$. The formulation of the perceptual loss can be expressed as:

$$\mathbf{L}_{pl}^{\Phi_j}(\tilde{y}, y) = \frac{1}{C_j H_j W_j} \|\Phi_j(\tilde{y}) - \Phi_j(y)\|_2^2. \quad (2)$$

In our work, we leverage the **relu2_2** feature from VGG-16, which is similar to the work in [19]. Consequently, in order to obtain a more visual pleasure results for visible watermark removal, we combine benefits of these two loss functions, which can keep the details of input information as well as the perceptual information. Thus, the objective function of our removal network is:

$$\mathbf{L}_{whole} = \mathbf{L}_{L1} + \alpha \mathbf{L}_{pl}^{\Phi, relu2.2}, \quad (3)$$

where $\alpha \geq 0$ is a weight for regularizing the effect of $L1$ loss and perceptual loss.

4 Experiments

In order to evaluate our proposed framework, we conduct comprehensive experiments on our large-scale visible watermark dataset introduced in Sect. 3.1. In this work, both components in our proposed framework, the watermark detection and removal modules, are evaluated and the experiments are conducted on a computer cluster equipped with NVIDIA Tesla K80 GPU with 12 GB memory. The experimental details of these two components are illustrated and analyzed individually. It should be noted that existed methods cannot handle images with unknown watermark patterns, thus they are not suitable for the case that we deal with in this paper.

4.1 Visible Watermark Detection

Settings. We presume the proposed watermark detection framework can take any recent deep learning algorithms for generic object detection. In our work, one two-stage method Faster RCNN [10] and two one-stage methods YOLOv2 [12] and RetinaNet [15] are adopted to verify our assumption. In order to make the generic object detector suitable for watermarks detection in images, we adapt the number of class to two (i.e. watermark or background), and follow the training strategy on object detection [10, 12, 15] to train our watermark detection networks.

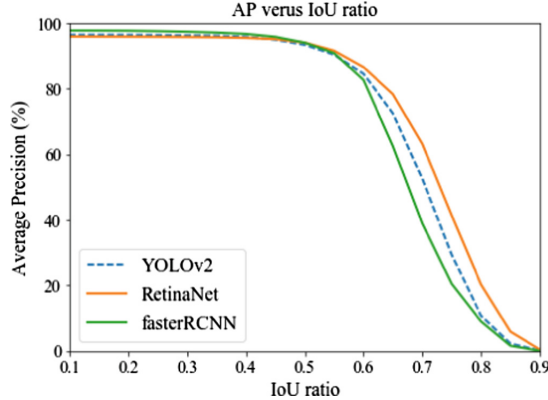


Fig. 4. Evaluation of Intersection over Union (IoU) parameter settings for watermark detection performance (AP)

Table 1. Evaluation of visible watermark detection

Method	AP _{50:95}	AP ₅₀	AP ₇₅
Faster RCNN	40.0	94.0	20.5
YOLOv2	69.9	93.4	29.4
RetinaNet	72.0	94.1	41.4

As we formulate visible watermark detection as an object detection task, we follow the standard object detection evaluation metric to validate the effectiveness of our visible watermark detector, which is the Average Precision (AP) under defined Intersection over union (IoU).

Results and Analysis. Figure 4 shows the AP curves versus IoU threshold of the watermark detection models using Faster RCNN [10], YOLOv2 [12] and RetinaNet [15]. From the figure, it is clear to see that the AP of the visible watermark detection models stays at around 100% when IOU is smaller than 0.4 and the difference between these three models are very small. This promising results imply that the visible watermark model which is obtained by finetuning existing object detection model on our visible watermark dataset can be effective on detecting unknown visible watermark patterns. With the IoU increasing, the AP curves drop dramatically. However, this has limited influence on our work as watermark detection does not require very precise location of watermark bounding box in real-world scenarios. Furthermore, it is evident that the watermark detection model using one-stage method RetinaNet improves AP significantly over Faster RCNN and YOLOv2. This indicates that the focal loss introduced in RetinaNet can result in a more precise detection results for the small and unapparent visible watermarks target.

To have a rounded analysis, we present the results of visible watermark detection in Table 1. The value of AP_{50:95}, AP₅₀ and AP₇₅ are listed, where AP_{50:95}



Fig. 5. Visualization of detection examples on our large-scale watermark dataset with RetinaNet. The red box with the predicted watermark confidence score shown on the top of the box is predicted by our watermark detection model using RetinaNet, while the blue box shown on the bottom of the blue box is the groundtruth with IoU ratio between the groundtruth box and the predicted one. (Color figure online)

is the average of AP under IoU threshold ranging from 0.5 to 0.95. These results validate the excellent performance of RetinaNet.

In order to evaluate the performance of our watermark detector, we visualize the watermark detection results of some testing examples in our collected dataset and show them in Fig. 5. The results in the figure indicate that our watermark detector is strong enough to detect those watermarks with different scales, transparency, location and various pattern from background clutter. It verifies that formulating the visible watermark detection as an object detection task is feasible.

4.2 Removal

Settings. For visible watermark removal, we build up our U-net with four down-sampling blocks. Specifically, the input patch and the ground truth one are

cropped from the marked image and the source one according to the predicted watermark bounding box of our watermark detection model using RetinaNet. Here, the center of both cropped marked and ground truth patches center at the center of the detected watermark bounding box and the size is 1.5 times larger than that of the predicted watermark bounding box to ensure that the watermark target can be included in the cropped patches. We further round the size of both cropped patches (i.e. height and width of the patches) to be a multiplier of 16, which is required to meet the input requirement of the U-net. During training, we adopt Adam optimization algorithm with initializing the learning rate as $2e-4$, and the batch size is set to be 1. The α for regulating perceptual loss and L1 loss is adjusted to $1e-6$.

The metrics which we adopt to evaluate the effects of watermark removal is the same as that of [4], including Peak Signal to Noise Ratio (PSNR) and Structural dissimilarity Image Index (DSSIM), both of which are adopted to measure the similarity between the predict watermark free patch and the ground truth one.

Table 2. Evaluation of visible watermark removal

Metrics	Input	Perceptual loss	L1 loss	Ours
PSNR	20.65	29.86	30.42	30.86
DSSIM	0.103	0.051	0.045	0.043

Results and Analysis. We calculate the average value of PSNR and DSSIM over the whole test set. Table 2 gives the PSNR and DSSIM of our model using different types of loss. As shown in Table 2, our removal model can have significant improvement in comparison over the input image. Besides, the results of the combination of the L1 loss and perceptual loss is shown to be better than those of single type of loss.

As shown in Fig. 6, despite the pattern of watermarks in images shown from the first row to the fourth row is quite diverse, our watermark removal algorithm performs well on removing visible watermarks. More specifically, some watermarks are some English words or letters, while some of them are the combination of English words, Logo and etc. However, our proposed method is able to extract the invariant feature of the watermarks and generate the image patches which is almost the same as the original ones. In addition, we report the removal results of our model using different sorts of loss, which are subtle distinct. The results in Fig. 6 indicate that our model using the loss combining the L1 loss as well as perceptual loss can exploit the strength of both loss to wipe out the visible watermarks and meanwhile keep the fine details of the source images, yielding powerful reconstruction performance.

We also conduct experiments to compare the performance of different architectures. Observing the results of the encoder-decoder architecture mentioned

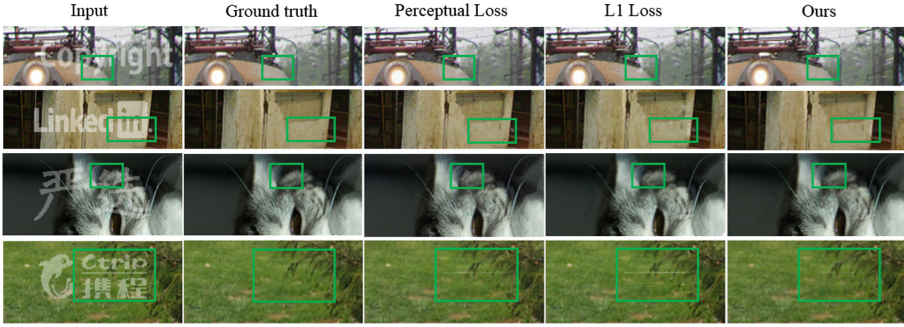


Fig. 6. Different losses induce different removal results. The last three column shows the output results trained under a different loss.

in [16] (The encoder-decoder is created by severing the skip connections in the U-Net), we find that it alters the global brightness and there exists local watermark residual in local area. Thus the watermarked patch is hard to be restored to get similar to its watermark free condition. The outputs of U-net architectures are more similar to the ground truth patches, which is applicable for our removal task. The results in Fig. 7 demonstrate that our U-net architecture is more effective, as it does not break surrounding information by allowing low-level information to be shortcut across the network.

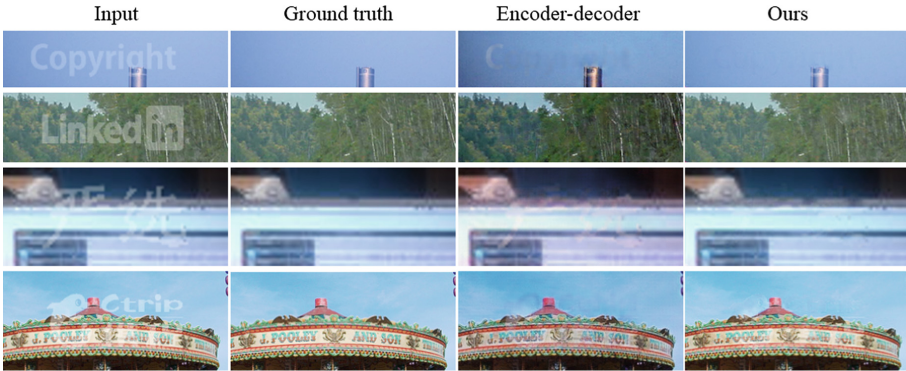


Fig. 7. Example results of different architectures and images of the groundtruth. These experimental results demonstrate that the U-net architecture can be more effective for visible watermark removal.

4.3 Discussions

Our experiments show that our proposed framework can effectively deal with the large-scale visible watermark tasks. For watermark detection, our watermarks detection model using one-stage method RetinaNet perform very well on

detecting visible watermark. During watermark removal, the size of bounding box is expanded to a little larger than the size of the detected watermark patch to alleviate the effect of partial detection, and our network can adaptively transform the marked pixels to watermark free ones and do not corrode the other pixels at the same time. Therefore, setting a small IoU threshold to capture the watermarked patches as much as possible, and then expanding and inputting these patches into our removal net, can ensure the performance of our proposed framework.

5 Conclusion

This paper presents a new deep learning based framework for large-scale visible watermark processing tasks, which consist of two components: (1) watermark detection, which is formulated as an object detection task. (2) watermark removal, which is transferred into an image-to-image translation problem. Besides, we build a large-scale visible watermark dataset for training and evaluating deep learning based framework for watermark detection, watermark removal and so on. In addition, extensive experiments are conducted to verify the feasible of our proposed pipeline. Experimental results show that our proposed framework is effective on watermark detection and removal.

Acknowledgment. Danni Cheng and Xiang Li equally contributed to this work. The authors would like to thank Dongcheng Huang and Xiaobin Chang’s valuable advice on paper writing.

References

1. Santoyo-Garcia, H., Fragoso-Navarro, E., Reyes-Reyes, R., et al.: An automatic visible watermark detection method using total variation. In: IWBF 2017 (2017)
2. Pei, S.C., Zeng, Y.C.: A novel Image recovery algorithm for visible watermarked images. *IEEE Trans. Inf. Forensics Secur.* **1**, 543–550 (2006)
3. Huang, C.H., Wu, J.L.: Attacking visible watermarking schemes. *IEEE Trans. Multimed.* **6**(1), 16–30 (2004)
4. Dekel, T., Rubinstein, M., Liu, C., et al.: On the effectiveness of visible watermarks. In: CVPR 2017 (2017)
5. Xu, C., Lu, Y., Zhou, Y.: An automatic visible watermark removal technique using image inpainting algorithms. In: ICSAI 2017 (2017)
6. Qin, C., He, Z., Yao, H.: Visible watermark removal scheme based on reversible data hiding and image inpainting. *Sig. Process.: Image Commun.* **60**, 160–172 (2018)
7. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI 2015 (2015)
8. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR 2014 (2014)
9. Girshick, R.: Fast R-CNN. In: ICCV 2015 (2015)
10. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS 2015 (2015)

11. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: CVPR 2016 (2016)
12. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: CVPR 2017 (2017)
13. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint (2018)
14. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
15. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: ICCV 2017 (2017)
16. Isola, P., Zhu, J.Y., Zhou, T., et al.: Image-to-image translation with conditional adversarial networks. In: CVPR 2017 (2017)
17. Pathak, D., Krahenbuhl, P., Donahue, J., et al.: Context encoders: feature learning by inpainting. In: CVPR 2016 (2016)
18. Everingham, M., Eslami, S.M.A., Van Gool, L.: The pascal visual object classes challenge: a retrospective. IJCV **111**(1), 98–136 (2015)
19. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43



Learning to Generate Realistic Scene Chinese Character Images by Multitask Coupled GAN

Qingxiang Lin¹, Lingyu Liang¹, Yaoxiong Huang¹, and Lianwen Jin^{1,2}(✉)

¹ School of Electronic and Information Engineering,
South China University of Technology, Guangzhou, China
lhlqx2014@gmail.com, lianglysky@gmail.com, hwang.yaoxiong@gmail.com,
lianwen.jin@gmail.com

² SCUT-Zhuhai Institute of Modern Industrial Innovation,
South China University of Technology, Zhuhai, China

Abstract. Scene text recognition, is challenging due to the large appearance variances of the scene character. Recently, deep learning technique has shown its power for scene text recognition, but it requires enormous annotated data for training and it is time-consuming to manually obtain abundant data for all the categories of characters. This paper proposes a new architecture, called multitask coupled generative adversarial network (MtC-GAN), for scene Chinese character recognition (SCCR). The MtC-GAN consists of coupled GAN networks for scene character style transfer and classifier networks trained by the style-transferred data generated by the coupled GAN. To make the generated data be realistic enough for SCCR, we train the multitask networks using a new loss function that combines the constrains of encoders, generators and classifiers simultaneously. Experiments show that the proposed MtC-GAN framework is general and flexible to improve the accuracy for SCCR.

Keywords: Scene Chinese character recognition
Generative adversarial networks · Multitask training

1 Introduction

Scene text recognition (STR) has been drawing ever-increasing research interests in recent years given its potential for many applications, such as autonomous driving [1, 2], license plate recognition [3, 4] and industrial automation [5, 6]. Although traditional optical character recognition has been extensively studied, naively adapting the technique to STR may fail to perform well, especially for scene Chinese character recognition (SCCR). The main challenge of SCCR lies in the large appearance variances of the scene character caused by style, font, resolution, illumination, projection transformation or partially occluded.

Recently, deep learning technique has been introduced into the field of STR [7–9]. The deep neural networks (DNN) consists of hierarchical nonlinear transformation, and is allowed to learn the feature and classifier with great invariant

and discriminate properties. The developed system with DNN structure obtains the state-of-the-art performance for SCCR. However, it requires enormous annotated data to train and fine-tune the DNN-based system. Although large-scale benchmark databases have been constructed for STR and SCCR [10], it is still time-consuming to obtain abundant labels, and the large categories of SCCR may also suffer from data imbalance. For instance, in the recently proposed CTW dataset [10], Chinese character samples of common categories can exceed the 17000 entries, whereas some rare categories contain only one sample. Therefore, it would be significant to generate scene Chinese character images for SCCR using DNN architecture.

The generation of scene Chinese character images can be divided into rule-based and learning-based methods. For the rule-based scheme, Campos et al. [11] generated English characters to train a character-level English scene text classifier; Jaderberg et al. [12] create a synthetic word data generator through physical rendering process to train a whole-word-based English scene text classifier; Gupta et al. [13] proposed a fast and scalable engine to generate synthetic images of text in clutter which further consider the local 3D scene geometry, and then train a text localisation network. The abovementioned methods which are limited by their rule-based nature seems to hardly simulate all the important variances in the real-world. For example, the work of [13] is limited by the segmentation and depth prediction of background images.

The learning-based method is mostly motivated by the GAN architecture [14], which can estimate the target distribution, and then generate similar images to the real ones. Although the previous X-GAN framework can have many advantages, it can't be ensured that each samples generated by GAN methods can preserve annotation information, and the naively synthetic data generated by GAN method may fail to improve the prediction performance due to these bad samples.

To tackle this problem, we propose a multitask coupled GAN framework for scene Chinese character recognition, which generates realistic scene Chinese character and improves the classification accuracy by the generated data simultaneously. The MtC-GAN consists of coupled GAN networks for scene character style transfer and classifier networks trained by the style-transferred data generated by the coupled GAN. To make the generated data be realistic enough for scene Chinese character recognition, we propose a new loss that combines the constrains of encoders, generators and classifiers simultaneously. Experiments show that the synthetic data by our method have great visual consistency to the realistic data. Furthermore, classifiers with different deep structures, like ResNet18 [15], ResNet34 [15] or VGG16 [16], can obtain apparent performance improvement, which indicate that the proposed multitask coupled GAN framework is general and flexible to improve the accuracy for SCCR.

The contributions of our work can be summarized as follows:

- A multitask coupled GAN learning framework for SCCR, which is general and flexible to generate realistic data and improve the accuracy of the classifier by generated data simultaneously without extra human annotation efforts;

- A new loss that combines the constrains of encoders, generators and classifiers to regularize the learning of the multitask coupled GAN.
- We qualitatively and quantitatively assess the classifier performance to demonstrate the effectiveness of the proposed method.

2 Related Works

Scene text image generation is a challenging task given the presence of complex background and font diversity. Many researchers have proposed the generation of realistic scene text images. Campos et al. [11] generated English character images to train a character-level English scene text classifier. Jaderberg et al. [12] create a synthetic word data generator through physical rendering process to train a whole-word-based English scene text classifier. Gupta et al. [13] proposed a fast and scalable engine to generate clutter-text synthetic images considering local 3D scene geometry, and then train a text localisation network. However, these methods are limited by their rule-based nature. For instance, the method in [13] is limited by the segmentation and depth prediction of background images. Unlike the abovementioned methods, we propose a learning-based method to generate realistic scene Chinese character images and further improve the recognition performance.

As one of the most considerable improvements on the research of deep generative models [17, 18], GANs [14] are being intensively studied by the deep learning and computer vision communities alike. A GAN basically consists of generator and discriminator networks, where the former generates samples to increase the discriminator error rate, and the latter aims to distinguish real from synthetic images. This adversarial training allows the generator to estimate the target distribution and then generate similar images to the real ones. Mathematically, the standard GAN training aims to solve the following optimization problem:

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

To extend the abilities of GANs, Mirza et al. [19] proposed a conditional GAN to direct data generation by conditioning both the generator and discriminator on additional information. This type of GAN has been successfully used in plenty of applications, such as image super-resolution [20, 21], image style transfer [22–25], domain adaptation [26], etc.

Furthermore, conditional GANs are suitable for image-to-image translation, which has been applied for different purposes including the generation of maps from aerial photos and colorization of grayscale images. Conditional GAN is well suited for this task and many researchers have achieved great success based on it. Likewise, Isola et al. [22] proposed the pix2pix model to learn the mapping from input to output images using paired images. Zhu et al. proposed CycleGAN [23] based on a cycle consistency loss to break the limit of training with paired images. Liu et al. [25] proposed an unsupervised image-to-image translation (UNIT) network assuming a shared latent space. Azadi et al. [27] proposed

the multi-content GAN(MCGAN) for few-shot font style transfer. Shrivastava et al. [28] proposed a simulated and unsupervised SimGAN to enhance the realism of an image simulator while preserving annotation data and demonstrated a high performance with no labeled real data. Zhao et al. [29] proposed a dual-agent GAN(DA-GAN) to enhance the realism of a face simulator output by using unlabeled real-face images while preserving identity information. Our proposed multitask coupled GAN combines the advantages of the UNIT network [25] and DA-GAN [29] to improve the quality of synthetic images and consequent classifier performance.

3 Multitask Coupled GAN

3.1 Source Data

We first propose a synthetic character generator that retrieves simple Chinese character images through font rendering, affine transformation, and perspective transformation. We denote the synthetic data generated in this way as source data \mathbf{x}_s . By using diverse TrueType and OpenType font files obtained from the Internet, we generate plenty of simple Chinese character images with annotation information. In addition, we use real image dataset published by Yuan et al. [10] and denote it as \mathbf{x}_t . We aim to simultaneously reduce the difference between \mathbf{x}_s and \mathbf{x}_t and improve the performance of a scene Chinese character classifier.

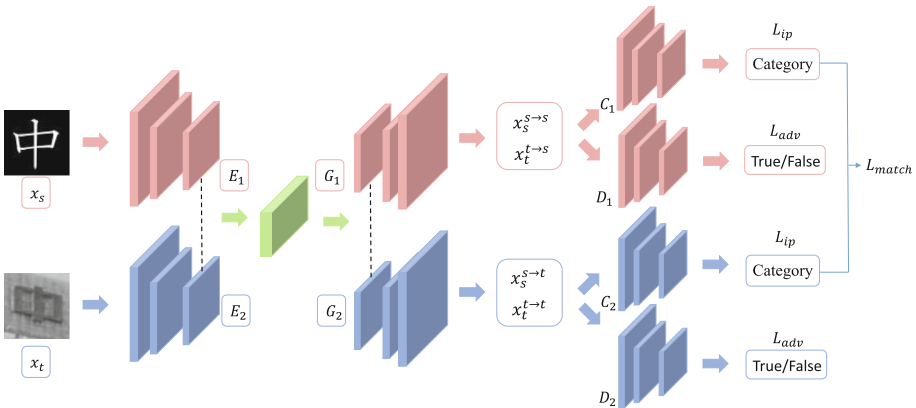


Fig. 1. Diagram of the proposed multitask coupled GAN architecture. E_1 and E_2 are two encoding functions that map images to latent codes. G_1 and G_2 are generation functions that map latent codes to images. D_1 and D_2 are adversarial discriminators for the respective domains. C_1 and C_2 are classifiers for the respective domains. L_{ip} , L_{adv} and L_{match} are the identity perception, adversarial, and matching losses, respectively. The dash lines denote weight sharing.

3.2 Coupled Generator

The same Chinese characters can present appearance variations in natural images arising from complex backgrounds and writing styles. Still, humans can easily recognize these characters, suggesting that the same characters written with different styles might share high-level semantic characteristics in the human brain. This semantic similarity can be represented by a map from characters with different styles into the same latent space, and an inverse map from a latent space into different domain images. Consequently, if the same characters with different styles are mapped into a latent space, we can generate corresponding images in two domains using autoencoders. To this end, we use concepts of coupled GAN [30] and UNIT network [25] to establish a shared latent-space assumption through a weight-sharing constraint. The architecture of the proposed MtC-GAN model is illustrated in Fig. 1 and relies on a UNIT network, where generator loss L_{unit} is formulated as:

$$L_{unit} = L_{VAE_1}(E_1, G_1) + L_{GAN_1}(E_1, G_1, D_1) + L_{CC_1}(E_1, G_1, E_2, G_2) + \\ L_{VAE_2}(E_2, G_2) + L_{GAN_2}(E_2, G_2, D_2) + L_{CC_2}(E_2, G_2, E_1, G_1) \quad (2)$$

where L_{VAE} denotes the variational autoencoder loss, L_{CC} denotes the cycle-consistent loss [23], L_{GAN} denotes the standard adversarial loss [14]. and D , G , and E denote adversarial discriminators, generators and encoders, respectively. More details on the loss functions can be found in [25]. The loss constraint can only add realism to synthesized images in appearance, but hardly preserves annotation information well. However, to use the synthesized data for improving classification performance, the synthesized images should preserve annotation information. Therefore, we include identity perception loss L_{ip} that is a multi-class cross-entropy loss to preserve annotation information. Then, we update the generator parameters by minimizing the following loss:

$$L_G = L_{unit} + \lambda_1 L_{ip} \quad (3)$$

where hyperparameter λ_1 control the weights of the objective terms. This combined loss both enhances the realism of synthetic images and preserves annotation data.

3.3 Multitask Discriminator

The discriminator aims to distinguish real from synthesized images. Its loss is given by:

$$L_{adv} = \log D_1(x_s) + \log(1 - D_1(G_1(E_2(x_t)))) + \\ \log D_2(x_t) + \log(1 - D_2(G_2(E_1(x_s)))) \quad (4)$$

In addition, we train a classifier to preserve label information of the generated data using identity perception loss L_{ip} defined as:

$$L_{ip} = \sum_n -Y_s \log D_{c_1}(x_s) + \sum_n -Y_t \log D_{c_1}(G_1(E_2(x_t))) + \sum_n -Y_t \log D_{c_2}(x_t) + \sum_n -Y_s \log D_{c_2}(G_2(E_1(x_s))) \quad (5)$$

where D_{c_1} and D_{c_2} are the probabilities of class n output by classifier C_1 and C_2 , respectively. Y_s and Y_t are the labels of \mathbf{x}_s and \mathbf{x}_t , respectively. The definitions above derive in a multitask training that preserves label information of the synthetic data. In addition, we can generate any amount of training data for training supervised models.

To further constrain classifiers C_1 and C_2 , we define a matching loss, formulated as:

$$L_{match} = \sum_i |D_{c_1}(x_s) - D_{c_2}(G_2(E_1(x_s)))| + |D_{c_2}(x_t) - D_{c_1}(G_1(E_2(x_t)))| \quad (6)$$

Where i is the class index. This loss improves the classifier performance. Likewise, we define another constraint in the generator to improve the quality of the generated data by training the discriminator to minimize combined loss:

$$L_D = L_{adv} + \gamma_1 L_{ip} + \gamma_2 L_{match} \quad (7)$$

where hyperparameters γ_1 and γ_2 weigh the corresponding objective terms.

We optimize MtC-GAN by alternatively optimizing multitask discriminator and coupled generator for each training iteration until the whole network converge.

4 Experiments and Results

We evaluated the performance of the proposed MtC-GAN mainly on the CTW dataset [10]. Although the most commonly used metric for determining the quality of generative models is the inception score [31], it does not suit our objective of using the generated data to improve the classifier performance. Instead we use two complementary evaluation metrics. First, similar to [28], we deploy the ‘Visual Turing Test’ to evaluate the visual quality of the generated images. Second, we use generated data to train a classifier, and compare the performance among classifiers with different generation methods.

4.1 GAN Training

We used a recently released Chinese text detection and recognition dataset, the CTW dataset [10]. It is split into training, validation and testing dataset, where the validation dataset was used for evaluating all the experiments. Similar

to [10], we only consider recognition of the top 1000 most frequently observed character categories. In addition, we evaluated a simple classifier to determine the enhancement provided by the generated images. Specifically, the classifier that we used is the ResNet18 [15], whereas the architecture of generator and discriminator was the same as that of the UNIT network [25]. The encoders consisted of 3 convolutional layers as the front-end and 4 basic residual blocks [15] as the back-end. The generators consisted of 4 basic residual blocks as the front-end and 3 transposed convolutional layers as the back-end. The discriminators consist of 6 convolutional layers. Then, an Adam solver [32] was adopted for the MtC-GAN with learning rate of 0.0002, $\lambda_1 = 1$, $\gamma_1 = 1, \gamma_2 = 5$.

4.2 Generated Image Quality

In this section, we deployed the ‘Visual Turing Test’ [28] to quantitatively evaluate the visual quality of the generated images and designed a simple user study where subjects were asked to classify images as being either real or synthetic. Each subject observed a random selection of 40 real and 40 synthetic character images that were randomly presented, and was asked to label the character images as either real or synthetic. We used the classification accuracy for quantitative evaluation, whose outcomes are shown in Table 1. The classification accuracy among subjects was 57%, which is very close to a random selection, i.e., 50%. Consequently, we considered that the subjects were unable to distinguish between real and synthetic images.

Table 1. Results of the ‘Visual Turing test’ where subjects classified real and synthetic images. The average classification accuracy among subjects was **57%**, close to the **50%** of random selection.

	Selected as real	Selected as synthetic
Ground truth real	225	175
Ground truth synthetic	169	231

Figure 2 shows examples of characters generated using the proposed method that served to quantitatively evaluate its outcomes.

4.3 Classifier Performance

The goal of this study was to use generated data for improving the classifier performance, and thus the classification accuracy was our main concern. Table 2 lists the classification accuracy using different generation methods. We can see that, naively learning from synthetic data can undermine classification accuracy due to the difference between synthetic and real image distributions, whereas the proposed MtC-GAN generation method achieves the best performance among

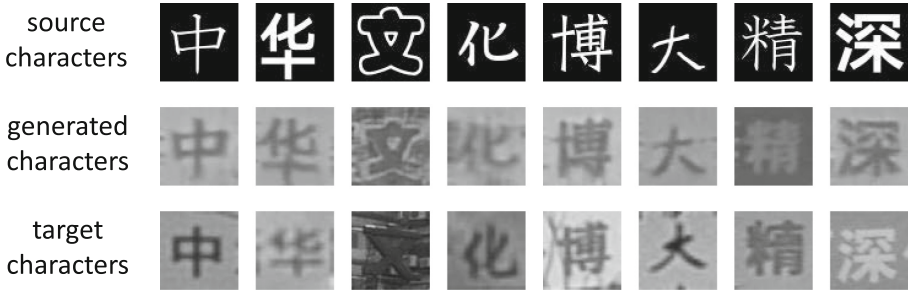


Fig. 2. The generated images using multitask coupled GAN. From top to bottom: source characters, generated characters, target characters.

Table 2. Classification accuracy of different generation methods

Generation method	Classification accuracy
Real data only	76.3%
Real data + source data(x_s)	75.5%
Real data + synthtext2014 [12]	78.5%
Real data + synthtext2016 [13]	78.2%
Real data + SimGAN [28]	77.2%
Real data + CycleGAN [23]	77.8%
Real data + UNIT [25]	78.5%
Real data + proposed MtC-GAN	80.7%

Table 3. Classification accuracy of different classifiers with and without the generated images

Classifier	Real data	Real data+MtC-GAN
ResNet18 [15]	76.3%	80.7%
ResNet34 [15]	78.5%	82.2%
VGG16 [16]	81.3%	83.5%

the compared methods, suggesting that multitask training can improve the classifier performance.

To further verify the effectiveness of the proposed method, we use different classifiers, whose accuracies are listed in Table 3. Every classifiers using data generated from the proposed MtC-GAN exhibits the best performance. Furthermore, the ResNet18 with multitask training can have better performance than the ResNet34 [15] without multitask training. It shows that if we can generate images which are realistic enough, we can train a shallow network enjoying the comparable performance with a deep one.

5 Conclusions

We propose a multitask coupled GAN (MtC-GAN) for realistic annotation-preserving image synthesis. The generated scene Chinese character images improve the performance of character classifiers. Both qualitative and quantitative evaluations demonstrate the effectiveness of the proposed MtC-GAN method and its superior performance. The experimental results also suggest that if we can generate images which are realistic enough, we can train a shallow network enjoying the comparable performance with a deep one.

Acknowledgement. This research was supported in part by GD-NSF (No. 2017A030312006), the National Key Research and Development Program of China (No. 2016YFB1001405), the National Natural Science Foundation of China (No.: 61673182, 61771199, 61502176), GDSTP (No.: 2014A010103012, 2017A010101027), GZSTP (No. 201607010227) and Fundamental Research Funds for the Central Universities (No. 2017BQ058).

References

1. Bojarski, M., et al.: End to end learning for self-driving cars. arXiv preprint [arXiv:1604.07316](https://arxiv.org/abs/1604.07316) (2016)
2. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3D object detection network for autonomous driving. In: IEEE CVPR, vol. 1, p. 3 (2017)
3. Björklund, T., Fiandrotti, A., Annarumma, M., Francini, G., Magli, E.: Automatic license plate recognition with convolutional neural networks trained on synthetic data. In: 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6. IEEE (2017)
4. Masood, S.Z., Shu, G., Dehghan, A., Ortiz, E.G.: License plate detection and recognition using deeply learned convolutional neural networks. arXiv preprint [arXiv:1703.07330](https://arxiv.org/abs/1703.07330) (2017)
5. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Meghini, C., Vairo, C.: Deep learning for decentralized parking lot occupancy detection. *Expert Syst. Appl.* **72**, 327–334 (2017)
6. Song, X., Kanasugi, H., Shibasaki, R.: DeepTransport: prediction and simulation of human mobility and transportation mode at a citywide level. In: IJCAI, pp. 2618–2624 (2016)

7. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4168–4176 (2016)
8. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(11), 2298–2304 (2017)
9. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: towards accurate text recognition in natural images. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5086–5094. IEEE (2017)
10. Yuan, T.-L., Zhu, Z., Xu, K., Li, C.-J., Hu, S.M.: Chinese text in the wild. *arXiv preprint [arXiv:1803.00085](https://arxiv.org/abs/1803.00085)* (2018)
11. De Campos, T.E., Babu, B.R., Varma, M.: Character recognition in natural images (2009)
12. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint [arXiv:1406.2227](https://arxiv.org/abs/1406.2227)* (2014)
13. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315–2324 (2016)
14. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)* (2014)
17. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. *arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114)* (2013)
18. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint [arXiv:1401.4082](https://arxiv.org/abs/1401.4082)* (2014)
19. Mirza, M., Osindero, S.: Conditional generative adversarial nets. *arXiv preprint [arXiv:1411.1784](https://arxiv.org/abs/1411.1784)* (2014)
20. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint* (2016)
21. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 318–333. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_20
22. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. *arXiv preprint* (2017)
23. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint [arXiv:1703.10593](https://arxiv.org/abs/1703.10593)* (2017)
24. Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGAN: unsupervised dual learning for image-to-image translation. *arXiv preprint* (2017)
25. Liu, M.-Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, pp. 700–708 (2017)
26. Ganin, Y., et al.: Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**(1), 2030–2096 (2016)

27. Azadi, S., Fisher, M., Kim, V., Wang, Z., Shechtman, E., Darrell, T.: Multi-content GAN for few-shot font style transfer. arXiv preprint [arXiv:1712.00516](https://arxiv.org/abs/1712.00516) (2017)
28. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 3, p. 6 (2017)
29. Zhao, J., et al.: Dual-agent GANs for photorealistic and identity preserving profile face synthesis. In: Advances in Neural Information Processing Systems, pp. 65–75 (2017)
30. Liu, M.-Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 469–477 (2016)
31. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
32. Kinga, D., Adam, J.B.: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)



A Recognition Method of the Similarity Character for Uchen Script Tibetan Historical Document Based on DNN

Xiaojuan Wang¹, Weilan Wang¹(✉), Zhenjiang Li¹, Yiqun Wang¹, Yuehui Han², and Zhanjun Hao³

¹ Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

wangweilan@xbmu.edu.cn

² College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, Gansu 730000, China

³ College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu, China




⁴ Library of Northwest Minzu University, Lanzhou, Gansu 730000, China

Abstract. In order to improve the similarity character recognition of Tibetan historical document, this paper applied the Depth Neural Network (DNN) to similar characters recognition of Tibetan historical document, and proposed a recognition method of the similarity character for Uchen Script Tibetan based on deep learning. The effective feature learning and recognition are automatically carried out by DNN. We also introduced a sample labeling method of Tibetan historical document of Uchen Script using unsupervised clustering and constructing sample sets of the similar characters. Compared with the traditional methods such as Support Vector Machine (SVM) and Naive Bayes Classifier (NBC) based on gradient features through simulation experiment, our method can achieve better performance. The proposed method can learn feature effectively and avoid the disadvantages of manual feature selection and extraction, and it can improve recognition rate greatly. With the increasing of training samples, the recognition rate was improved more significantly. The experimental results show that the proposed method used for similar characters of Tibetan historical document Uchen Script recognition, higher recognition rate can be obtained.

Keywords: Deep neural network (DNN) · Deep learning
Convolutional neural network (CNN) · Tibetan
Similar character of Uchen script

1 Introduction

The characters of Tibetan historical document cover modern Tibetan and Sanskrit Tibetan, so the number of characters is more than 7,000. The similarity between characters of Tibetan historical document is high and there are a lot of similar characters, such as “འཇ”, “ཇ”, “ཇལ”, “ཇལ”, “ཇལ”, “ཇལ”, “ཇལ”, “ཇལ”, etc., which bring a larger technical

difficulty to character recognition. In addition, many Tibetan historical documents are carved on the woodblock, which was engraved by hands, so the nicks are usually uneven. Therefore, the late manual inkiness is uneven, for example, the deep groove has less ink, leading to a loss part strokes of character of historical documents; Or a loss of strokes caused by the Image preprocessing of the ancient books, for example, “ཨ” “མ” “ལ” are changed into “” “” “”, which undoubtedly increases the difficulty of character recognition of Tibetan ancient books. At present, there is a lack of researches on the image and character recognition of Tibetan ancient books.

SVM method [1], hidden Markov model [2] and so on are more widely used in character recognition. Convolution neural network is a deep neural network which has a local connection between layers and which was put forward by American scholar LeCun. After the appearance of convolution neural network (CNN), using a variety of types of deep neural network models to analyze and recognize documents has become a research hotspot in this field. CNN has been successfully used in many areas, such as the recognition of handwritten digits, English characters, Chinese character and so on. Among 107 papers collected in ICFHR meeting held in late October 2016, whose image analysis and retrieval [3], text line segmentation [4], feature extraction [5], classification recognition processing [6] and other links involved in Chinese, English, Japanese, Mongolian, Arab, Bangladesh, etc., and more than half of the papers applied the deep learning technology. The Tibetan language includes modern Tibetan language (also known as Tibetan language or local Tibetan language) and Sanskrit Tibetan language (the Tibetan transferring form of Sanskrit). The print form of modern Tibetan characters has been studied a lot, such as professor Ou Zhu at Tibet University, professor Huang Heming at Qinghai Normal University, professor Li Yongzhong at Jiangsu University of Science and Technology, etc. And the team of professor Ding Xiaoqing at Tsinghua University studied, researched and developed the Tibetan character recognition system of practical multifont printing of more than 592 characters [7, 8], which has been well applied. The literature [9–13] shows that, for handwritten character recognition, the statistical characteristics of characters are the best, and for the off-line handwritten Chinese character recognition, gradient feature has a high recognition rate [14–16]. The researchers successfully applied the convolution neural network to digit recognition [17, 18] and character recognition [19, 20] in the natural scene, and pointed out that the convolution neural network could learn the characteristics which are better than artificial design [21, 22]. The literature [23] applied the deep convolution neural network to the recognition of offline handwritten similar characters, and the recognition rate is more significantly improved than traditional method. Therefore, this thesis proposes to use the deep convolution neural network to conduct the recognition of similar Tibetan characters. In contrast, there is no report about the application of deep convolutional neural network in the character recognition research of Tibetan ancient books.

Due to the irreproducibility of Tibetan ancient books, sample extraction of Tibetan characters can only be extracted from the document and image itself of Tibetan ancient books, and the project team has realized the preprocessing, binarization and layout analysis of document and image of Tibetan ancient books, and completed the document character segmentation. Due to the printing requirement of “soft character fine

alignment and fine carving” in the Phyi dar of Tibetan Buddhism, most of the Buddhist texts adopted Uchen Script. The striking feature of Uchen Script is that the top stroke of each letter is horizontal and straight, and the base line of the character arrangement is on a straight line. See Fig. 1. The baseline (baseline 1, baseline 2, etc. expressed by the dotted line in Fig. 1) is adopted to further segment into the vowel part above the baseline. For example, baseline 1 is adopted to express the character “ལྷོ”, “ལྷོ” and so on above the baseline; The part under the baseline, such as “ལྷོ”, “ལྷོ”, “ལྷོ”, etc. There are fewer types of characters above the baseline, about a dozen types, and there are also fewer types of similar characters. This thesis mainly studies the similar characters of the characters under the baseline.

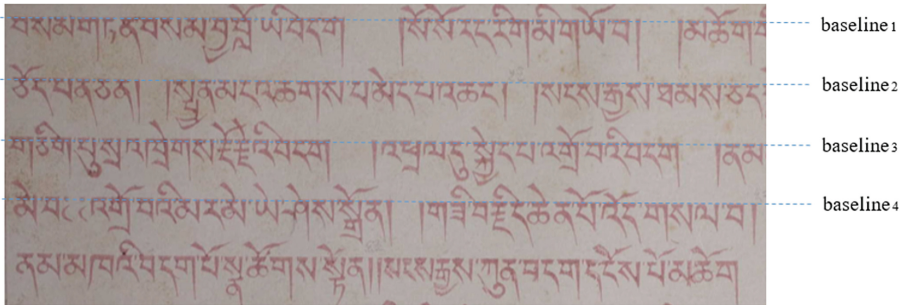


Fig. 1. Document image of Tibetan ancient books (a part)

2 Construct Sample Set of Similar Characters

In view of the current situation that there is no character sample of Tibetan ancient books, the following methods are proposed to classify and label the similar character sets.

In view of the Tibetan characters which have been segmented early, first of all, their characteristics are extracted, and three features about extraction in this paper are:

- (1) Gradient 8 direction characteristics (64 D)

First of all, the character image of Tibetan ancient books is normalized to 136×50 , and in order to ensure the less distortion of the image, bicubic interpolation is adopted for the deformation process. Then the uniform grid of 4×2 is used to evenly divide the original image into 8 small grids according to the size, and then the gradient feature of character pixels in each small grid is calculated. Then, the gradient is decomposed into 8 directions in accordance with the method of Bai to form 8 D gradient direction characteristics [24], and then 8 small grids features are combined to get 64-dimensional gradient direction characteristics.
- (2) Features of 8×8 grid (64 D)

In the first place, the character image of Tibetan ancient books is transformed into 64×64 , and in order to ensure a less distortion of the image, the deformation process adopts bicubic interpolation. Then, the original image is evenly divided

into 64 small grids by using the even grid of 8×8 , and later, the percentage of the characters in each small grid in the total pixel is calculated, and the characteristics of 64-dimension are obtained.

(3) Peripheral features of characters (64 D)

The grids which are divided and extracted by using feature (2) to continue to extract the pixel periphery features from top to bottom, from bottom to top, from left to right and from right to left. The features of four directions are combined into one-dimensional features, and 64 small grids have a total of 64-D features.

After integrating the above three characteristics, there are a total of 192 D feature dimensions. Through principal component analysis, the dimension is reduced to 80 D features. k-means clustering is used to record the filename of each character and the corresponding relationship of the distance of each centroid. According to the sorting characters in the class, the former k characters which are divided into the same class and which are in a close range are divided into similar characters, constituting a set of similar characters. MATLAB is used to copy the image of similar characters in the same file, and the distance information is added before the image's original file name. Then, according to the sort of file name, the image of the same category of characters can be gathered as far as possible (Fig. 2).

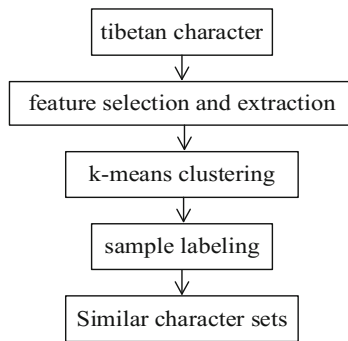


Fig. 2. Construction process of similar character set of Tibetan Uchen script

3 Convolution Neural Network (CNN)

Convolution neural network (CNN) is a neural network which is specially used to deal with similar network structure data, such as image data which can be considered as a two-dimensional pixel grid. CNN shows a high recognition rate in 2 D image recognition application, and its network structure is highly invariant to translation, scaling, tilting or other forms of deformation. CNN directly conducts the learning and character classification for the characteristics of original image, and it doesn't need too much pre-processing and feature extraction of the original character image, so it is an end-to-end recognition system, which effectively avoid the defects of losing the details of similar characters caused by artificial feature extraction and feature selection in advance. This thesis adopts the following CNN network structure, as shown in Fig. 3.

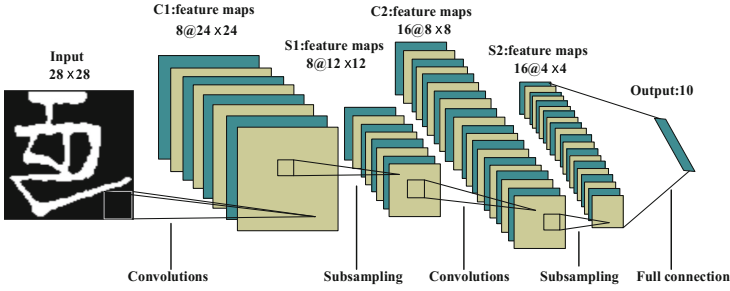


Fig. 3. CNN network structure

Convolution neural network is composed of the convolution layer and the sampling layer, and each layer is composed of multiple feature maps. Each pixel (neuron) of convolution layer is connected with a local area of the upper layer, and it can be viewed as a local feature detector. Each neuron can extract primary visual features such as direction line segments, angular point, etc. At the same time, this local connection makes the network have fewer parameters, which is beneficial to training. There is usually a sampling layer behind the convolution layer, in order to reduce the resolution of the image, and the network have a certain displacement, scaling and distortion invariance. For the convolution layer, the feature graph of the previous layer is conducted with a convolution operation with multiple group of convolution masks and then the feature graph of the layer is obtained through the activation function. The calculation form of the convolution layer is as follows:

$$a_j^l = \sigma \left(\sum_{i \in M_j} a_i^{l-1} * w_{ij}^l + b_j^l \right) \quad (1)$$

In Eq. (1), l is the number of layers where the convolution layer is; w is convolution kernel, which is a template of 5×5 . b is setover, and σ is activation function, that is $1/(1 + e^{-x})$. M_j represents an input feature graph of the upper layer.

The sampling layer is to sample the characteristics of the upper convolution layer and get the same number of feature graphs. The training of convolution neural network is the same as that of traditional neural network, and it adopts stochastic gradient descent. The input layer is a character image of Tibetan ancient books, whose size is 28×28 . C1 layer is the first convolution layer, which has eight feature graphs of 24×24 , and one pixel (node or neuron) in each feature graph is interconnected with a region of 5×5 corresponding to the input layer. S1 layer is a lower sampling layer containing 8 feature graphs of 12×12 , and each node in the feature graph is interconnected with a region of 2×2 corresponding to the feature graph in the C1 layer. C2 is the second convolution layer with 16 feature graphs, and the size of each feature graph is 8×8 . The connection between S1 and C2 plays an important role in feature extraction. S2 is the second sampling layer with 16 feature graphs, and the size of each feature graph is 4×4 . The last layer is the output layer with 10 nodes, corresponding to the output category, and it has a full connection with S2 layer.

4 Experiment and Result Analysis

4.1 Experiment Data

In this paper, the experimental data is the two groups of similar characters under the baseline of Tibetan characters, and each group contains 10 Tibetan character categories. The first group is a set of similar characters formed by Tibetan vertical stacks, and it is composed of “ཨ”, “ཉ”, “མ”, “ལ”, “ཚ”, “ཛ”, “ཞ”, “ཟ”, “འ” and “ཡ”. It is represented by G1, and there are a total of 5215 experimental samples.

The second group is a set of similar characters which are composed of complete consonant characters, and it is composed of “ཀ”, “ཁ”, “ག”, “གྷ”, “པ”, “ཕ”, “བ”, “ཇ”, “མ”, “ཉ”, “ཏ”, “ཐ”, “ད”, “ཎ”, “ཏ”, “ཡ”. It is represented by G2, and there are a total of 24,700 experimental samples.

In order to compare the performance of CNN in the recognition of Tibetan similar characters, CNN is compared with Naive Bayes Discriminant classifier and support vector machine classifier. For Naive Bayes discriminant and SVM classification, first of all, gradient 8 direction features described in Sect. 2 are extracted to get 64 D feature vector of each sample, and then the feature vector is used to discriminate and classify. For CNN, the image of the Tibetan characters is directly compressed to the image with a resolution of 28×28 , so as to reduce the parameters of CNN, and thus improve the training speed of the network.

4.2 Experiment Process

In the network training process shown in Fig. 3, the error reverse transform and the gradient random descent method are adopted to update the parameter w and b .

$J(w, b)$ is used to express the error function, and the expression of updating parameters with the gradient descent method is as follows:

$$w := w - \alpha \frac{\partial J(w, b)}{\partial w} \quad (2)$$

$$b := b - \alpha \frac{\partial J(w, b)}{\partial b} \quad (3)$$

α is the descent rate control parameter, and the selection of α in the experiment is determined by adopting the test method. Finally, selecting $\alpha = 1.5$ as the descent rate parameter of the system.

In order to observe the influence of different α on recognition rate, first of all, other parameters are fixed, for example, the times of circuit training are 30, because smaller number of circuit training times can save the training time, but it is enough to reflect the impact of α on the recognition rate. Different α and corresponding identification error rate are shown in Table 1.

The value of α during the experimental process is conducted according to the order from top to bottom in Table 1. The error rate in Table 1 shows that the error rate is the smallest when $\alpha = 1.5$, and it is 0.2339.

Table 1. Different α and corresponding recognition error rate

α	Error rate
0.01	0.7440
0.25	0.7440
0.6	0.3706
0.9	0.2817
1.5	0.2339
2	0.2798
1.8	0.2716
1.6	0.2651
1.4	0.2679

4.3 Experimental Results and Analysis

The experiment adopts CNN network structure shown in Fig. 3 and uses 64 D gradient feature to conduct Naive Bayes and SVM classification. In this paper, G1 and G2 sets are conducted with K-fold cross validation ($K = 10$), namely, each similar set is evenly divided into 10 parts: T1, T2, T3..... T10. Each part is taken as a test set each time, and the other 9 parts are regarded as the training set. The error rate results of G1 and G2 sets are shown in Tables 2 and 3 respectively. The experimental results show that, compared with Naive Bayes and SVM recognition method, the method based on deep neural network has a lower error rate. The reason for the poor performance of SVM and Naive Bayes is that the identification information of similar Tibetan characters is lost in the process of feature extraction.

Table 2. A comparison of error rate of 10-fold cross-validation on G1 set

Classifier	NBC	SVM	CNN
Error rate of T1	0.1288	0.0250	0.0212
Error rate of T2	0.1288	0.0327	0.0192
Error rate of T3	0.1308	0.0308	0.0269
Error rate of T4	0.1288	0.0327	0.0173
Error rate of T5	0.1385	0.0423	0.0154
Error rate of T6	0.1115	0.0365	0.0231
Error rate of T7	0.1385	0.0212	0.0154
Error rate of T8	0.1212	0.0250	0.0154
Error rate of T9	0.1231	0.0346	0.0192
Error rate of T10	0.1654	0.0404	0.0231
Average error rate	0.1315	0.0321	0.0196

The experimental results show that, compared with Naive Bayes and SVM recognition method, the method based on deep neural network has a lower error rate.

Table 3. A comparison of error rate of 10-fold cross-validation on G2 set

Classifier	NBC	SVM	CNN
Error rate of T1	0.0526	0.0158	0.0117
Error rate of T2	0.0530	0.0154	0.0134
Error rate of T3	0.0453	0.0109	0.0097
Error rate of T4	0.0555	0.0170	0.0134
Error rate of T5	0.0951	0.0146	0.0121
Error rate of T6	0.0632	0.0166	0.0162
Error rate of T7	0.0567	0.0142	0.0105
Error rate of T8	0.0551	0.0153	0.0109
Error rate of T9	0.0579	0.0117	0.0130
Error rate of T10	0.0587	0.0178	0.0117
Average error rate	0.0593	0.0149	0.0123

The reason for the poor performance of SVM and Naive Bayes is that the identification information of similar Tibetan characters is lost in the process of feature extraction.

In order to illustrate the recognition performance of this paper method, The average error rate comparison of different classifiers on G1 and G2 sets is shown in Fig. 4.

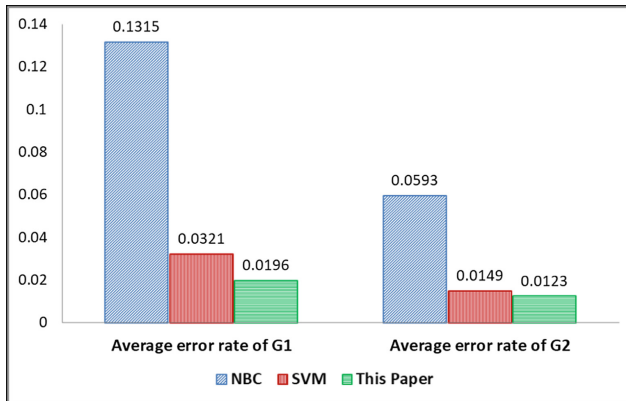
**Fig. 4.** Error rate of different classifiers on G1 and G2 set

Figure 4 shows this paper's method does not need human intervention in the process of training and recognition, is a kind of end-to-end approach, as well as under the condition of less training samples to achieve ideal effect.

Figures 5 and 6 shows the error curve of T10 of G1 and T10 of G2. It can be seen that CNN has smaller error in similar character recognition with the increase of the iterations.

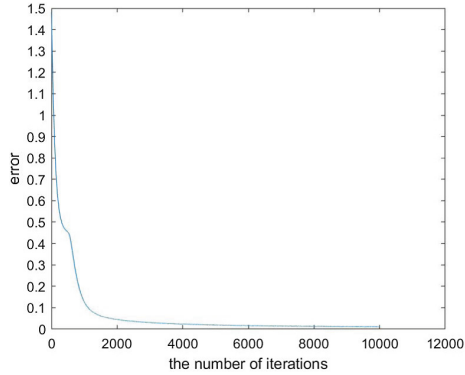


Fig. 5. T10 of G1 error curve

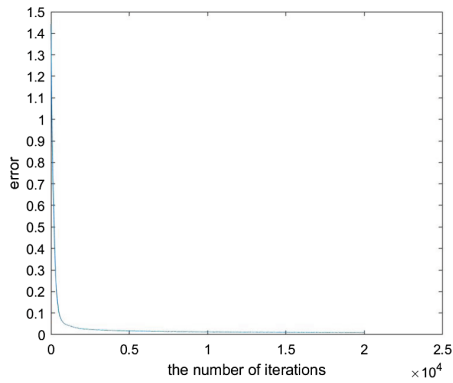


Fig. 6. T10 of G2 error curve

To further the robustness and stability of the network, In this paper randomly selects 1/10 of the sample from category of G2 set to form the test sample set (Te), and the number of test set is 2,470. In addition, it randomly selects five training sample sets (Tr1, Tr2, Tr3, Tr4 and Tr5) which doesn't include the test sample of Te, and the size are 1, 2 times, 3 times, 5 times and 9 times of test sample respectively, and The number of training sample sets is 2470, 4940, 7410, 12350, and 22230. The recognition error rate of these five sets of data is shown in Table 4.

Table 4. A comparison of error rate of different training samples in G2 set

Classifier	NBC	SVM	CNN
Error rate of Tr1-Te (2470-2470)	0.0628	0.0190	0.0510
Error rate of Tr2-Te (4940-2470)	0.0789	0.0202	0.0255
Error rate of Tr3-Te (7410-2470)	0.0846	0.0182	0.0227
Error rate of Tr4-Te (12350-2470)	0.0494	0.0153	0.0166
Error rate of Tr5-Te (22230-2470)	0.0526	0.0158	0.0146

Table 4 shows that with the increase of the sample size, the error rate of the recognition method based on the deep neural network gradually decreases, but the error rate of NBC and SVM method fluctuates up and down. It's clear that the network is more stable for the different sample collection, and the system has more robust robustness.

5 Conclusion

This thesis proposes that using convolution neural network to automatically learn and recognize the characteristics of similar characters of Uchen Script in Tibetan ancient books. At the same time, the similar characters of Tibetan ancient books constructed in this paper are adopted to train the model parameters, and the experimental results show that, compared with the traditional methods: (1) Deep convolution neural network can automatically learn the effective features and identify them from the pixel level, which avoids losing details caused by artificial selection and extraction of features and improves the recognition rate; (2) With the increase of the number of training samples, deep convolution neural network has a remarkable performance in reducing the error recognition rate, and the increase of training samples has an obvious effect on enhancing the recognition rate of deep neural network.

Acknowledgment. This work was supported by the National Science Foundation (No. 61772430), Program for Leading Talent of State Ethnic Affairs Commission, the Fundamental Research Funds for the Central University of Northwest Minzu University (No. 31920170142), and also supported by the Gansu Provincial first-class discipline program of Northwest Minzu University.

References

1. Gaur, A., Yadav, S.: Handwritten Hindi character recognition using k-means clustering and SVM. In: International Symposium on Emerging Trends and Technologies in Libraries and Information Services, pp. 65–70. IEEE (2015)
2. Sharma, A., Kumar, R., Sharma, R.K.: HMM based online handwritten Gurmukhi character recognition. *Mach. Graph. Vis.* **19**(4), 439–449 (2010)
3. Sudholt, S., Fink, G.A.: PHOCNet: a deep convolutional neural network for word spotting in handwritten documents. In: 15th ICFHR, pp. 277–282 (2016)
4. Moysset, B., Louradour, J., Kermorvant, C., Wolf, C.: Learning text-line localization with shared and local regression neural networks. In: 15th ICFHR, pp. 1–6 (2016)
5. Krishnan, P., Dutta, K., Jawahar, C.V.: Deep feature embedding for accurate recognition and retrieval of handwritten text. In: 15th ICFHR, pp. 289–294 (2016)
6. Sun, Z., Jin, L., Xie, Z., Feng, Z., Zhang, S.: Convolutional multi-directional recurrent network for offline handwritten text recognition. In: 15th ICFHR, pp. 240–245 (2016)
7. Wang, W., Ding, X., Chen, L., Wang, H.: Research on modern Tibetan language recognition in print. *Comput. Eng.* **29**(3), 37–39 (2003)
8. Pan, W.S., Jin, L.W., Feng, Z.Y.: Recognition of Chinese characters based on multiscale gradient and deep neural network. *J. Beijing Univ. Aeronaut. Astronaut.* **41**(4), 751–756 (2015)

9. Chen, K., Seuret, M., Wei, H., Liwicki, M., Hennebert, J., et al.: Ground truth model, tool, and dataset for layout analysis of historical documents. In: Proceedings of SPIE-IS&T, vol. 9402 940204-2. <http://proceedings.spiedigitallibrary.org>. Accessed 19 May 2015
10. Wei, H., Chen, K., Ingold, R., et al.: Hybrid feature selection for historical document layout analysis. In: 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 87–92. IEEE (2015)
11. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. *Int. J. Doc. Anal. Recognit.* **9**(2), 123–138 (2007)
12. Kesiman, M.W.A., Valy, D., Burie, J.C., Paulus, E., Sunarya, I.M.G.: Southeast Asian palm leaf manuscript images: a review of handwritten text line segmentation methods and new challenges. *J. Electron. Imaging* **26**(1), 1–15 (2017)
13. Xiao, X., Yang, Y., Ahmad, T., Jin, L., Chang, T.: Design of a very compact CNN classifier for online handwritten Chinese character recognition using DropWeight and global pooling. In: ICDAR (2017)
14. Le, A.D., Nakagawa, M.: Training an end-to-end system for handwritten mathematical expression recognition by generated patterns. In: ICDAR (2017)
15. Wu, Y.-C., Yin, F., Chen, Z., Liu, C.-L.: Handwritten Chinese text recognition using separable multi-dimensional recurrent neural network. In: ICDAR (2017)
16. LeCun, Y., Boser, B., Denker, J.S., et al.: Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems, Denver, United States, pp. 396–404 (1990)
17. Netzer, Y., Wang, T., Coates, A., et al.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning, Granada, Spain (2011)
18. Sermanet, P., Chintala, S., LeCun, Y.: Convolutional neural networks applied to house numbers digit classification. In: Proceedings of IEEE International Conference on Pattern Recognition, Tsukuba, Japan, pp. 3288–3291 (2012)
19. Coates, A., Carpenter, B., Case, C., et al.: Text detection and character recognition in scene images with unsupervised feature learning. In: Proceedings of IEEE International Conference on Document Analysis and Recognition, Beijing, China, pp. 440–445 (2011)
20. Wang, T., Wu, D.J., Coates, A., et al.: End-to-end text recognition with convolutional neural networks. In: Proceedings of IEEE International Conference on Pattern Recognition, Tsukuba, Japan, pp. 3304–3308 (2012)
21. Jin, L., Zhong, Z., Yang, Z., et al.: Application of deep learning in handwritten Chinese character recognition. *J. Automat.* **42**(8), 1125–1141 (2016)
22. Liu, C.L.: Normalization-cooperated gradient feature extraction for handwritten character recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(8), 1465–1469 (2007)
23. Zhao, Y., Tao, D., Zhang, S., et al.: Similar Chinese character recognition based on deep neural network under big data. *J. Commun.* **321**(9), 184–189 (2014)
24. Bai, Z.L., Huo, Q.: A study on the use of 8-directional features for online handwritten Chinese character recognition. In: Proceedings of the 8th International Conference on Document Analysis and Recognition, pp. 262–266. IEEE, Seoul (2005)



Research on the Method of Tibetan Recognition Based on Component Location Information

Yuehui Han^{1,2}, Weilan Wang^{1(✉)}, Yiqun Wang¹,
and Xiaojuan Wang¹

¹ Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730000, Gansu, China
wangweilan@xbmu.edu.cn

² College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730000, Gansu, China

Abstract. The recognition of Tibetan is of great significance to the study of Tibetan culture while the progress of Tibetan character recognition is lagging behind. Especially when there are not a large number of available training samples, Tibetan character recognition is very difficult. So we propose a recognition method for Tibetan characters based on component location information without a large number of training samples. The proposed method includes three main parts: (1) The segmentation of character and the extraction of component which contain location information in the character; (2) Features extraction and classifier design; (3) The superposition of component after recognition and the retrieval of character. The testing results are: the recognition rate of single component is 98.4%, the recognition rate of multilevel component is 97.2%. It indicates that the method has a good effect on the recognition of Tibetan character, and it is helpful for the recognition of Tibetan documents.

Keywords: Tibetan recognition · Character segment
Component combination · Classifier design

1 Introduction

Tibetan is a minority nationality character which is used by 5 million Tibetan people in China. There are two views on the origin of Tibetan character: One view is that the Tibetan was created by a minister Tumi Sabza of Srongtsen Gampo's in the seventh Century. Another view is that the Tibetan was evolved from Zhang zhung character. Tibetan is a special kind of phonetic character, whose longitudinal unit is a character, and a character consists of at most 4 components. Syllables are the basic spelling units. Each syllable consists of at most 4 characters, as shown in Fig. 1.

Compared with other languages, the progress of Tibetan recognition research is relatively backward. However, the gap is gradually narrowing under the efforts of a lot of scholars.

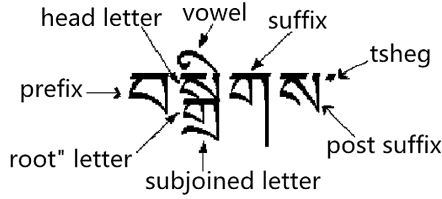


Fig. 1. Example of Tibetan structure

In Printed Tibetan: Hua Wang carried on the preliminary study of Tibetan recognition from the preprocessing, text line segmentation, feature selection and classifier design [1]. By using the segmentation method based on the connected domain and the extraction of the stroke feature based on the grid, Zhu Ou increased the recognition rate of the Tibetan [2]. In order to improve the recognition rate, Yulei Wang extracted the features of Tibetan characters based on Fractal Moments and improved rough mesh method [3]. Yuzhen Baima proposed projection method based on network lattice which is suitable for Tibetan recognition [4]. Wei Zhou proposed a Tibetan recognition method based on geometry analysis of component [5]. In Handwritten Tibetan: Heming Huang established the first off-line handwritten Tibetan recognition system [6]. Xiaojuan Cai proposed a feature extraction algorithm for off-line handwritten Tibetan characters based on multi projection normalization, which further improved the recognition rate [7]. By using HMM based on stroke type and the position relation between strokes to improve the recognition performance [8], Weilan Wang designed a complete online handwritten Tibetan recognition system [9], proposed a Tibetan Sanskrit handwritten sample generation method based on component combination [10]. Longlong Ma proposed a semi-automatic component annotation method for online handwritten Tibetan character database [11], a Tibetan component representation learning method for component-based online handwritten Tibetan character recognition [12], and a component segmentation-based recognition method for online handwritten Tibetan syllables [13]. We propose a recognition method for Tibetan characters based on component location information without a large number of training samples. The rest of this paper is organized as follows.

Section 2 introduces printed Tibetan characters and components. Section 3 gives the component segmentation method. The method of feature extraction and classifier design is given in Sect. 4. Section 5 gives recognition process and result analysis. Section 6 offers concluding remarks.

2 Tibetan Characters and Tibetan Components

Tibetan is a special kind of alphabetic writing that a character contains 1 to 4 components which are superposed up and down. Most Tibetan recognition work is based on characters, while the recognition work based on components is rarely. There are 534 printed Tibetan characters used frequently, while 231 components in totally. And the 231 component contains 51 single components, 180 deformation combination

components. As for non-single that changes have taken place in the deformation combination, so we take combination components as a whole, as shown in Table 1 and Fig. 2. In fact based on components is a very useful method for Tibetan recognition work especially when the training sample is insufficient. Tibetan characters have strict distribution rules, which can help separation component easily. Based on component can also help reduce the number of classification. Character is recognized by retrieving Tibetan characters database after the components are recognized.

Table 1. Example of Tibetan characters database.

ID	Tibet	TibetOrder	Sort	Code
144	མཚ	82	1	41
145	མཚ	82	2	3
146	མཚ	82	3	161

Table 1 is a character example in Tibetan characters database, “TibetOrder” is the sequence number of the character in database, “Tibet” is a character, “ID” is the database record number, “Sort” is the layer information of a component in a character, and “Code” is the sequence number of component in the template. Figure 2 is all Tibetan components which contain 51 single components and 180 deformation combination components.

We proposed a recognition method for Tibetan characters based on components location information. The stages of the proposed method are shown as follow.

- (1) After the size transformation, the segmentation of the above vowel, the segmentation of the below vowel and the segmentation of intermediate component, the component containing location information are obtained.
- (2) Feature extraction and classifier design.
- (3) Calculate the matching degree using the Euclidean distance, screen out the top-ten matching degree and the corresponding components.
- (4) According to the recognition result of each component, retrieve and find out the corresponding character in database.

3 Component Segmentation

Component segmentation based on the writing standard of Tibetan character, which follow the sequence of above vowel, below vowel and intermediate component. The component segmentation process is shown in Fig. 1.

In Fig. 3, “Above” indicates above vowel, “Below” represent below vowel and “Single” indicates single intermediate component, “Double” refers to double intermediate component.

༥	༦	༧	༨	༩	༩	༱	༲	༳	༴	༵	༶	༷	༸	༹	༺	༻	༼	༽	༾	༿	
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿
༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿	༿

Fig. 2. All Tibetan components we used.

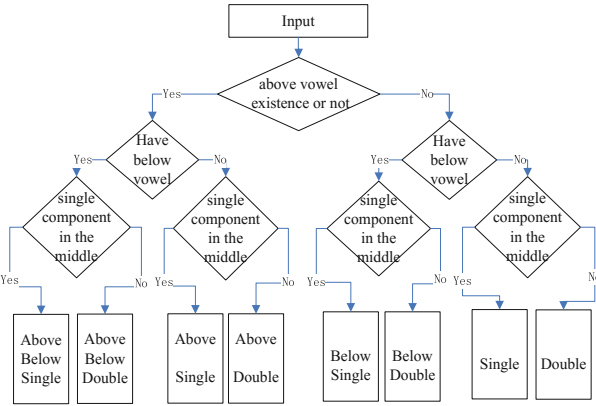


Fig. 3. Component segmentation process

3.1 Above Vowel Segmentation

The above vowel is located on the baseline of the Tibetan character, the top 1/4 part of an image. The above of the baseline is empty without above vowel. The specific algorithms are as follows.

Step 1: Above vowel judgment.

The statistical number of handwriting points in the 1/5 section above the image, and the numbers is replaced by “sup”. Column projection on the 1/5 section above the image, Statistical the numbers that Greater than zero, and the numbers is replaced by “tnum”. The method of judgment is shown in Fig. 4, “Cnum” represent the numbers of columns. Experimental verification, when T is 5, there is the best result.

Step 2: Find the segmentation point.

Image line projection, and the point near “Rnum/4”, which has minimum projection value and has the maximum rate is the segmentation point. “Rownum” indicate the numbers of lines.

Step 3: Above vowel segmentation.

Image segmentation based on segmentation point. Example of above vowel segmentation is shown in Fig. 5.

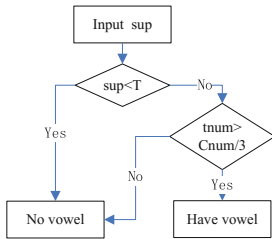


Fig. 4. Above vowel judgment

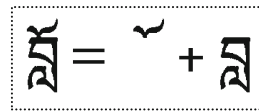


Fig. 5. Example of above vowel segmentation

3.2 Below Vowel Segmentation

The below vowel is located in the underneath, 1/4 part of image. The specific algorithms are as follows.

Step 1: Below vowel judgment.

The statistical number of handwriting points in the bottom 1/5 section of the image is replaced by *sdown*. Column projection on the 1/5 section bottom the image, Statistical the numbers that Greater than zero, and the numbers is replaced by “dnum”. The method of judgment is shown in Fig. 6.

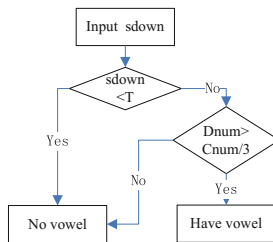


Fig. 6. Below vowel judgment

The numbers of “Cnum” indicate the number of columns.

Step 2: Find the segmentation point.

Projection in the right half of the image, and the point near “4*Rnum/5”, the segmentation point is supposed to have minimum projection value. “Rnum” indicate the numbers of lines.

Step 3: Below vowel segmentation.

Starting from the right side of the image, if connected to a below vowel, disconnect based on the segmentation point. If not connected, search the segmentation path along the contour of below vowel. Figure 7 is the Example of below vowel segmentation.

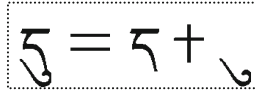


Fig. 7. Example of below vowel segmentation

Intermediate component Segmentation

There are only one or two layers of components in middle part, after above vowel segmentation and below vowel segmentation. The specific algorithms are as follows.

Step 1: Judgment of the number of layers.

After removing the above vowel and below vowel, assume the number of handwriting points in the top half of the image is N , in the bottom half of the image is M . Single component if $M/N < T_3$, the middle part is called single component, and it is called double component under the condition of $M/N > T_3$. The experiment proves that the result is best when M is 0.9.

Step 2: Find the segmentation point.

Projection the image, and the point near the middle position of the image, which has minimum projection value is the segmentation point.

Step 3: Intermediate component segmentation.

Image segmentation is based on segmentation point. Example of intermediate component segmentation is shown in Fig. 8.

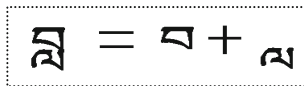


Fig. 8. Intermediate component segmentation

3.3 Special Circumstances Process

- (1) Sometimes the segmentation of above vowels may makes mistakes, as is shown in Fig. 9. In this case we can use the minimum rectangle to extract the correct top component. As is shown in Fig. 10.
- (2) Sometimes the deformation combination of some components will be considered as a single component, which is shown in Fig. 11. So we consider the result of the deformed combination as a component and increase the number of components in the template.

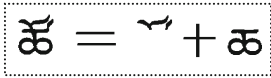


Fig. 9. Error segmentation example

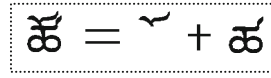


Fig. 10. Correct segmentation example

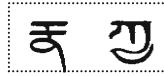


Fig. 11. Component deformation combination

4 Feature Extraction and Classifier Design

4.1 Component Feature Extraction

168 features are extracted altogether, and the images involved are original component image, remove position information image, skeleton image and edge image. As is shown in Fig. 12(a)–(d). All image normalization, 100 rows and 50 columns.

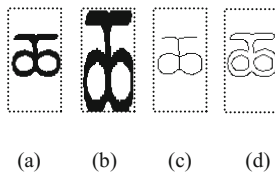


Fig. 12. (a) (b) (c) (d) Image used to extract feature

The feature extraction algorithm of the component is as follows.

Step 1: Feature extraction of original image

The original image refers to the component image come form template or character segmentation. The original image contains the location information of the component distribution. And the distribution information of different components is different. Four features are extracted from the original image: The ratio of black pixel points, the number of rows with black pixel points, the position of first and the last row with black pixel points.

Step 2: Feature extraction of remove position information image

After minimum rectangle frame processing, image extends to the original size. And the image is divided into 16 parts using an elastic grid. 23 features are extracted from the remove position information image: The ratio of black pixel points, position of grid line and the position of first black pixel point per line in each part.

Step 3: Feature extraction of skeleton image

After skeleton processing of the original image, we get the skeleton image. 41 features are extracted from the skeleton image: rough periphery and inner profile.

Step 4: Feature extraction of edge image

After edge processing of the original image, we get the edge image. And the image is divided into 25 parts averagely. Statistical directional line information in each part and 100 features are extracted.

4.2 Classifier Design

Euclidean distance is used to calculate the matching degree between the test components and the components in the template. D_i indicate the matching degree between the test components and the i -th components in the template. And the range of number “ i ” is 1 to 231. As shown in (1).

$$D_i = \sum_{j=1}^m (x_j - x_{i,j})^2 \tag{1}$$

Where m indicate the total number of feature values, x_j and $x_{i,j}$ represents the j -th feature value of test component and the j -th feature value of i -th components in the template.

5 Analysis of Experimental Results

The method is carried out after line and character segmentation. Figure 13 is a part of the Tibetan document image. Figure 14 is line segmentation results. Figure 15 is the recognition results of Fig. 14(a), and the results are “གདན་ས་བཟ་ཤིས་ལྷན་པོ།”. It can be seen from the recognition example that our method has a satisfactory recognition result. For the experiments 100 Tibetan printed document images are used, and the recognition rate of single component is 98.4%, the recognition rate of multilevel component is 97.2%.

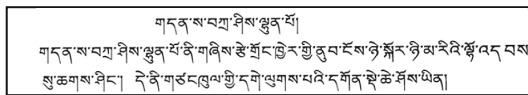
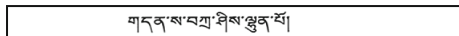
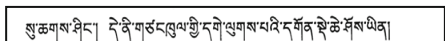
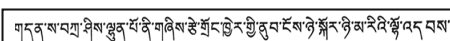


Fig. 13. Tibetan printed document example



(a)



(b)

(c)

Fig. 14. (a) (b) (c) line segmentation results

Character	Component segmentation result	Recognition results Top1 ->Top10
ག	ག	ག ག ག ག ག ག ག ག ག ག
ང	ང	ང ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
ཉ	ཉ	ཉ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
.	.	. ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
མ	མ	མ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
.	.	. ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
འ	འ	འ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
ཨ	ཨ	ཨ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
.	.	. ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
ཀ	ྱ ག	ཀ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
མ	མ	མ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
.	.	. ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
ཞ	ཞ ཏ	ཞ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
ཉ	ཉ	ཉ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
.	.	. ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
ཏ	ཏ འ	ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ
		ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ ཏ

Fig. 15. Recognition example

After analysis, we can see that there are two main reasons for wrong recognition. One is caused by the line or character segmentation error, which may cause some character information to lose or increasing noise. And another is caused by components segmentation error, which segmentation points are judged mistakenly. Wrong segmentation point cause wrong segmentation results and lead to wrong recognition results certainly, which is the reason why multi-layer character recognition rate is lower than single-layer character.

ཀློག་གཞི་གསལ་པར་མཚན་ལྡན་ལྟར་བཟང་པོར་བཞག་ནེ་ན་མཚན་ལྟགས་ལྡོམ་ལ་མཚན་པར་……
ཐྱེད་གསུམ་མ་མུལ་གྱི་རོང་རལ་ལྱང་བུའི་མཉམ་སྦྲེག་ལ་མཚན་པར་མཉམ་སྦྲེག་སྒྲིབ་ལ་མཚན་པར་……
མཉམ་བཟང་དུ་ཉིན་གྱི་ལོ་དུང་ལྔ་ལྟེང་མཉམ་སྦྲེག་སྒྲིབ་ལ་མཚན་པར་མཉམ་སྦྲེག་བཟང་པོར་བཞག་ནེ་པོ་……
དཔལ་གནོད་སྤང་ལོ་མཚན་སྦྲེབ་པོ་ལྟེང་མཉམ་སྦྲེག་ལ་མཚན་པར་མཉམ་སྦྲེག་བཟང་པོར་བཞག་ནེ་པོ་……
གསུམ་ལ་མཚན་པར་བཟང་པོར་བཞག་……

Fig. 16. Black body

ཀློག་གཞི་གསལ་པར་མཚན་ལྡན་ལྟར་བཟང་པོར་བཞག་ནེ་ན་མཚན་ལྟགས་ལྡོམ་ལ་མཚན་པར་……
ཐྱེད་གསུམ་མ་མུལ་གྱི་རོང་རལ་ལྱང་བུའི་མཉམ་སྦྲེག་ལ་མཚན་པར་མཉམ་སྦྲེག་སྒྲིབ་ལ་མཚན་པར་……
མཉམ་བཟང་དུ་ཉིན་གྱི་ལོ་དུང་ལྟེང་མཉམ་སྦྲེག་སྒྲིབ་ལ་མཚན་པར་མཉམ་སྦྲེག་བཟང་པོར་བཞག་ནེ་པོ་……
དཔལ་གནོད་སྤང་ལོ་མཚན་སྦྲེབ་པོ་ལྟེང་མཉམ་སྦྲེག་ལ་མཚན་པར་མཉམ་སྦྲེག་བཟང་པོར་བཞག་ནེ་པོ་……
གསུམ་ལ་མཚན་པར་བཟང་པོར་བཞག་……

Fig. 17. Long body

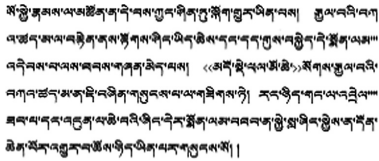


Fig. 18. Round body

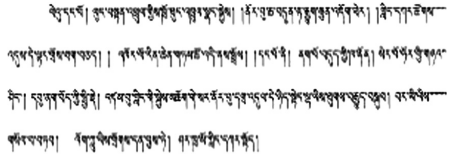


Fig. 19. Bamboo body

We also tested four other Tibetan fonts that include Black body (see Fig. 16), Long body (see Fig. 17), Round body (see Fig. 18) and Bamboo body (see Fig. 19), which recognition rate is 96.3%, 92.1%, 95.8% and 93.3% in the 50 sets of test samples. From the test results, we can see that the recognition effect of Black body and Round body is better than Long body and bamboo body. This is because that the change of Long body and Bamboo body is larger than that of Black body and Round body compared with the commonly used Tibetan fonts, which is the template we use. So it is easy to make mistakes when components are segmented, which lead to the component contain noise or some information lost. And then the result of the character recognition is wrong. Although these Tibetan fonts are slightly different from the commonly used Tibetan fonts, but the recognition rate has not been greatly affected. This also can prove that the characteristics extracted are effective.

6 Conclusions

This paper propose a recognition method for Tibetan characters based on component location information when lack a large number of training samples. The main work includes: the extraction of component which contain location information, features extraction based on four kinds of images, classifier training, superposition of component and the retrieval of character based on component location information database. The single-layer character recognition rate for this method is 98.4%, and 97.2% for multi-layer character. It is found that the effect of component segmentation directly affects the recognition of character. So the optimization of component segmentation algorithm is the focus of further research.

Acknowledgements. This work was supported by the National Science Foundation (No. 61772430), the Program for Leading Talent of State Ethnic Affairs Commission, the Fundamental Research Funds for the Central University of Northwest Minzu University (No. 31920170142), and also supported by the Gansu Provincial first-class discipline program of Northwest Minzu University.

References

1. Wang, H., Ding, X.Q.: Multi-font printing Tibetan character recognition. *J. Chin. Inf. Process.* **17**(6), 47–52 (2003)
2. Drup, N., Ren, P., Sanglangjie, D.: Study on printed Tibetan character recognition. *Comput. Eng. Appl.* **48**(1), 55–62 (2009)
3. Li, Y.Z., Wang, Y.L., Liu, Z.Z.: Study on printed Tibetan character recognition technology. *J. Nanjing Univ.* **48**(1), 55–62 (2012)
4. Baima, Y.Z.: Research on feature extraction of Tibetan characters. *Comput. Knowl. Technol.* **28**(1), 6362–6364 (2013)
5. Zhou, W., Chen, L.: Tibetan recognition based on geometric shape analysis. *Comput. Eng. Appl.* **48**(18), 201–205 (2012)
6. Huang, H.M.: Research on recognition of off-line handwritten Tibetan character, pp. 19–34. Southeast University (2014)
7. Cai, X.J., Huang, H.M.: Feature extraction of offline handwritten Tibetan characters based on multiple projections. *Comput. Technol. Dev.* **26**(3), 93–96 (2016)
8. Liang, B., Wang, W.L., Qian, J.J.: Application of hidden Markov model in on-line recognition of handwritten Tibetan characters. *Microelectron. Comput.* **26**(4), 98–100 (2009)
9. Research on online handwritten Tibetan recognition input: W.L. Wang. *Sci. Technol. Achiev. China* **11**, 36–38 (2012)
10. Wang, W.L., Lu, X.B., Cai, Z.Q.: Online handwritten sample generated based on component combination for Tibetan-Sanskrit. *J. Chin. Inf. Process.* **31**(5), 64–73 (2017)
11. Ma, L.L., Wu, L.: Semi-automatic Tibetan component annotation from online handwritten Tibetan character database by optimizing segmentation hypotheses. In: 12th International Conference on Document Analysis and Recognition, pp. 1340–1344 (2013)
12. Ma, L.L., Wu, J.: A Tibetan component representation learning method for online handwritten Tibetan character recognition. In: 14th International Conference on Frontiers in Handwriting Recognition, pp. 317–322 (2014)
13. Ma, L.L., Wu, J.: Online handwritten Tibetan syllable recognition based on component segmentation method. In: 13th International Conference on Document Analysis and Recognition. pp. 46–50 (2015)



Research on Text Line Segmentation of Historical Tibetan Documents Based on the Connected Component Analysis

Yiqun Wang¹, Weilan Wang^{1(✉)}, Zhenjiang Li¹, Yuehui Han^{1,2},
and Xiaojuan Wang¹

¹ Key Laboratory of China's Ethnic Languages and Information Technology
of Ministry of Education, Northwest Minzu University,
Lanzhou 730000, Gansu, China

wangweilan@xbmu.edu.cn

² College of Mathematics and Computer Science, Northwest Minzu University,
Lanzhou 730000, Gansu, China

Abstract. Text line segmentation is one of the critical content in handwriting documents recognition especially in the historical documents' analysis and recognition. Because of the low quality and the complexity of these documents (background noise, scattered character, touching components between consecutive lines), automatic text line segmentation remains to be a hot spot for researching. In this paper we propose a new method to segment the text line from the historical Tibetan scripture "kangjur" of the Beijing version on the paper by means of woodcut. This method first performs document image skew detection and correction, using projection profiles to get the baseline of text line, then the connected component is allocated to text line according to the location relationship. For some connected components, analyzing their location and sharp to assign these connected components correctly. This method using connected component instead of pixels, avoiding the noise generated by splitting characters. Experiments show that this method is effective in copes with touching text lines and promising in text line segmentation from historical Tibetan document.

Keywords: Historical Tibetan document · Kangjur · Text line segmentation
Component analysis · Location · Sharp

1 Introduction

The Tibetans have a large number of historical documents; most of them are stored in temples. Those historical documents are exist in the form of scriptures for a very long time. It is urgent to protect and reuse them by using digital technology because of the deterioration of the quality of the historical documents. Using Optical Character Recognition (OCR) technology to converts the historical Tibetan documents into text files. The text files stored in the services is not only appropriate preserved but also convenient for reusing those precious historical documents. In document processing field, the segmentation is essential for document recognition which it needs several steps of binarization, layout analysis, text blocks extraction, text lines and words segmentation

and character recognition. The degraded historical documents (e.g. ink stains, torn pages, overlapped/touching character, broken stroke etc.) make a challenge for the text line segmentation task. The variation of the interline distance and the baselines undulation between lines or even along the same text line. The touching characters between adjacent text lines appear frequently in Tibetan documents. The whole characters may be divided into several parts because of broken stroke. All above greatly complicates the task of the text lines segmentation from historical handwritten document.

In this paper, we focus on the extraction of text lines from historical Tibetan documents and we propose a method based on the analysis of the location and shape of the connected component. This method cannot totally solve the problem of segmentation, but we try to reduce the error as much as possible to extract text line complete. For text line extraction of historical Tibetan documents, a few researches have been done such as: based on baseline detection method [1] and contour curve tracking method [2]. Other common text line extraction methods also include: projection-based method [3], Hough-transform [4], smearing method [5], clustering approach [6, 7].

In [1] the baseline is getting by template matching, pruning the salient strokes and closing operation, then touching characters is detecting and splitting, the text-line is extracted according to baseline and split position, this method can deal with the touching characters and fluctuating text lines. However, this method does not consider broken strokes, so it is inadequate for some historical document image with a large number of broken strokes.

In [2] the text line segmentation method based on contour tracking is proposed. The text line is extracted by the contour from the document image which comes from the constructed connected component. The method combine the barycentre coordinates of the connected component to form the curve line and the separated components are assigned to the corresponding text line by the barycentre gravity later. The text line is obtained by the contour curve of the text line. This method is innovative but the performance is not satisfactory when a document image with many touching characters is segmented.

Projection-based method [3] is most commonly used for the text line segmentation especially in printed or slightly document. The projection value is computed by summing the values of pixel in the foreground in horizontal axis of each line. The text lines is segmented by straight lines with suitable positions and directions, this method is not suitable for historical Tibetan document as there is no obvious line gap. According to the layout of the Tibetan Scripture “Kangjur”, the direction of the text lines is approximately horizontal parallel, so this method can be used to find the baseline of the text lines.

Hough-based method [4] is proper to detect text lines which are usually parallel in certain areas. Smearing method [5] enlarged area of black pixels, the white space between the black pixels is filled with black pixels if their distance is within a pre-defined threshold. But this method is not suitable for historical Tibetan document. Because some vertical stroke is overlong that smearing horizontal will produce more touching components.

Clustering method [6, 7] usually divides a picture into several connected components, blocks or other units according to some features, and then aggregates these units to form alignments according to some rules. Considering that there are a lot of touching

characters in historical Tibetan documents, it is very difficult to assign the characters to the correct text lines in this way. Thus, this method is not suitable for text line segmentation from historical Tibetan documents.

The paper is organized as follows: in Sect. 2, our method is described. In Sect. 3 the proposed method to segment text lines is detailed. Section 4 present the experimental results and discuss. Section 5 describes conclusions and future work.

2 Our Method

Tibetan character can be regarded as a kind of string composed of basic characters and characters in the vertical direction [8] (see Fig. 1). The authentic historical Tibetan document not only have lots of touching characters between adjacent lines as the height of the character is inconsistent but also have lots of broken strokes than other languages. The touching characters between adjacent line, the separated upper and lower vowels and the broken strokes make the text-line segmentation more complex (see Fig. 2). At present, there is no satisfactory segmentation method for the authentic historical Tibetan document of wooden printing.

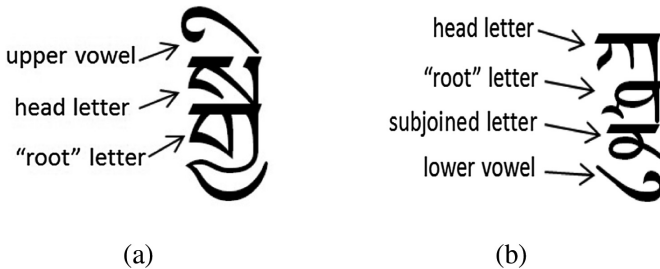


Fig. 1. (a) Character with upper vowel (b) character with lower vowel.

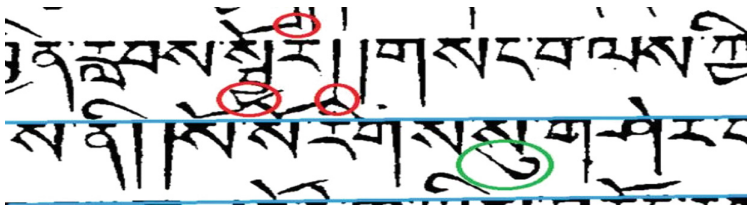


Fig. 2. Partial image with slanted baseline, separated character, overlong and touching characters.

We can see the characters in the historical Tibetan document are very close to each other because of the limited area of the document and there is no obvious gap between adjacent lines. The historical document images have large number of touching and overlapped characters and variety of broken strokes which are the main challenge to extract text lines accurately.

In order to extract the text-line completely from handwritten, degraded, historical Tibetan documents, we present a text-line segmentation method which combine the row projection location analysis and shape analysis of connection components. Our method stems from the idea that the text line is composed of a set of location related components. The task of text line segmentation is to find such a set of components and extract them from the document image to form a text line.

Our method detects the input document image whether the image is skew or not and perform skew correction if it is. Then the position of baseline is obtained using projection method as the text line is approximately horizontal after skew correction.

The connected component is allocated to text line according to the location relationship between the component and segmentation line by their location information. For some connected components, it is difficult to assign them to the corresponding text line only depends on location. Generally speaking, these components are broken strokes, separate vowels, symbols, touch characters, noise, and so on. Therefore, it is necessary to make a further analysis of the location and shape of these connected components in order to correctly determine their attributes. Combining location and shape information to determine which text line these connected components should belong to will be more accurate, especially for complex documents. At last, the components belong to the same alignment are merged to recover the text line.

Here is the architecture we extract text-lines from Tibetan historical documents shown in Fig. 3.



Fig. 3. The text line segmentation process.

Our method includes four stages:

1. Pre-processing: We detect whether the input image is skew. If the image is skewed, the skew correction is done to make the text lines in the image horizontally parallel. Then, the information about height of character is got which will be used to estimate the feature of characters in next stage. At last the position of baseline is detected using the projection method.
2. Location analysis: According to the baseline position we obtained before, the text line region is extracted from the input image as a rectangle, and divide the region into upper part and lower part according to the baseline position of the current baseline. The upper part is undoubtedly part of the current text line, but the lower part contains some components of the next text line. Next, the projection method is used to find the optimal segmentation line (SL) which is the row's location with minimum pixels in the lower part. Then the connected component in the lower part is divided into three classes according to whether it intersects with the SL. Some connected components are belongs to current text line or next text line certainly but

the others cannot easily determine which text line they belong to, so further analysis is needed.

3. Location and Shape analysis: By judging whether there is intersection point with SL, the connected components with uncertain attribution is divided into one class. By analyzing the location information and shape information of the connected component in this class, we classify it into the correct text lines, especially for the touching characters between the text lines, we use some features and rules to detect and separate them.
4. Image merging: Through the Location and Shape analysis (LSA) of the connected components, the connected components belonging to the current text line have been marked out. Combining these connected components to form the lower part of the current text line, and then splicing the upper and lower parts to form a complete image of the current text line.

3 Text Line Segmentation

The proposed text line segmentation method base on the projection, location and shape analysis of connected components for historical Tibetan handwritten document deals with the following challenges: (i) parts of neighboring text lines may be connected; (ii) overlong character and touching character in text line; (iii) the separated vowel may be appeared either above or below the text line and (iv) the broken strokes of characters in text line. The work flow of the text line segmentation is shown in Fig. 4.

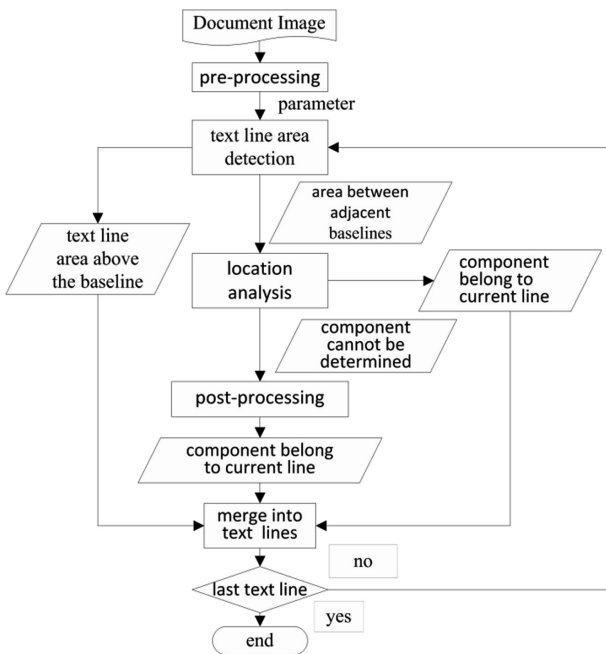


Fig. 4. Proposed method framework.

3.1 Pre-processing

The pre-processing stage consists of three steps. First, whether the input image is skew is detected, the document image is skew corrected if the image is skew. The angle of the skew correction is determined by the length of the border detection line, the method rotates the image from -2 angle to +2 angle at step 0.1, and detects the sum of the length of the edge lines of the four borders, the maximum sum corresponding angle is the correction angle. An example is shown in Fig. 5. Then, average character height (AH) and the average component height (ACH) for the whole document image are calculated and the bordering box is removed. Last, the baseline position of each text line is obtained by row projection profile method, and the number of locations equals the number of text lines. An example is shown in Fig. 6.

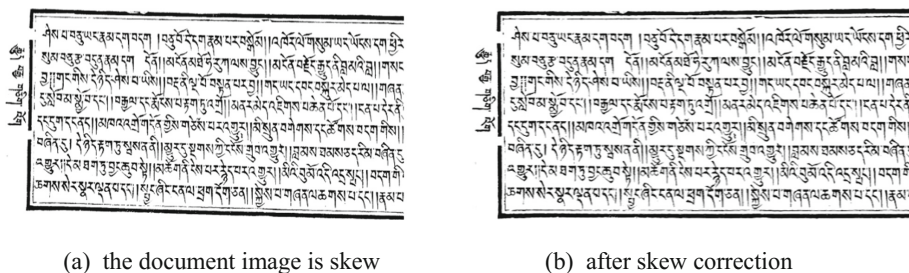


Fig. 5. The input document image is skew (a) and the document image after skew correction (b).

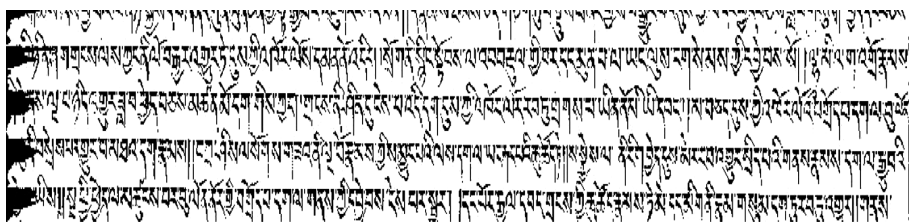


Fig. 6. Row projection diagram of binary document image.

3.2 Location Analysis

This stage includes two steps. At the first step, the projection method is used to get the initial row position of the text line that is the line of beginning (LB) then extract the area between LB and the baseline location of the current text line as the upper part of the current text line image, and this part is denoted as “upper image” (see Fig. 7. black part). The next step will analyze the image (“lower part”) between the baseline of current text line and the next baseline. Firstly, the statistical method is used to find the optimal segmentation line (SL) which is the row’s location with minimum pixels. Next,

by using relative location relations between components and SL, the connected components domain is divided into three sub-domains, which are denoted as “Subsetcur”, “Subsetlow” and “Subsets” respectively.

“Subsetcur” contains all components which totally are located above the SL (see Fig. 7. green part) and “Subsetlow” contains all components which are located below the SL (see Fig. 7. blue part). “Subsets” contains all the components which have the intersected points with the SL, this subset have various components that need to be analyzed in different manners by the proposed method in the next stage (see Fig. 7. red part).

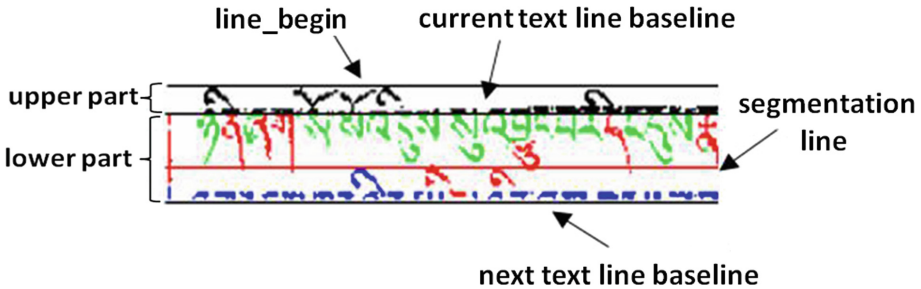


Fig. 7. An example of partitioning the connected components by the relationship between the component and the segmentation line. The black part is upper image, the lower part is the region between current baseline and next text line’ baseline, the green part means “subsetcur”, the red part means “subsets” and the blue part means “subset low”. (Color figure online)

3.3 Sharp-Analysis

This stage analyzes the location and shape of the components which is in the “subsets” to determine whether it belongs to the current text line or not. The categories of these components in the “subsets” are separated into upper vowels and lower vowels, broken strokes ,overlong characters, touching and overlapped characters, and bar shaped connected components. All connected components in “subsets” have a common property that they intersect with SL, in other words, SL divides these connected components into upper and lower parts. In order to assign connected components to the corresponding text lines accurately, we need to extract some features of these connected components, such as the height of connected components (H), the height above the SL (HA), the pixel per row for the part above the SL (PPRA), the height below the SL (HB), the pixel per row for the part below the SL (PPRB), and the ratio of the foreground area to the minimum rectangular bounding area (RFB).

The PPRA is calculated as follows: (value 1 for foreground and 0 for background pixels)

$$PPRA = \sum_{x=1}^{width} \sum_{y=1}^{HA} I(x, y) / HA \text{ if } I(x, y) = 1 \quad (1)$$

The PPRB is calculated as follows:

$$PPRB = \sum_{x=1}^{width} \sum_{y=1}^{HB} I(x,y)/HB \text{ if } I(x,y) = 1 \quad (2)$$

The RFB is defined as follows:

$$RFB = \sum_{x=1}^{width} \sum_{y=1}^{height} I(x,y)/width * height \text{ if } I(x,y) = 1 \quad (3)$$

The location and shape analysis (LSA) procedure consists of two steps. At the first step, the feature obtained above are used to determine whether the connected components lying in subsets are belong to the current text line or not according to the following conditions.

In the first step, the method take advantage of the feature we obtained above and the average character height (AH) and the average component height (ACH) which are got at first stage to classify them into three categories by rules. The first category have the connected components which are assigned to the next text line. One category consists of components that in this step cannot determine the attribution of text lines, and these components will be analyzed shapes in the next step. The last category includes the components of the current text line, usually consisting of overlong characters, symbols, and touching characters. The touching character will be segmented and retain the component belonging to the current text line.

The broken strokes and separated vowels were selected by conditions 4. The condition is described follow:

$$H < ACH \quad (4)$$

The connected component is belongs to the current text line, if some features satisfy the condition below:

$$(H > AH) \text{ and } (HA > HB) \quad (5)$$

Identify the connected component with height exceeds the height threshold which is defined as:

$$HT = 1.5 * AH \quad (6)$$

The connected components which satisfied the above conditions include the overlong characters (see Fig. 8, a b c), the touching characters(see Fig. 8, d e) and the bar-shaped connected components which generally are Tibetan character symbol(see Fig. 8, f).

The bar-shaped connected components usually are symbol which is belong to current text line. Such component will be selected if the following condition is satisfied:

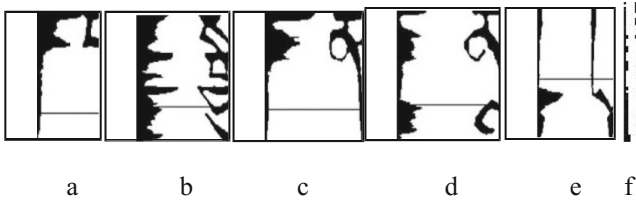


Fig. 8. The image of overlong character, the overlong characters (a b c), the touching characters (d e) and the bar-shaped connected components (f).

$$RFB > 0.5 \tag{7}$$

The touching character are as long as the overlong character (see the Fig. 6a b c and d e). Choose the touching character according to the following constraint.

$$PPRB > 1.2 * PPRA \tag{8}$$

The LSA first step work flow is shown in Fig. 9

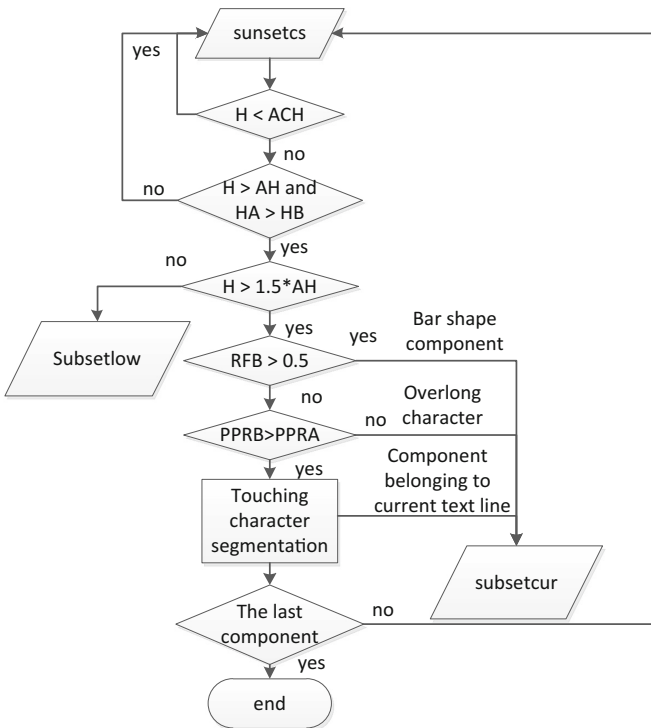


Fig. 9. The LSA first step work flow

The second step continues to deal with connected components still in subsets, which are broken strokes (see Fig. 10, a b c d), separated lower vowels belonging to the current text line (see Fig. 10, e f g), and upper vowels belonging to the next text line (see Fig. 10, h i j).

This step has three works to do:

1. Calculate the centroid and the skeleton of connect components, then detect the intersection between skeleton and the line located by the centroid, and calculated the numbers and the coordinate positions of the intersected points.
2. For the connected components with only one intersected point (see Fig. 11 a b c d), move it from subsets to the subsetcur if its centroid position is above the segmented line, or it belongs to subsetlow if its centroid position is below the segmented line.
3. For the connected components with two intersected points, the skeleton is segmented into the upper part and the lower part according to the line located by the centroid and the coordinate of two intersected points. The number of pixels in the two parts is counted respectively. Connected components are assigned to subsetcur if the pixels in the lower part is more than that in the upper part (see Fig. 11 e f g), otherwise, the connected components will belong to subsetlow(see Fig. 11 h i j).

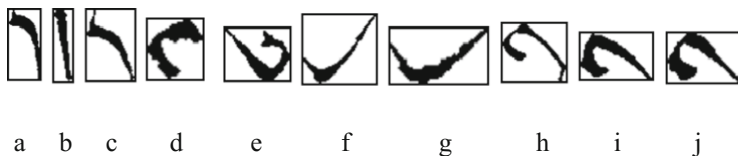


Fig. 10. The connected components of broken strokes, separated lower vowels and separated upper vowels

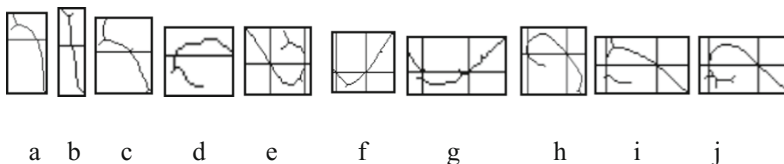


Fig. 11. The skeleton diagram with line located by the centroid

3.4 Merging Image

Since all the connected components that belong to the current text line have been marked in the subsetcur, so the lower part of the current text line is generated by the subsetcur. The complete image of the current text line image is got by merge the upper part and lower part. The input image subtracts the current text line image from the position of the LB to produce a image that is the input image for the next text line.

4 Experimental Results and Discussion

The experimental dataset are from the historical Tibetan scripture “kangjur” of the Beijing version on the paper by means of woodcut. The scripture “kangjur” of the Beijing version have more than 60 thousand images, the dataset just have 1696 text lines from 212 images which is selected at random. The method presented in this paper is implemented in matlab.

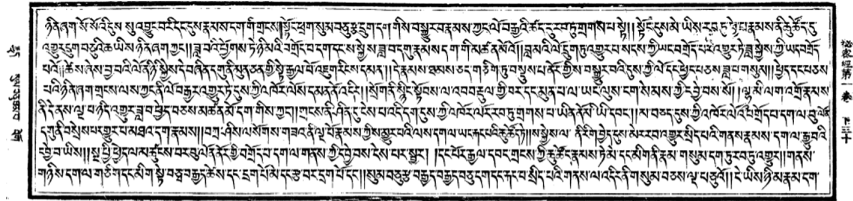


Fig. 12. The input image

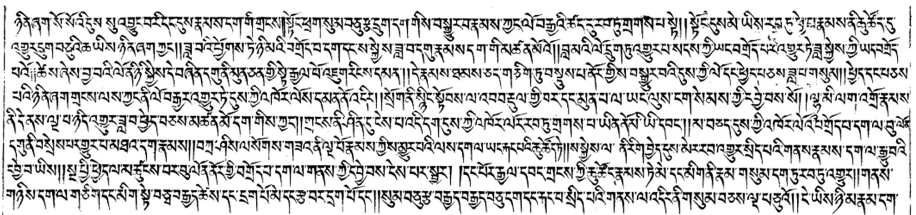


Fig. 13. The image after remove the bounding box

Figure 12 is an original historical Tibetan script image. This method performs image skew detection and correction, using projection profiles to get the baseline of text line, then the bounding box is removed. Figure 13 is the document image without bounding box. Figure 14 gives the text line segment results.

Let N be the number of all text lines, G_j the set of all points inside the ground truth region, R_j the set of all points inside the corresponding result region. The detection rate (DR) and the recognition accuracy rate(RA) are defined as follows:

$$DR = \frac{G \cap R}{G}, RA = \frac{G \cap R}{R} \tag{9}$$

Because text line segmentation is an important part of OCR recognition system, the ideal situation is that the text lines only contains all the components belonging to the text line, and it does not lose any component and does not have any component that do not belong to them. Therefore, we propose completeness rate to measure the segmentation effect. The definition of integrity is as follows:

$$CR = \frac{\sum N_i}{N} \quad N_i = 1, \text{ if } G_i = R_i, \text{ otherwise } N_i = 0 \quad (10)$$

Table 1 shows the performance of contour curve tracking method and our method. Comparing with the contour curve tracking method, our method has a considerable improvement in each evaluation value.

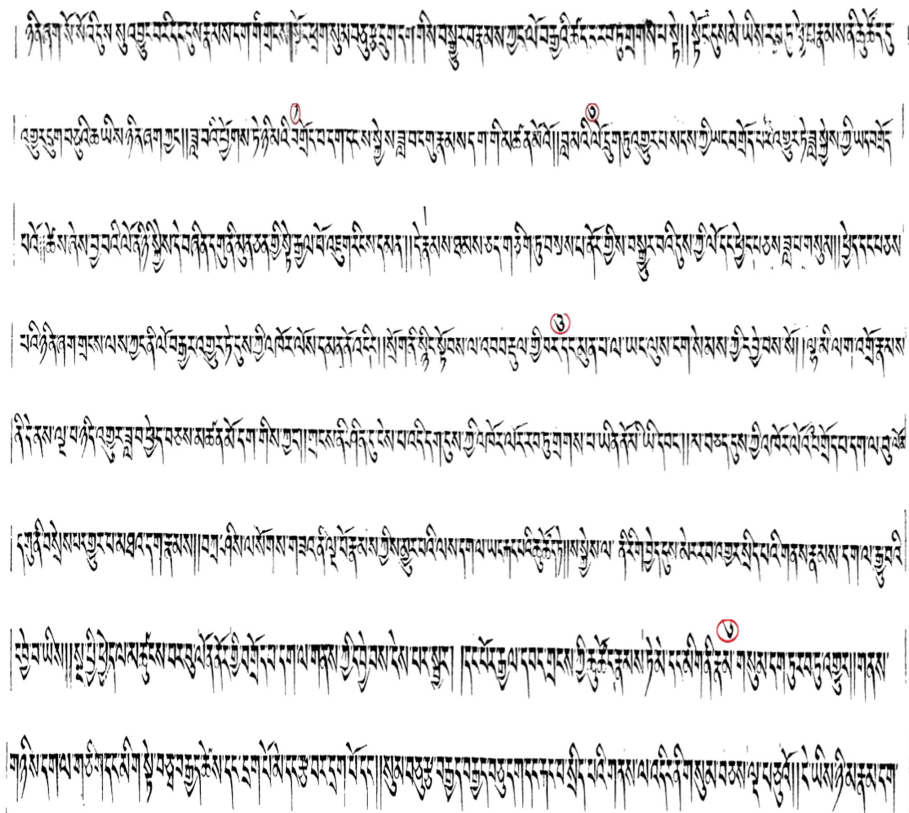


Fig. 14. Result of text line segmentation

Table 1. The performance of contour curve tracking method and our method

Method	N	DR	RA	CR
Contour curve tracking	2196	82.79%	80.09%	33.23%
Our method	2196	91.17%	90.23%	37.51%

Experimental results show that almost all the components belonging to the wrong text line are caused by broken strokes and separated lower vowels. And this method is very efficient to detect the touching characters in adjacent text lines. There are 874 touching characters in the dataset of the 212 pictures, and 840 of them are detected successfully, the touching character's detect ratio is 98.4%.

5 Conclusion and Further Work

Text line segmentation is still one of the most challenging topics in document image analysis. In this paper, we present a text line segmentation method for handwritten historical Tibetan documents based on connected components analysis. This method correct the skew document image, gets the reasonable baseline position by the contour projection, and obtains the text line region by the baseline position from the document image. The connected component's attribution is decided by analyzing the location and shape. The method is suitable for text segmentation from complex layout document image and can overcome the slightly fluctuation of text line. Although the algorithm is reasonably designed and many features about location and shape are analyzed, there are still many wrong parts in the extracted text line image.

Low completeness rate of text line segmentation is not only caused by strict standards, but also by the real historical handwritten documents that is more complicated because of the high frequency of separated vowel characters, broken strokes, and touching characters.

Through experiments, we get the following conclusions for the text line segmentation task for the degraded Tibetan historical document image of wooden printing: (i) the method based on the connected component analysis is feasible for text line segmentation. (ii) it is necessary to correct the skew document image for text line segmentation. (iii) the problem of touching and overlapped characters in text line segmentation of historical Tibetan documents can be solved effectively. (iv) it is not enough to make use of a few features to identified the shape of character.

The focus of future work is to study the shape recognition algorithm of similar vowels and broken strokes. Another issue is to research the better segmentation algorithm for touching and overlapped character.

Acknowledgments. This work was supported by the National Science Foundation (No.61772430), the Program for Leading Talent of State Ethnic Affairs Commission, the Fundamental Research Funds for the Central University of Northwest Minzu University (No. 31920170142), and also supported by the Gansu Provincial first-class discipline program of Northwest Minzu University.

References

1. Li, Y., Ma, L., Duan, L., Wu, J.: A text-line segmentation method for historical tibetan documents based on baseline detection. In: Yang, J., et al. (eds.) CCCV 2017. CCIS, vol. 771, pp. 356–367. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-7299-4_29
2. Zhou, F., Wang, W., Lin, Q.: A novel text line segmentation method based on contour curve tracking for tibetan historical documents. *Recogn. Artif. Intell.* **32**(10), 1854025 (2018). *Image Processing*
3. Manmatha, R., Rothfeder, J.L.: A scale space approach for automatically segmenting words from historical handwritten documents. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1212–1225 (2005)
4. Bar-Yosef, I., Hagbi, N., Kedem, K., Dinstei, I.: Line segmentation for degraded handwritten historical documents. In: 10th ICDAR, pp. 1161–1165 (2009)
5. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. *Int. J. Doc. Anal. Recogn.* **9**(2), 123–138 (2007)
6. Garz, A., Fischer, A., Bunke, H., Ingold, R.: A binarization-free clustering approach to segment curved text lines in historical manuscripts. In: *International Conference on Document Analysis and Recognition*, pp. 1290–1294 (2013)
7. Zahour, A., Likforman-Sulem, L., Boussalaa, W., Taconet, B.: Text line segmentation of historical arabic documents. In: *9th International Conference Document Analysis and Recognition*, vol. 1, pp. 138–142 (2007)
8. Baima, Y.Z.: Research on feature extraction of tibetan characters. *Comput. Knowl. Technol.* **9**, 6362–6364 (2013)



Online Handwriting Tibetan Character Recognition Based on Two-Dimensional Discriminant Locality Alignment

Zhengqi Cai^{1,2} and Weilan Wang^{1(✉)}

¹ Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730000, Gansu, People's Republic of China

caizhengqi@l26.com, wangweilan@xbmu.edu.cn

² College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730000, Gansu, People's Republic of China

Abstract. Discriminant Locality Alignment (DLA) has been successfully applied in handwriting character recognition. In this paper, a new manifold based subspace learning algorithm, which is called Two-dimensional Discriminant Locality Alignment (2DDLDA) algorithm, is proposed for online handwriting Tibetan character recognition (OHTCR). The proposed algorithm integrates the idea of DLA and two-dimensional feature extraction algorithm. At first, extracting direction feature matrix and edge feature matrix of Tibetan character respectively, they are together formed original feature matrix. Then, in part optimization stage, for each character sample, a local patch is built by the given sample and its neighbors, and an object function is designed to preserve local discriminant information. Third, in whole alignment stage, the alignment trick is used to align all part optimizations to the whole optimization. The projection matrix can be obtained by solving a standard eigen-decomposition problem. Finally, a SMQDF classifier is used training and recognition. Experimental results demonstrate that 2DLDA is superior to LDA and IMLDA in terms of recognition accuracy. In addition, 2DLDA can overcome the matrix singular problem and small sample size problem in OHTCR.

Keywords: Online handwriting recognition · Tibetan character recognition
Two-dimensional discriminant locality alignment (2DDLDA) · Subspace learning

1 Introduction

With the acceleration of Tibetan information process, the demand of Tibetan character recognition system is becoming more and more prominent. At present, handwriting Tibetan character recognition (HTCR) has made great progress in both research and practical application [1–6]. However, the recognition of Tibetan character is different from handwriting recognition of other languages, it poses a special challenge due to a complex structure, wide varieties in writing style, a large character set and many instances of highly similar characters. Figure 1 illustrates some samples of handwriting

Tibetan character. Unconstrained online HTCR is still an open problem remaining to be solved, for it is still challenging to reach high recognition rate.

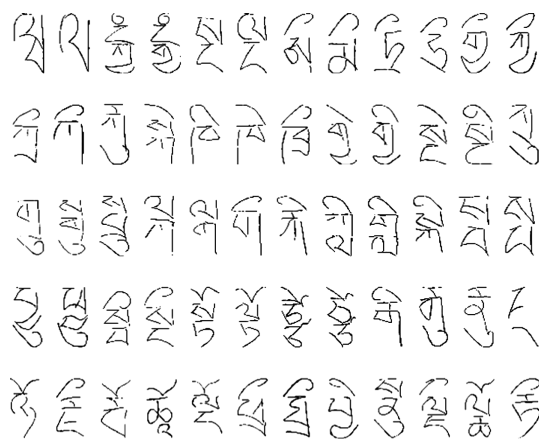


Fig. 1. Some samples of handwriting Tibetan character

As shown in Fig. 1, many Tibetan characters are almost identical to other characters except for only a small difference. However, the small difference can be lost during the feature extraction process. So the discriminate information extraction is crucial for improvement of the recognition performance. In OHTCR, dimensionality reduction is the process of transforming data from a high dimensional space to a low dimensional space to reveal the intrinsic structure of the distribution of data. It plays a crucial role in the field of computer vision and pattern recognition as a way of dealing with the “curse of dimensionality”. In past decades, a large number of dimensionality reduction algorithms have been proposed and studied. Among them, principal components analysis (PCA) [7] and Fisher’s linear discriminant analysis (LDA) [8] are two of the most popular linear dimensionality reduction algorithms.

PCA maximizes the mutual information between original high dimensional Gaussian distributed data and projected low dimensional data. PCA is optimal for reconstruction of Gaussian distributed data. However, it is not optimal for classification problems. LDA overcomes this shortcoming by utilizing the class label information. It finds the projection directions that maximize the trace of the between-class scatter matrix and minimize the trace of the within-class scatter matrix simultaneously. However, LDA is only a suboptimal model which suffers from the class separation problem. The objective of LDA can be formulated as maximizing the sum of all the pairwise distances between different classes, which will overemphasize the large distance of the already well-separated classes, and confuse the small distance classes that are close in the original feature space. Li and Yuan [9] proposed a new method of feature extraction using two-dimensional linear discriminant analysis (2DDLDA), and directly uses the matrix to extract the discriminant feature without a vectorization

procedure, it has a great advantage over the one-dimensional method in calculation and processing efficiency.

Zhang [10] proposed a local linear dimensionality reduction algorithm called discriminative locality alignment (DLA). The DLA takes into account the locality of samples, deals with the nonlinearity of the samples distribution, and preserves the discriminability of classes as well. However, the DLA algorithm is based on the vector space, and the data must be vectorized during calculation, which destroys the spatial distribution characteristics and structure information of the data. Based on the stability and effectiveness of DLA algorithm in recognition performance, in this paper, we combine the idea of DLA algorithm with two-dimensional feature extraction algorithm, and proposes a two-dimensional discriminative locality alignment (2DDLA) algorithm to improve the recognition performance in OHTCR.

The rest of paper is organized as follows. Section 2 introduces two-dimensional discriminant locality alignment (2DDLA) algorithm for extracting discriminative features for OHTCR and details the basic formulation. Section 3 introduces SMQDF classifier. We perform experiments in Sect. 4 to show the effectiveness of the proposed method and Sect. 5 gives concluding remark.

2 Two-Dimensional Discriminative Locality Alignment

2DDLA aims to extract discriminative information from patches. To achieve this goal, one patch is first built for each sample. Each patch includes a sample and its within-class nearest samples and its between-class nearest samples. Then an objective function is designed to preserve the local discriminative information of each patch. Finally, all the part optimizations are integrated together to form a global coordinate according to the alignment trick. The projection matrix can be obtained by solving a standard Eigen decomposition problem.

2.1 Part Optimization

Suppose we have a set of samples $X = [X_1, X_2, \dots, X_N]$ from C different classes, $X_i \in \mathbb{R}^{m \times n}$. For a given sample X_i , we select k_1 nearest neighbors from the samples of the same class with X_i and name them as the neighborhoods of a same class: $X_{i^1}, \dots, X_{i^{k_1}}$, we also select k_2 nearest neighbors from samples of different classes with X_i , and name them as neighborhoods of different classes: $X_{i_1}, \dots, X_{i_{k_2}}$. By putting them together, we can build the local patch for X_i as $\Pi_i = [X_i, X_{i^1}, \dots, X_{i^{k_1}}, X_{i_1}, \dots, X_{i_{k_2}}]$. For each patch, the corresponding output in the low-dimensional space is denoted by $\Gamma_i = [Y_i, Y_{i^1}, \dots, Y_{i^{k_1}}, Y_{i_1}, \dots, Y_{i_{k_2}}]$.

In the low-dimensional space, we expect that distances between the given sample and its within-class samples are as small as possible, while distances between the given sample and its between-class samples are as large as possible. So we have

$$\arg \min_{Y_i} \sum_{j=1}^{k_1} \|Y_i - Y_{i_j}\|_F^2 \quad (1)$$

$$\arg \max_{Y_i} \sum_{p=1}^{k_2} \|Y_i - Y_{i_p}\|_F^2 \quad (2)$$

where $\|\cdot\|_F$ denotes the Frobenius norm for a matrix.

Since the patch formed by the local neighborhood can be regarded approximately linear, we formulate the part discriminator by using the linear manipulation as follows:

$$\arg \min_{Y_i} \left(\sum_{j=1}^{k_1} \|Y_i - Y_{i_j}\|_F^2 - \beta \sum_{p=1}^{k_2} \|Y_i - Y_{i_p}\|_F^2 \right) \quad (3)$$

where β is a scaling factor ($\beta \in [0,1]$). The coefficients vector is defined as:

$$W_i = \left[\overbrace{1, \dots, 1}^{k_1}, -\overbrace{\beta, \dots, -\beta}^{k_2} \right] \quad (4)$$

Then the Eq. (3) reduces to:

$$\begin{aligned} & \arg \min_{Y_i} \left(\sum_{j=1}^{k_1} \|Y_i - Y_{i_j}\|_F^2 W_i(j) + \sum_{p=1}^{k_2} \|Y_i - Y_{i_p}\|_F^2 W_i(p + k_1) \right) \\ &= \arg \min_{Y_i} \left(\sum_{j=1}^{k_1+k_2} \|Y_{F_i\{j\}} - Y_{F_i\{j+1\}}\|_F^2 W_i(j) \right) \\ &= \arg \min_{Y_i} \text{tr}(\Gamma_i(L \otimes I_n)(\text{diag}(W_i) \otimes I_n)(R \otimes I_n)(\Gamma_i)^T) \\ &= \arg \min_{\Gamma_i} \text{tr}(\Gamma_i T_i \Gamma_i^T) \end{aligned} \quad (5)$$

where $\text{tr}(\cdot)$ denotes the trace operator, the operator \otimes denotes the Kronecker product of matrix, $F_i = \{i, i^1, \dots, i^{k_1}, i_1, \dots, i_{k_2}\}$ is the index set for the i th patch, $e_{k_1+k_2} = [1, \dots, 1]^T \in \mathbb{R}^{k_1+k_2}$, $I_{k_1+k_2}$ is a $(k_1+k_2) \times (k_1+k_2)$ identity matrix, $R = [-e_{k_1+k_2}, I_{k_1+k_2}]^T$, $L = \begin{bmatrix} -e_{k_1+k_2}^T \\ I_{k_1+k_2} \end{bmatrix}$, and

$$T_i = (L \otimes I_n)(\text{diag}(W_i) \otimes I_n)(R \otimes I_n) \quad (6)$$

2.2 Whole Alignment

After the part optimization step, we unify the optimizations together as a whole one by assuming that the coordinate for the i 'th patch $\Gamma_i = [Y_i, Y_{i^1}, \dots, Y_{i^{k_1}}, Y_{i_1}, \dots, Y_{i_{k_2}}]$ is selected from the global coordinate $\Gamma = [Y_1, Y_2, \dots, Y_N]$, such that $\Gamma_i = \Gamma S_i$, where $S_i \in \mathbb{R}^{(N \times n) \times ((k_1+k_2+1) \times n)}$ is the selection matrix and an entry is defined as follows:

$$(S_i)_{pq} = \begin{cases} I_n & \text{if } p = F_i(q) \\ 0_n & \text{else} \end{cases} \quad (7)$$

Then Eq. (7) can be rewritten as

$$\arg \min_{\Gamma} \text{tr}(\Gamma S_i T_i S_i^T \Gamma^T) \quad (8)$$

By summing over all the part optimizations described as Eq. (8), we can obtain the whole alignment as

$$\begin{aligned} & \arg \min_{\Gamma} \sum_{i=1}^N \text{tr}(\Gamma S_i T_i S_i^T \Gamma^T) \\ &= \arg \min_{\Gamma} \text{tr}((\Gamma \sum_{i=1}^N S_i T_i S_i^T) \Gamma^T) \\ &= \arg \min_{\Gamma} \text{tr}(\Gamma L \Gamma^T) \end{aligned} \quad (9)$$

where $L = \sum_{i=1}^N S_i T_i S_i^T \in R^{N \times N}$ is the alignment matrix.

To obtain the linear and orthogonal projection matrix W , such as $Y = W^T X$, Eq. (9) is deformed as follows:

$$\arg \min_W \text{tr}(W^T X L X^T W), \text{ s.t. } W^T W = I \quad (10)$$

The transformation matrix W that minimizes the objective function is given by the minimum eigenvalue solution to the standard eigenvalue problem,

$$X L X^T P = \lambda P \quad (11)$$

3 SMQDF

3.1 MQDF

MQDF [11] classifier's discriminate function is formulated as

$$\begin{aligned} f(Y, \omega_j) &= \sum_{i=1}^k \frac{[(Y - \mu_j)^T \zeta_i^{(j)}]^2}{\lambda_i^j} + \sum_{i=k+1}^m \frac{[(Y - \mu_j)^T \zeta_i^{(j)}]^2}{\lambda} \\ &+ \sum_{i=1}^k \log \lambda_i^{(j)} + \sum_{i=k+1}^m \log \lambda \quad j = 1, 2, \dots, C \end{aligned} \quad (12)$$

where, Y is input feature vector, m is line number of feature matrix, μ_j denotes the mean vector of class ω_j , $\lambda_i^{(j)}$ and $\zeta_i^{(j)}$ denote the i th larger eigenvalue and the corresponding eigenvector of the covariance matrix of class ω_j , respectively. k is the number of dominant principal eigenvectors that are kept in MQDF, λ is experiment parameter.

We can obtain the classified result based on the following criterion: If $f(Y, \omega_i) = \min_{1 \leq j \leq C} f(Y, \omega_j)$ (C is class number), then we believe that input pattern Y belongs to the ω_i class.

3.2 SMQDF

MQDF classifier is widely used in the area of character recognition. However, it only applies to feature vector, and it is not appropriate for feature matrix. For this reason, SMQDF (second modified quadratic discriminate function) classifier is generated by improving MQDF classifier, its discriminate function as shown in the follow formula (13). We take it as a baseline classifier.

$$f(Y, \omega_j) = \sum_{i=1}^{m-1} \frac{((Y - \mu_j)^T \zeta_i^{(j)})^T ((Y - \mu_j)^T \zeta_i^{(j)})}{\lambda_i^{(j)}} + \frac{((Y - \mu_j)^T \zeta_m^{(j)})^T ((Y - \mu_j)^T \zeta_m^{(j)})}{\lambda} + \sum_{i=1}^{m-1} \log \lambda_i^{(j)} + \log \lambda \quad j = 1, 2, \dots, C \quad (13)$$

where, Y is feature matrix, m is positive integer, When classifies, Y belongs to the class whose $f(Y, \omega_i)$ is minimum. To compensate for the estimation error of parameters on limited training samples, the minor eigenvalues are replaced with a constant λ . It can be set to a class-independent constant or class-dependent constant. Here we set λ to be class-independent for its superior performance. λ is computed by

$$\lambda = \frac{1}{c * d} \sum_{j=1}^c \sum_{i=1}^d \lambda_i^{(j)} \quad (14)$$

4 Experiment Results

4.1 Experiment Data

We evaluated the recognition performance of 2DDLA on a databases of handwritten Tibetan characters, collected by our group, contains the handwriting samples of 7240 characters, 5000 samples per class [12]. In order to reduce the computing cost, we only selected 562 frequently used characters, 2000 samples per class for training and 500 samples per class for testing.

For character image pre-processing and feature extraction, we adopt the same methods as in [6]. Each character image is normalized to 48×96 pixels, the

directional features and edge features are extracted. The resulting 60×12 feature matrix is projected onto a 12×12 subspace learned by 2DDLA, then the baseline classifier SMQDF is designed on this 12×12 subspace.

4.2 Choice of Parameters for 2DLDA

Since the parameters setup for 2DDLA is essential for its performance, we carried out the 2DDLA parameter optimization experiments before for OHTCR. In the model of 2DDLA, there are three parameters: k_1 , k_2 and β , where k_1 is the number of the samples from identical class in the given patch, k_2 is the number of the samples from other classes in the same given patch, and parameter β is the scale parameter. In order to find a proper range for the dominant parameters k_1 , k_2 and β in 2DDLA, we will investigate the effects of the three model parameters on the recognition rates in the validation phase based on our collected database.

Suppose n is the training sample number in each class ($n = 2000$), N is the total training sample number ($N = 562 \times 2000 = 1024000$), and C is class number ($C = 562$). Then, k_1 and k_2 could be chosen in the range of $[1, n-1]$ and $[0, N-n]$ respectively. Therefore, $1 \leq k_1 \leq 1999$, $0 \leq k_2 \leq 1022000$, and $0 \leq \beta \leq 1$.

To evaluate the effects of the three model parameters, firstly, we analyze the effect of the scale parameter β , by fixing patch building parameters k_1 and k_2 to arbitrary values. For a given pair parameters k_1 and k_2 , we can obtain the recognition rate curve with respect to β , as shown in Fig. 1. Base on the figures, we observe that the best recognition rates are obtained when β is neither too small nor too larger.

Secondly, we analyze the effects of patch building parameters k_1 and k_2 , by fixing scale parameter $\beta = 0.1$. When we vary k_1 and k_2 simultaneously, the best recognition rate with the corresponding to β can be acquired. Table 1 shows that the details of the best recognition rate. Figure 2 shows that best recognition rate with the corresponding $k_1 = 50$ and $k_2 = 300$ in this experiment (Fig. 3).

Table 1 shows that, the best combination of k_1 , k_2 , and β are $k_1 = 50$, $k_2 = 300$, $\beta = 0.1$ and $k_1 = 100$, $k_2 = 300$, $\beta = 0.3$, with the corresponding accuracy 99.38%. Considering the computing cost, in the following experiments, we use the best combination of k_1 , k_2 , and β is 50, 300, 0.1 respectively.

4.3 Evaluation Experiments

To evaluation the performance of 2DDLA in SHCCR, we compare the performance of 2DDLA, LDA, IMLDA [13] and 2DLDA in terms of recognition rate over SMQDF classifier [6]. The experimental results are summarized in Table 2. We can see that the proposed method obtains higher top 1 and top 10 recognition rate than other method.

From Table 2, it is shown that the recognition rates of 2DDLA are significantly higher than that of IMLDA and 2DLDA respectively. It also shows the discriminate information extraction performance is very competitive in OHTCR.

To illustrate the effects of the 2DDLA, Fig. 4 shows some sample that are mis-recognized by IMLDA and 2DLDA, but can be corrected by 2DDLA.

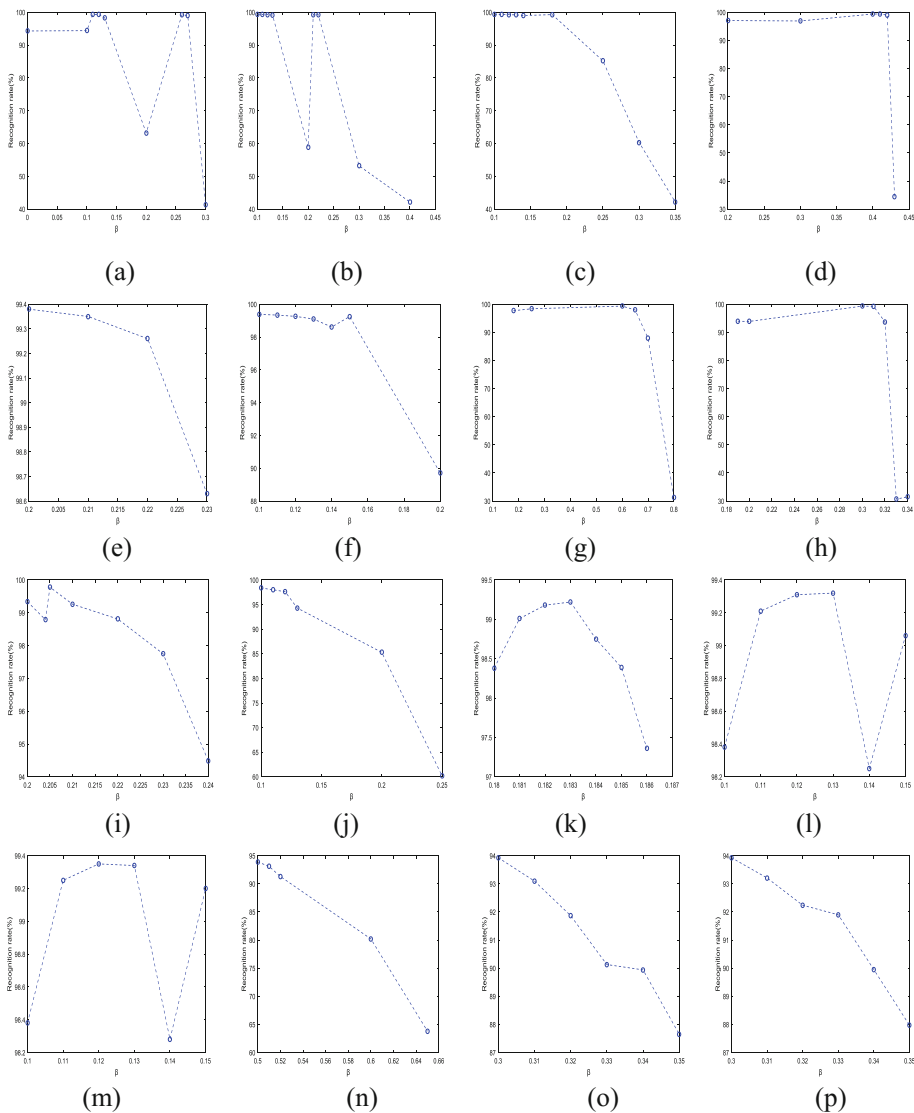


Fig. 2. For a given pair parameters k_1 and k_2 , the best recognition rate with the corresponding β . (a) $k_1 = 32$, $k_2 = 100$, (b) $k_1 = 32$, $k_2 = 200$, (c) $k_1 = 32$, $k_2 = 300$, (d) $k_1 = 50$, $k_2 = 100$, (e) $k_1 = 50$, $k_2 = 200$, (f) $k_1 = 50$, $k_2 = 300$, (g) $k_1 = 80$, $k_2 = 100$, (h) $k_1 = 80$, $k_2 = 200$, (i) $k_1 = 80$, $k_2 = 300$, (j) $k_1 = 100$, $k_2 = 300$, (k) $k_1 = 100$, $k_2 = 500$, (l) $k_1 = 100$, $k_2 = 800$, (m) $k_1 = 100$, $k_2 = 990$, (n) $k_1 = 300$, $k_2 = 500$, (o) $k_1 = 300$, $k_2 = 800$, (p) $k_1 = 300$, $k_2 = 990$, (q) $k_1 = 500$, $k_2 = 800$, (r) $k_1 = 500$, $k_2 = 1000$, (s) $k_1 = 500$, $k_2 = 1200$, (t) $k_1 = 500$, $k_2 = 1500$, (u) $k_1 = 800$, $k_2 = 1500$, (v) $k_1 = 800$, $k_2 = 2500$, (w) $k_1 = 800$, $k_2 = 3800$, (x) $k_1 = 1000$, $k_2 = 300$

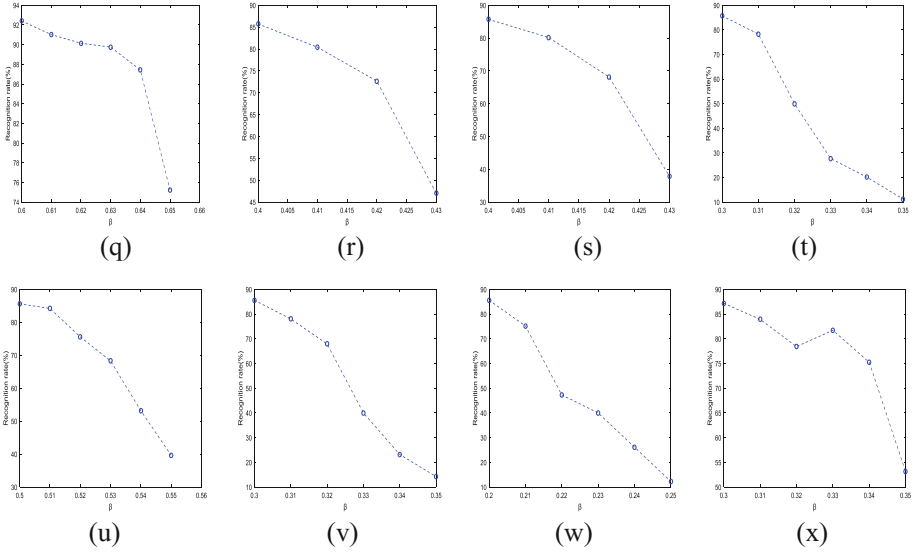


Fig. 2. (continued)

Table 1. Best recognition rate (%) with the corresponding k_1 and k_2 .

K_1	K_2	β	Recognition rates
32	100	0.11	99.36
32	200	0.11	99.36
32	300	0.11	99.36
50	100	0.41	99.37
50	200	0.21	99.37
50	300	0.10	99.38
80	100	0.61	99.34
80	200	0.31	99.34
80	300	0.21	99.34
100	300	0.30	98.38
100	500	0.18	99.18
100	800	0.12	99.31
100	990	0.12	99.35
300	500	0.50	93.93
300	800	0.30	93.93
300	990	0.30	93.93
500	800	0.60	92.41
500	1000	0.40	85.73
500	1200	0.40	85.72
500	1500	0.30	85.73
800	1500	0.50	85.62
800	2500	0.3	85.62
800	3800	0.2	85.62
1000	3000	0.3	87.19

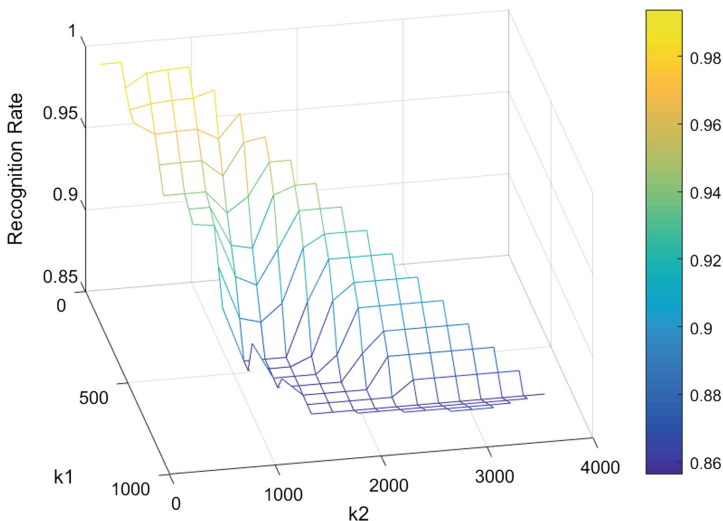


Fig. 3. Recognition rate vs. k_1 and k_2

Table 2. Best recognition rates (%) of three methods.

Methods	Top 1	Top 10
IMLDA	55.56	92.73
2DLDA	83.73	98.21
2DDLA	85.9	99.38

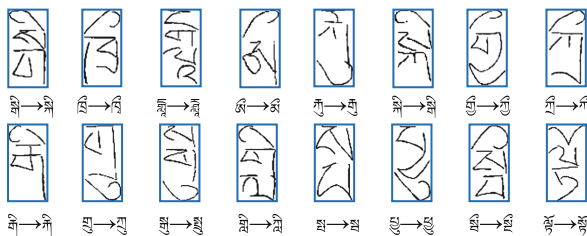


Fig. 4. Some misrecognized by SMQDF, but corrected by compound distance method

5 Conclusion

In this paper, we present a new method, called two-dimensional discriminative locality alignment (2DDLA). Compare with IMLDA and 2DLDA, the proposed method has better recognition rate. It inherits all the advantages of DLA and can overcome the matrix singular problem and small sample size problem in OHTCR.

Acknowledgment. This work was supported by the Fundamental Research Funds for the Central University of Northwest Minzu University (No. 31920170142), the Program for Leading Talent of State Ethnic Affairs Commission, the National Science Foundation (No. 61375029), and supported by the Gansu Provincial first-class discipline program of Northwest Minzu University.

References

1. Wang, W., Ding, X., Qi, K.: Study on similar character in Tibetan character recognition. *J. Chin. Inf. Process.* **16**(4), 60–65 (2002)
2. Wang, H., Ding, X.: Multi-font printed Tibetan character recognition. *J. Chin. Inf. Process.* **17**(6), 47–52 (2003)
3. Wang, W., Qian, J., Duojie, Z., Ma, M., Qi, K., Duo, L., et al.: A method of online handwritten Tibetan characters recognition. State Intellectual Property Office of the Peoples Republic of China (2011). ZL200910128595.8
4. Huang, H., Da, F., Hang, X.: Wavelet transform and gradient direction based feature extraction method for offline handwritten Tibetan letter recognition. *J. Southeast Univ.* **30**(1), 27–31 (2014)
5. Ma, L.L., Wu, J.: A Tibetan component representation learning method for online handwritten Tibetan character recognition. In: *Proceedings of the 14th ICFHR*, pp. 317–322 (2014)
6. Wang, W., Qian, J., Wang, D., Duojie, Z.: Online handwriting recognition of Tibetan characters based on the statistical method. *J. Commun. Comput.* **8**(2011), 188–200 (2011)
7. Jolliffe, I.T.: *Principal Component Analysis*. Springer, New York (1986). <https://doi.org/10.1007/978-1-4757-1904-8>
8. Fisher, R.A.: The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188 (1936)
9. Li, M., Yuan, B.: 2D-LDA: a statistical linear discriminant analysis for image matrix. *Pattern Recognit. Lett.* **26**(5), 527–532 (2005)
10. Zhang, T., Tao, D., Li, X., et al.: Patch alignment for dimensionality reduction. *IEEE Trans. Knowl. Data Eng.* **10**(2), 433–439 (2009)
11. Kimura, F., Takashina, K., Tsuruoka, S., Miyake, Y.: Modified quadratic discriminant functions and its application to Chinese character recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **9**(1), 149–153 (1987)
12. Wang, W., Lu, X., Cai, Z., Shen, W., Fu, J., Caike, Z.: Online handwritten sample generated based on component combination for Tibetan-Sanskrit. *J. Chin. Inf. Process.* **31**(5), 64–73 (2017)
13. Qian, J., Wang, J.: A novel approach for online handwriting recognition of Tibetan characters. In: *Proceedings of IMECS 2010*, pp. 1–4 (2010)



Complex Printed Uyghur Document Image Retrieval Based on Modified SURF Features

Aliya Batur¹, Patigul Mamat², Wenjie Zhou¹, Yali Zhu¹,
and Kurban Ubul¹(✉)

¹ School of Information Science and Engineering,
Xinjiang University, Urumqi 830046, China
kurbanu@xju.edu.cn

² School of Mathematics and Information,
Hotan Normal College, Hotan 848000, China

Abstract. As an important part of information retrieval, it is important to improve the accuracy of document image retrieval system. This paper proposes a document image retrieval method based on modified SURF features. Firstly, FAST+SURF features are extracted from the image, and then the similarity degree is retrieved by using different kinds of distances and matching points respectively. With the change of size, angle and illumination, the FLANN bidirectional matching and KD-Tree +BBF matching are implemented for its feature points; finally, based on these two kinds of retrieval methods, various Uyghur document image databases that have been collected and retrieved are searched. The experimental results indicated that both search methods can achieve accurate search requirements, but in computational complexity based on the matching number of retrieval is more convenient. At the same time, the comparison experiment proves that the proposed method is superior to the original feature in the retrieval time.

Keywords: SURF feature · FALNN bidirectional match
KD-Tree and BBF match · Complex document image retrieval

1 Introduction

With the rapid development of multimedia information technology, document images have become the main information resource, which also causes the explosive growth of document image. How to obtain document image content efficiently has become a hot research topic in domestic and overseas research. Xiaoxiao et al. [1] compared 64-dimensional vectors to describe the feature points that were more suitable for image data processing. Two modified SVM algorithms were used to extract information from matched images and compare with traditional SVM algorithm. Zhao et al. [2] first extracted the 64-dimensional SURF feature points, and based on the FLANN algorithm for bidirectional matching, matching pairs for PROSAC analysis, excluding mismatched pairs to improve the image matching accuracy, and effectively reduce the matching time. Cheon et al. [3] proposed an enhanced Fast Robustness Feature (e-SURF) algorithm to save memory and increase speed. Zhang et al. [4] proposed an

modified matching algorithm based on SURF (Speeded Up Robust Features) feature point matching, which combined SURF and RANSAC (Random Sample Consensus) algorithm. Chen et al. [5] proposed to improve the detection of SURF key points, extract the feature points of the image detail region, and achieve accurate matching based on KD-Tree bidirectional matching. Luo et al. [6] modified the SURF descriptor using the DAISY descriptor, and matched the target image with nearest neighbor distance ratio (NNDR), with a maximum matching rate of 95.78%. Wang et al. [7] proposed a robust feature (SURF) based on improved accelerated fast image matching algorithm, The RELIEF-F algorithm is used to reduce and simplify the improved SURF descriptors to achieve image registration, and finally the improved algorithm is verified by the experiments of real-time and robustness.

This paper analyzes the Uyghur complex document image without layout analysis, proposes to the modified SURF features to achieve the key points extraction, and to achieve effective retrieval from the large-scale image database. The algorithmic flow of this paper is shown as in Fig. 1.

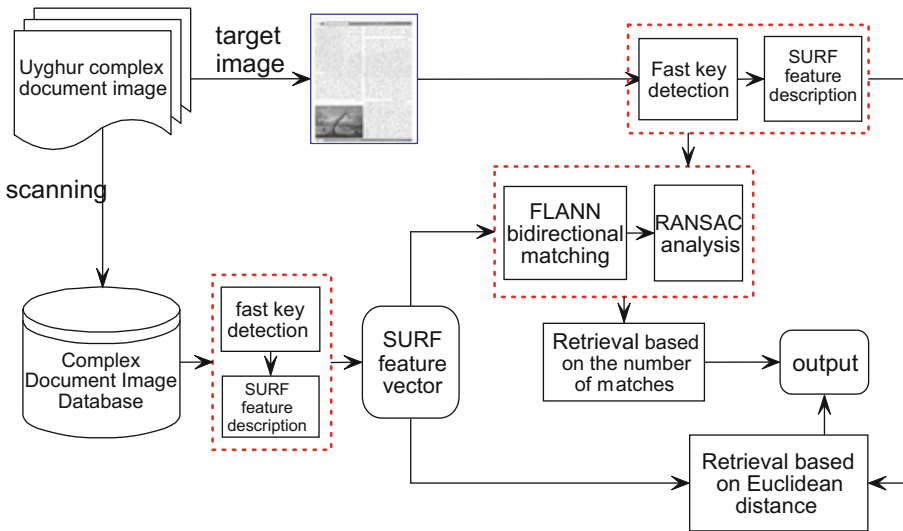


Fig. 1. The block diagram of Uyghur printed complex document image retrieval based on modified SURF feature.

2 Fast and SURF Feature Extraction

The process of fast robust feature extraction (SURF) is similar to SIFT, and consists of two parts: key point detection and feature description. However, it maintains the same image size and changes the size and scale of the box filter in multiples relationship. Based on the integral image, the proportional space is filtered so that the feature detection takes much less time than SIFT. And the key points detected in the scale space have the size translational robustness. In the feature description, the Haar wavelet

response value in the fan-shaped area is calculated, the main direction of the key points is determined, and the computational complexity is reduced.

However, shortening time parameter is not ideal for the complicated document images of text and video. Therefore, in order to quickly detect the key points of the image in the complex layout, the author makes full use of pixel gray level information, detects the corners based on the FAST algorithm, and describes the sub-description with the SURF descriptor to form the 64 dimension FAST and SURF feature, effectively shortening the features Extraction time [8]. The Flow chart of modified SURF feature key point detection is shown in the following Fig. 2.

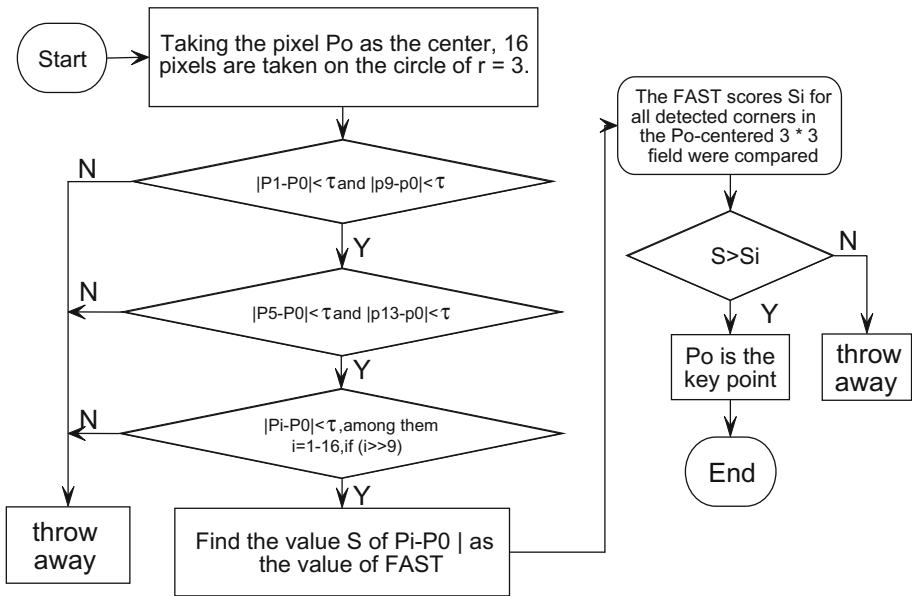


Fig. 2. Flow chart of modified SURF feature key point detection

3 Fast and SURF Feature Matching Analysis

To improve the matching speed of Uyghur complex document images, the author implements two-way fast approximate nearest neighbor (FLANN) matching for different layout images, and compares the results with KD-Tree and BBF matching results, from the performance of matching system to establish a retrieval system, and realize the effective retrieval of Uyghur complex document images.

3.1 Bidirectional FLANN Match

Due to the SURF feature vector is a high-dimensional vector, the matching process is equivalent to the nearest neighbor search problem in high-dimensional space, and the operation is complex. Therefore, this paper starts from the rapidness of FLANN

matching, and matches in two directions successively to get the location information of matching pair. By comparing the location of the matching point to determine whether it is correct. In order to effectively remove the mismatched point pairs, the author uses the RANSAC algorithm to calculate the distance between the matched points and the projection matrix, and compares it with the threshold value, effectively eliminating the outer points and improves the matching accuracy. The original image FALNN bidirectional matching results are shown in Fig. 3.

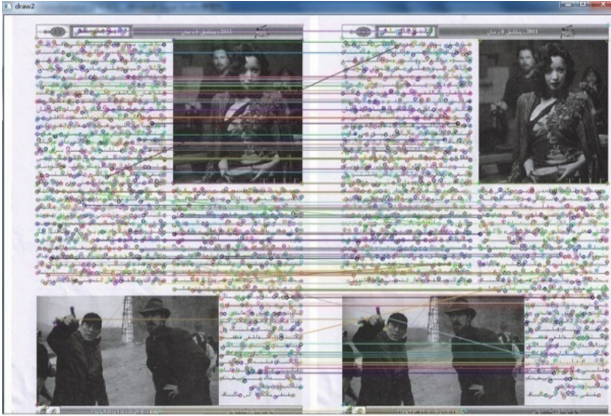


Fig. 3. Schematic diagram of modified SURF features bidirectional FLANN matching

3.2 KD-Tree and BBF Match

KD-Tree is a tree structure for realizes K-nearest neighbor search and matching in large-scale high-dimensional eigenvector space. Its research mainly consists of two parts, namely, the establishment of tree structure and the nearest neighbor search. With the increase of image feature vector dimension, the KD-Tree search ability is greatly reduced. Therefore, starting from the modified KD-Tree, this paper finds the nearest neighbor distance within the limit of maximum backtracking times, and compares the distance ratio with a predetermined threshold to determine whether it is a matching key point [9]. In this paper, the process of improving KD-Tree matching by improving 64-dimensional SURF features is shown in Fig. 4.

The matching efficiency of the matching system under different transformation conditions is evaluated by the matching rate, the correct matching rate and the false matching rate, and its mathematical expression is as follows:

$$\text{Match rate} = \frac{\text{The total number of matched pairs}}{\text{The total number of feature points detected}} \quad (1)$$

$$\text{Correct match rate} = \frac{\text{The total number of correct matched}}{\text{The total number of matched pairs}} \quad (2)$$

$$\text{Mismatch rate} = \frac{\text{The total number of error matched}}{\text{The total number of matched pairs}} \quad (3)$$

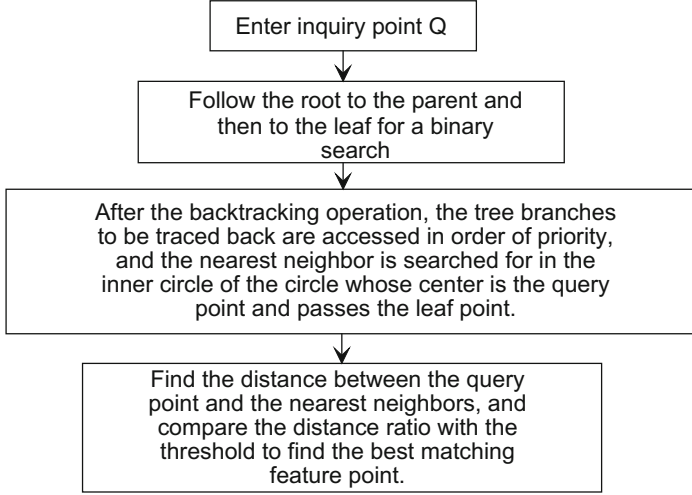


Fig. 4. The description of modified KD-Tree match

4 Uyghur Complex Document Image Retrieval Method

In this paper, the distance-based similarity measure and the matching number-based similarity measure are used. Four eigenvector distance similarity measures algorithms are selected and they are as follows:

$$\text{Euclidean distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (4)$$

$$\text{Manhattan distance} = |x_1 - x_2| + |y_1 - y_2| \quad (5)$$

$$\text{Chebyshev distance} = \max(|x_1 - x_2|, |y_1 - y_2|) \quad (6)$$

$$\text{Cosine distance} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}} \quad (7)$$

In the retrieval system that based on the matching number, the correct number of matches between the document image and each image in the image database is calculated from the correct number of matches, and then the number of correct matches between each image in the image database is sorted and sorted in descending order to effectively retrieve the document image. The more similar complex document images, the greater the number of matches. The calculation of retrieval system is:

$$\text{Retrieval rate} = \frac{N - S}{N - 1} * 100\% \quad (8)$$

Where N is the size of the complex document image database, and S is the position number of the target document image output in the retrieval system window.

5 Experimental Results and Analysis

5.1 Experimental Data

Collection of Uyghur complex layout of books, magazines, documents, scanned with a resolution of 100 dpi to form a depth of 8. bmp format of weaving complex document images, construction of 1000 complex document image database. The system is in 4 GB memory, Windows7_64 bit operating system environment, and Visual Studio 2010 programming.

5.2 Matching Analysis Under Various Transformation Conditions

The original SURF feature detection relies on the choice of Octaves, Intervals, and thresholds. Under the different thresholds (Octaves, Intervals, Init-sample, THRES), the number of feature points to be acquired varies greatly. To test and verify the feasibility of FAST and SURF features, the original SURF features were extracted at (4, 4, 2, 0.0004f) thresholds for complex text documents, in order to obtain the same layout with different sizes, and compared with FAST and SURF Features for performance analysis. The experimental results are shown in Table 1.

Table 1. Number of FAST+SURF key points and time statistics of different sizes image under different threshold [10].

Image size	Feature					
	Performance	SIFT	SURF	FAST (50) +surf	FAST (100) +surf	FAST (150) +surf
803 × 1145	Key points	3276	3537	9767	7195	4960
	Occupation time (S)	30.405	15.866	0.021	0.01	0.009
1606 × 2290	Key points	10320	11516	22028	9414	9299
	Occupation time (S)	105.450	51.141	0.031	0.019	0.017
3212 × 4581	Key points	27820	38115	52967	17764	7839
	Occupation time (S)	250.492	162.491	0.082	0.04	0.035

In order to detect the robustness of the extracted features to the rotation, scale and illumination transformation, the modified SURF eigenvectors of the Uyghur complex document image with the size of 1606×2290 were extracted under different transformations. Based on FLANN bidirectional matching, KD-Tree and BBF matches the number of exact match pairs. When the threshold $\gamma = 0.1$, the results of FLANN bidirectional matching and KD-Tree and BBF under the dimensional transformation are shown in Table 2.

Table 2. Uyghur document image different matching results of FAST+SURF features under scale transform condition [10].

	FLANN bidirectional matching			KD-Tree and BBF match		
	Whole	1:1/2	1:1/4	Whole	1:1/2	1:1/4
The total number of key points	9414	4369	2082	9414	4369	2082
The total number of matching pairs	1145	454	200	9335	4264	2017
Correctly matched pairs	759	363	172	7582	4151	573
Correct match rate (%)	66.29	79.96	86	81.22	97.35	28.40

As can be seen from Table 2 that the image area decreases, the number of feature points detected decreases, and the total number of matches also cut back. Therefore, for thousands of key points, the stability of FLANN bidirectional matching is stronger than KD-Tree and BBF matching. To test and verify the rotation invariance of the selected features, the complex document images are rotated anticlockwise or clockwise in different angular ranges, and matching based on different matching algorithms. The experimental results are shown in Table 3.

Table 3. Two kinds of FAST (100)+SURF feature points matching results under Uyghur document image rotation transform

FLANN bidirectional matching					
	0°	$+5^\circ$	$+10^\circ$	-5°	-10°
The total number of key points	9414	10373	10614	9339	10423
The total number of matching pairs	1145	1213	1164	1070	1254
Correctly matched pairs	759	791	757	752	785
Correct match rate (%)	66.29	65.21	65.03	70.28	62.60
KD-Tree and BBF match					
	0°	$+5^\circ$	$+10^\circ$	-5°	-10°
The total number of key points	9414	10373	10614	9339	10423
The total number of matching pairs	9335	6044	6691	6005	6794
Correctly matched pairs	7582	1802	2112	2015	1845
Correct match rate (%)	81.22	29.81	31.56	33.56	27.16

Rotating the image of a complex text document in the anti-clockwise or clockwise direction can enlarge the image area. Therefore, the number of the detected key points is appropriately increased and the position of the key point is changed. From Table 3, it can be seen that FLANN bidirectional matching performance is better than KD-Tree and BBF matching under the condition of rotation transformation. In order to verify the robustness of the feature under light conversion conditions, the brightness of the original document image is adjusted. The experimental results are shown in Table 4.

Table 4. Two types of FAST (100)+SURF feature points matching results with Uyghur document image illumination transform

FLANN bidirectional matching					
	0	20	40	-20	-40
The total number of key points	9414	9882	9411	8963	9647
The total number of matching pairs	1145	1073	1139	999	952
Correctly matched pairs	759	749	758	803	784
Correct match rate (%)	66.29	69.80	66.55	80.38	82.35
KD-Tree and BBF match					
	0	20	40	-20	-40
The total number of key points	9414	9882	9411	8963	9647
The total number of matching pairs	9335	4892	4931	5591	5789
Correctly matched pairs	7582	3556	1869	3499	3281
Correct match rate (%)	81.22	72.69	37.90	62.58	56.68

The change of illumination is the lightness and darkness of the image. From Table 4, it can be seen that the KD-Tree and BBF matching performance is better than FLANN matching under the key point matching under light conversion conditions, and the matching number is large and the matching rate is high.

5.3 Analysis of Search Results

Due to the large size of the original image collected, the number of feature points obtained by feature extraction is too large, which has a great influence on the number of final matching points. Therefore, in order to assess the performance of the retrieval system, two modifications were made to the overall Uyghur complex document image database by compressing each image and cutting each image into 256 * 256 size, as shown in Fig. 5, and constructed two kinds of Uyghur complex document image database.

In Fig. 5, Fig. 5(b) is sheared image from Fig. 5(a). For the above two improved Uyghur complicated document image databases, based on the number of matches, Euclidean distance and cosine distance similarity measures, the user-specific target document images are retrieved. The retrieval test results are shown in Tables 5 and 6.



(a) Compressed image sample

(b) Sheared image sample

Fig. 5. The sample instance of modified database

Table 5. The statistical results of the sheared Uyghur document image retrieval experiment

Retrieve performance indicators	Match the number of search	Euclidean distance search	Cosine distance search
Retrieval rate	100%	100%	100%
Total search time (s)	1000	854	861
Average index time (s)	1	0.854	0.861

Table 6. The statistical results of the compressed Uyghur document image retrieval experiment

Retrieve performance indicators	Match the number of search	Euclidean distance search	Cosine distance search
Retrieval rate	100%	100%	100%
Total search time (s)	1636	599	607
Average index time (s)	1.636	0.599	0.607

It can be seen from Table 5 above that all three retrieval methods in the cut-structured Uyghur complex document database achieve a retrieval rate of 100%, but the search occupancy time is different. The matching needs to find the nearest neighbor and the next nearest neighbor matching point of each key point, and it need to compare the distance ratio with the first threshold to determine whether they match. Therefore, the system consumes more time than the distance similarity metric retrieval algorithm. The experimental results of compressed Uyghur document image are indicated in Table 6.

It can be seen from Table 6 that the retrieval system based on the number of matches consumes more time than the distance similarity metric retrieval system. For the two databases, matching number based retrieval system, the more the number of image feature points is, the greater the number of matching and the greater the system matching index time. In terms of similarity measure of distance between feature

vectors, although the number of vector images in compressed image is larger, the output target document image can be searched within a shorter time than the cut image.

In this paper, in order to further validate the effectiveness of the FAST and SURF algorithm proposed in this paper, a cut-file image database of 256 * 256 size Arabic, Chinese, Tibetan and natural images is collected, each of which has a size of 1000 frames. The sample example is shown in Fig. 6 below:

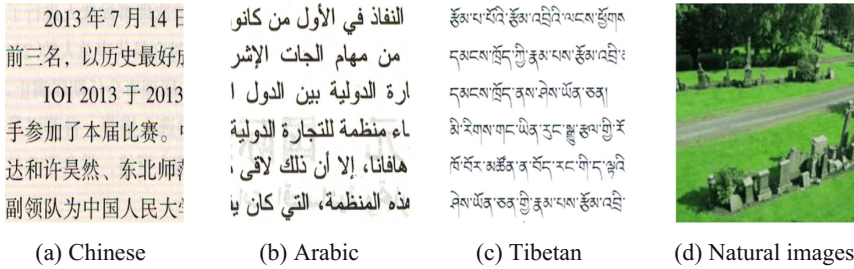
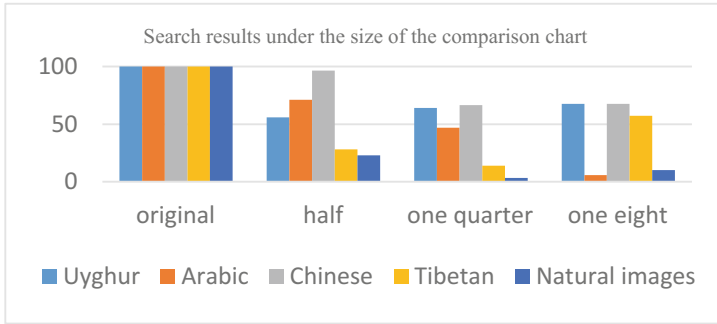


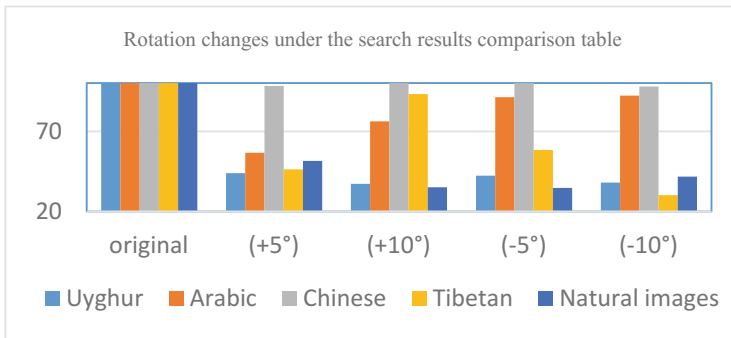
Fig. 6. Comparative experimental database sample diagram

A number of examples of the experimental sample were transformed, such as size (2, 4, 8), illumination (20, 40, 60, -20, -40, -60), and rotation angle (5°, 10°, -5°, -10°) transformation, the retrieval results under different transformations are compared with the retrieval experiments of the Uyghur-cut complex document images, Validate the validity of the retrieval algorithm. The comparison result of the experimental results of retrieving the output target image is shown in Fig. 7(a) to (c).

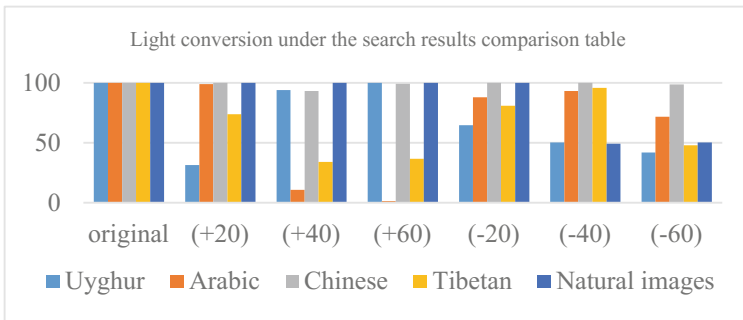
As can be seen from Fig. 7, the letters of Uyghur, Arabic and Tibetan are more irregular than those of Chinese characters, and the differences in the gray-level values of the neighborhood pixels vary greatly. The Chinese language has horizontal and vertical Coherence; the difference in gray value is small. Therefore, the retrieval rate of Chinese query images after many transformations is larger than that of other databases. There are many transformations on the query, and the average indexing time for finding the target image based on the modified retrieval system is 0.013 (0.018), 0.041 (middle), 0.043 (hide) and 0.003 (natural) respectively. Compared with the average retrieval time of the retrieval system of the original features, it is 35.38 (original), 27.81 (a), 15.61 (middle), 16.05 (hide), 123.33 (natural) times. It can be seen that the retrieval system of FAST+SURF features makes it easy to find the target image quickly and accurately, which shows that this article proposes the effective and reliable method of improving ideas.



(a) Comparison of experimental results retrieved in five databases under dimensional transformation



(b) Comparison of experimental results retrieved in five databases under rotation transformation



(c) Comparison of experimental results retrieved in five databases under light conversion

Fig. 7. Comparison of modified FAST and SURF retrieval platform under various transformations experimental results comparison chart

6 Conclusion

In order to make up for the gap in Uyghur complex document image research, this paper proposes a document image retrieval method which is to match retrieval of printed Uyghur composite document images using SURF and the modified SURF features. It is combined the FAST corner detection and SURF description, and two kinds of matching of the selected 64-dimensional feature vectors are performed, and the matching ratio is compared under the condition of size, rotation and light conversion to analyze the performance of the two matching systems. In the end, two retrieval systems were proposed, that is, retrieval scheme based on multiple distance metrics and matching number. The original 100 document images, 1000 compressed images and 1000 document images are retrieved respectively. The matched number of searches takes more time than the distance-based search, but it has a good retrieval rate. Therefore, the focus of the further work is to reduce the retrieval time while ensuring the high retrieval rate of the system.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61563052, 61363064, 61163028), and Hotan Normal University Scientific Research Plan Project (No. 1076515160).

References

1. Xiaoxiao, M.A., Gang, Y.U., Changchun, L.I.: A data processing algorithm for unmanned aerial vehicle images based on SURF and SVM. *J. Henan Polytech. Univ.* (2017)
2. Zhao, L.L., Geng, G.H., Kang, L.I., A-Jing, H.E.: Images matching algorithm based on SURF and fast approximate nearest neighbor search. *Appl. Res. Comput.* **30**(3), 921–923 (2013)
3. Cheon, S.H., Eom, I.K., Ha, S.W., Yong, H.M.: An enhanced SURF algorithm based on new interest point detection procedure and fast computation technique. *J. R.-Time Image Process.* 1–11 (2016)
4. Zhang, H.M., Yang, L., Li, M.L.: Improved SURF algorithm and its application in seabed relief image matching. **12**, 05017 (2017)
5. Chen, J., Han, X.: Image matching algorithm combining FAST-SURF and improved k-d tree nearest neighbor search. *J. Xian Univ. Technol.* (2016)
6. Luo, N., Sun, Q.S., Chen, Q., Ze-Xuan, J.I., Xia, D.S.: Image matching algorithm combining SURF feature point and DAISY descriptor. *Comput. Sci.* **41**, 286–290 (2014)
7. Wang, D., Yan, S., Ming, M.: A fast image matching algorithm based on improved SURF. In: *Tenth International Conference on Computational Intelligence and Security*, pp. 3643–3647 IEEE (2015)
8. Weisheng, A.N., Rangming, Y.U., Yuling, W.U.: Image registration algorithm based on FAST and SURF. *Comput. Eng.* (2015)
9. Dong, H., Han, D.Y.: Research of image matching algorithm based on SURF features. In: *International Conference on Computer Science and Information Processing*, pp. 581–584 IEEE (2012)
10. Batur, A.: Research on Uyghur printed complex document image retrieval based on local feature. *Xinjiang University* (2017)

11. Ren, K., Hu, M.: Color image registration algorithm based on improved SURF. *J. Electron. Meas. Instrum.* (2016)
12. Ma, Y.L.S.: Research on image based on improved SURF feature matching. In: *Seventh International Symposium on Computational Intelligence and Design*, pp. 581–584. IEEE (2015)
13. El-Gayar, M.M., Soliman, H., Meky, N.: A comparative study of image low level feature extraction algorithms. *Egypt. Inform. J.* **14**(2), 175–181 (2013)
14. Huang, L., Chen, C., Shen, H., He, B.: Adaptive registration algorithm of color images based on SURF. *Measurement* **66**, 118–124 (2015)
15. Zheng, C., Jin, W., Fang, F., Tang, C., Ling, Y.: Robust visual tracking algorithm based on structural multi-scale features adaptive fusion in co-training. In: *International Conference on Information Science and Control Engineering*, pp. 588–592. IEEE (2016)
16. Pandey, R.C., Singhm, S.K., Shukla, K.K., Agrawal, R.: Fast and robust passive copy-move forgery detection using SURF and SIFT image features. In: *International Conference on Industrial and Information Systems*, pp. 1–6. IEEE (2015)
17. Darve, N.R., Theng, D.P.: Image processing on eye image using SURF feature extraction. **3297**, 2738–2741 (2015)
18. Horak, K.: Classification of SURF image features by selected machine learning algorithms. In: *International Conference on Telecommunications and Signal Processing*, pp. 636–641 (2017)
19. Shanmugam, B., Rathinavel, R., Perumal, T., Subbaiyan, S.: An efficient perceptual of CBIR system using MIL-SVM classification and SURF feature extraction. *Int. Arab. J. Inf. Technol.* **14**(4), 428–435 (2017)



Deep Word Association: A Flexible Chinese Word Association Method with Iterative Attention Mechanism

Yaoxiong Huang¹, Zecheng Xie¹, Manfei Liu¹, Shuaitao Zhang¹,
and Lianwen Jin^{1,2}(✉)

¹ School of Electronic and Information Engineering,
South China University of Technology, Guangzhou, China
hwang.yaoxiong@gamil.com, zcheng.xie@gamil.com, manfei.l.liu@gamil.com,
z.shuaitao@gamil.com, lianwen.jin@gamil.com

² SCUT-Zhuhai Institute of Modern Industrial Innovation,
South China University of Technology, Zhuhai, China

Abstract. Word association is to predict the subsequent words and phrase, acting as a reminder to accelerate the text-editing process. Existing word association models can only predict the next word inflexibly through a given word vocabulary or a simply back-off N-gram language model. Herein, we propose a deep word association system based on attention mechanism with the following contributions: (1) To the best of our knowledge, this is the first investigation of an attention-based recurrent neural network for word association. In the experiments, we provide a comprehensive study on the attention processes for the word association problem; (2) An novel approach, named DropContext, is proposed to solve the over-fitting problem during attention training procedure; (3) Compared with conventional vocabulary-based methods, our word association system can generate an arbitrary-length string of words that are reasonable; (4) Given information on different hierarchies, the proposed system can flexibly generate associated words accordingly.

Keywords: Word association · Attention mechanism
Recurrent neural network · Chinese · DropContext

1 Introduction

Given a word, phrase, or sentence of arbitrary length, word association requires a machine to predict the following word, phrase, or even sentence that the user would like to express, acting as a reminder to accelerate the text-editing process. Word association is widely used in daily life, such as text input to smartphones, the auto-fill of fields in a web browser, and question/answer systems, which can not only save time and effort but also prevent spelling errors by providing users with a list of the most relevant words. Specifically, when a word is input by a user, the word association system provides a list of candidate words for the user

to select and then updates the associated word list until the user has finished the text editing task.

In the community, methods have been presented for the advancement of word association. Generally, custom systems use a vocabulary or statistical information for word association. PAL [1], the first word association system, predicted the most frequent words that match the given words, completely ignoring any useful context information. Profet [2] (for Swedish) and WordQ [3] (for English) used both word unigrams and bigrams to improve the word association but still suffered from a lack of context information, which would easily lead to syntactically inappropriate words. Considering the inflexibility of the above-mentioned systems, an approach that models the complex context information of the given words is significantly important for the word association problem. In recent years, neural networks [4–6] have demonstrated outstanding ability in language models (LMs). In particular, recurrent neural network LMs (RNNLMs) [7] use long-term temporal dependencies without a strong conditional independence assumption. As RNNLMs become more popular, Sutskever et al. [8] developed a simple variant of the RNN that can generate meaningful sentences by learning from a character-level corpus. Zhang and Lapata [9] have conducted some interesting work and use RNNs to generate Chinese poetry. Furthermore, the ability to train deep neural networks provides a more sophisticated method of exploiting the underlying context information of the sentence, thereby making the prediction more accurate [10].

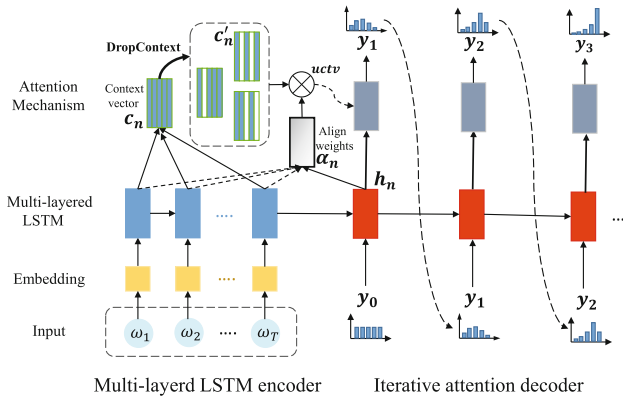


Fig. 1. The proposed word association system consists of two parts: (1) a multi-layered LSTM encoder that learns a hierarchy of semantic features from the input text corpus $w = w_1, \dots, w_T$. and (2) an iterative attention decoder module (with DropContext) that iteratively updates attentions and refines current predictions. Note that y_0 is uniform distribution and y_N predicts the finally results.

LSTM has the ability to remember the past information, but it is quite limited and thus easily leads to prediction failure [11]. Therefore, the attention

mechanism has gained popularity recently in training neural networks [12]; it allows models to learn the alignments between different modalities. The alignments may be between the frame level and text in the speech recognition task [13], or between the source words and translation in the neural machine translation problem [14], allowing the network focus more on the important part of the input. To the best of our knowledge, it is the best choice for natural language processing, e.g., word association problem.

The performance of the current neural network is highly dependent on the greedy learning of model parameters via many iterations based on a properly designed network architecture [15]. During the training phase, it is easy to encounter a problem of over-fitting. Many previous works have been dedicated to solving this problem, e.g., Dropout [16] and DropConnect [17]. Nevertheless, they were not appropriate for the attention mechanism.

Inspired by the aforementioned papers and works, we proposed a word association system that integrates multi-layered LSTM with iterative attention mechanism. The primary contributions of the network can be summarized as follows:

- Attention mechanism is integrated to allow the proposed system to iteratively review context information as well as historical prediction.
- A novel training strategy, namely DropContext, is proposed to alleviate the over-fitting problem during the learning process.
- Given certain information of different hierarchies, the network can generate words of arbitrary length, flexibly. The richer the information provided, the more meaningful words are associated.
- The effectiveness of the proposed system is validated not only by word association on huge Chinese corpus, but also by a poem generating experiment.

The remainder of this paper is organized as follows: Sect. 2 presents a system overview. Section 3 describes the results and performance evaluation of our proposed model. Section 4 summarizes our work.

2 System Overview

Given the training text corpus $\mathbf{w} = w_1, \dots, w_T$ in V , where V is the word dictionary, our word association system f , aims to minimize the loss function $L(\mathbf{w})$ as the negative log probability of correctly predicting all the associated words in the text corpus:

$$L(\mathbf{w}) = -\frac{1}{T} \sum_t \log f(w_t, w_{t-1}, \dots, w_{t-n+1}; \theta) + R(\theta) \quad (1)$$

where T is the total length of the corpus and $R(\theta)$ is a regularization term. Figure 1 describe the detailed architecture of our word association system. Given the training corpus $\mathbf{w} = w_1, \dots, w_T$, we first project each the word w_t in the corpus to a distributed feature vector in the word embedding layer. The multi-layered LSTM then sequentially takes these embeddings as well as the past

hidden state as input and outputs the corresponding context vector. Next, part of the context vector is randomly discarded in the DropContext layer. Finally, the updated context vector and final hidden state of the encoder are fed into the iterative attention decoder, iteratively updates the attentions and refines the current predictions. At the end of the decoder, the fully connected layer with a *softmax* layer will produce a probability distribution over all the words in the vocabulary.

2.1 Word Embedding

Word embedding is the concept of projecting each word in a vocabulary to a distributed word feature vector. Word embedding plays an important role in language modeling [18]. As pointed out by Bengio et al. [4], word embedding helps a network to fight the curse of dimensionality with distributed representations. Through word embedding, semantically similar words, such as ‘cat’ and ‘dog’, are expected to have a similar embedding feature; thus, a training sample that contains ‘cat’ can easily be projected to the case of ‘dog’ and vice versa. Accordingly, word embedding reduces the number of training samples requirement and, more importantly, alleviates the curse of dimensionality. Additionally, word embedding, i.e., the feature vector of each word, is directly learned from the corpora and is naturally trained with neural networks, such as RNN and LSTM, in an end-to-end manner. Given the advantages of word embedding, we used it for word representation at the bottom of our word association system, as shown in Fig. 1, to be jointly trained with the encoder and iterative attention decoder.

2.2 Iterative Attention Decoder (IAD)

In the previous works, the attention-based decoder only ‘glance’ at the source information once, and may make an inappropriate decision. Therefore, we herein employ an iterative attention decoder to our system, giving us a chance to ‘view’ the source information again and refine the current predictions.

From the multi-layered LSTM encoder, we obtain the source hidden state \mathbf{c}_n with a T dimension, which is the same as the number of the input words. Additionally, a current target hidden state \mathbf{h}_n is output from the decoder. Therefore, we can formulate the iterative attention decoder as:

$$\mathbf{y}_n = \text{IAD}(\mathbf{c}_n, \mathbf{y}_{n-1}) \quad (2)$$

where \mathbf{y}_{n-1} is the last output of the IAD system. Note that, when $n = 1$, \mathbf{y}_0 is uniform distribution, and Eq. (2) is updated for N times in the form of a recurrent neural network.

Inspired by the work of Luong [12], we attempt to employ a context vector \mathbf{c}_n that captures relevant input information to aid in the prediction of \mathbf{y}_n , and Eq. (2) can be executed in two step:

(1) We calculate the aligned weights α_n according to the source context vector \mathbf{c}_n and the current target hidden state \mathbf{h}_n :

$$\alpha_n^s = \frac{\exp(\gamma_n^s)}{\sum_{t=1}^T \exp(\gamma_n^t)} \quad (3)$$

where s is the dimension index of both α_n and γ_n . Here, the content-based score γ_n^t can be denoted as:

$$\gamma_n^t = \mathbf{v}_a^\top \tanh(\mathbf{W}_a [\mathbf{h}_n^\top; \mathbf{c}_n^t]) \quad (4)$$

Note that, both \mathbf{v}_a^\top and \mathbf{W}_a are learnable parameters and $[\cdot]$ is the concatenation operation. Subsequently, we adopt the soft attention mechanism [19] where the updated context vector uctv_t is defined as the weighted sum of the source context vector.

$$\text{uctv}_t = \sum_{t=1}^T \alpha_n^t \mathbf{c}_n^t \quad (5)$$

(2) The decoder iteratively updates the attentions and refines the current predictions using a recurrent neural network:

$$\mathbf{y}_n = \text{RNN}(\text{uctv}_t, \mathbf{y}_{n-1}) \quad (6)$$

where the RNN is implemented by a variant of recurrent neural network: Gated Recurrent Unit (GRU) [20]. Compared with LSTM, GRU only contains two gating units that modulate the flow of information, therefore, costing lower consumption.

In the last time step, the fully connected layer with a *softmax* layer will produce a probability distribution over all the words in the vocabulary.

2.3 DropContext (DC)

To overcome the over-fitting problem of attention model, we propose DropContext, a new training strategy, to enhance the efficiency of the learning process of attention model, as shown in the black dotted line in Fig. 1.

Suppose that we have the source context vector \mathbf{c}_n , which is a set of T-dimensional vectors, thus we can update the context vector with DropContext layer:

$$\mathbf{c}'_n = \text{DC}(\mathbf{c}_n) \quad (7)$$

Many attempts have been performed to execute the DropContext layer in our early work, considering the balance between performance and consumption. Our DropContext layer is implemented in two steps. First, we construct a T-dimensional drop-mask \mathbf{M} , which is randomly initialized by the drop-ratio θ :

$$\mathbf{M} = \{m_t = \mathbb{I}\{\zeta > \theta\}, t = 1, 2, \dots, T\} \quad (8)$$

where $\mathbb{I}\{\cdot\} = 1$ when the condition is true and otherwise zero. It is noteworthy that ζ can follow any distribution, e.g., Gaussian distribution or exponential distribution. In this paper, ζ follows a uniform distribution.

Subsequently, we update the source context vector by the element-wise product between \mathbf{c}_n and \mathbf{M} :

$$\mathbf{c}'_n = \mathbf{c}_n \odot \mathbf{M} \quad (9)$$

We have to claim that, after introducing the DropContext layer, we only need to replace \mathbf{c}_n with \mathbf{c}'_n in Eqs. (4) and (5) for the iterative attention decoder.

2.4 Word Association

By integrating the multi-layered LSTM encoder and iterative attention decoder with the prediction layer, from the bottom to the top, we construct a word association system. Formally, the word association system employs the chain rule to model joint probabilities over word sequences:

$$p(w_1, \dots, w_N) = \prod_{i=1}^N p(w_i | w_1, \dots, w_{i-1}) \quad (10)$$

where the context of all the previous words is encoded with LSTM and updated as the predicted word is added. The probability of words is generated through the *Softmax* layer.

The process of associating words of arbitrary length is shown in Fig. 2. Our word association system takes the words of a given sequence as the input. The system then associates the next word by generating a probability distribution over all the given words, as the number upon the black lines shown in Fig. 2. Therefore, we can sort the predicted words in descending order of probability.

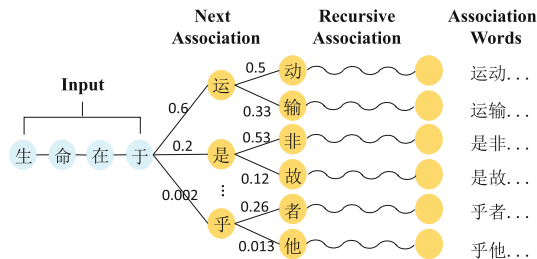


Fig. 2. Schematic diagram of word association. Given the beginning words as input, our word association system predicts a list of candidate words. By recursively adding these candidate words into the input, our word association system can associate sentence of arbitrary length, which is syntactically reasonable. Note that, the numbers upon the black lines represent the probability of the next word.

We adopt the first or top three in the list as the input for the next time step, and associate the following words in the same way. Finally, the system provides candidate associated sentences and their own probability. As described in Fig. 2, after taking the the initial words, our word association system produces a list of candidate words. By associating words in a recursive manner, our word association system manages to generate syntactically reasonable sentences of arbitrary length.

3 Experiments

3.1 Dataset

There is lack of benchmark dataset for the research on word association. Typically, researchers employ their own text corpus to generate the language model. To present an objective evaluation of our word association system, we use two publicly available text corpora, CLDC corpus [21] collected by the Institute of Applied Linguistics, and the Three Hundred Tang Poems (THTP corpus) [22].

For the CLDC corpus, we extracted the available data and filtered extremely rare Chinese characters and characters in other languages. The dataset contains 3455 classes and is divided into two groups, with approximately 70% of data used for training and the remainder for testing. Consequently, the training set contains 59,019,610 words and the test set contains 25,294,119 words.

The THTP corpus consists of 310 poems written by 77 famous poets during the Tang dynasty. For convenience, the punctuation has been removed from the poems. The dataset has approximately 20,000 words and consists of 2,497 classes, including a special symbol that indicates the end of a sentence.

3.2 Implementation Details

The proposed multi-layered LSTM encoder consists of two layers with the hidden size of 512, which are unrolled for 10 steps. Additionally, we also use dropout with probability 0.5 for our LSTMs. Besides, the iterative attention decoder is implemented with an attention-based GRU, whose hidden size is 512. To strike a balance between performance and consumption, we set the maximum iteration N as 3 for the little performance gain with larger N . We train the system in an end-to-end manner using stochastic gradient descent with a weight decay of 0.0005, momentum of 0.9, and gradient clipping set to 10. The initial learning rate is set to 0.1, followed by a polynomial decay of power 0.5.

In this paper, we use the canonical performance metric of language models, namely the perplexity [23], to evaluate our word association system. Perplexity measures the average number of branches of the predicted text, the reciprocal of which can be seen as the average probability of each word. Formally, perplexity is calculated as:

$$\text{perplexity} = \sqrt[\kappa]{\frac{1}{e^{(-\sum \log(p(w)))}}} \quad (11)$$

where $\mathbf{p}(w)$ is the probability of each word in the test set and K is the total number of words that appeared in the test set. It is noteworthy that the word association system with a low perplexity generally performs better than those with a higher perplexity. Besides, we also perform many visualizations of the experiment result, which are more obvious.

3.3 Effectiveness of the DropContext Layer

In this section, we perform a detailed analysis on the performance of our proposed DropContext method. In Table 1, we compare the performance of the system with different drop-ratios. When the drop-ratio is 0.0, no DropContext is available in our model and it is set as the baseline in our experiments. As the drop-ratio increases, the gap between train loss and test loss became smaller, and the system performance improves, i.e., the perplexity and testing loss of the system decreases. We can conclude that, by introducing the DropContext, the over-fitting during the training procedure can be alleviated. However, the system performance decreases afterward when the drop-ratio is larger than 0.4. This is because when the drop-ratio is too large, too much context information will be discarded in the training procedure, which will confuse the decoder and render our system difficult to converge.

Table 1. Influence of drop-ratio

Drop-ratio	0.0 (baseline)	0.2	0.4	0.6	0.8
Train loss	2.63	4.13	4.37	4.45	4.89
Test loss	4.79	3.92	3.86	3.89	4.42
Perplexity	120.36	50.40	47.46	48.91	83.10

3.4 Effectiveness of the Iterative Attention Decoder

In this section, we compare the proposed iterative attention model with a regular LSTM-based model similar to that reported by Merity et al. [5]. The regular LSTM-based model consists of two LSTM layers, with the hidden size of 512, which is the same as the multi-layer LSTM encoder in our system. The difference between the regular LSTM-based model and our model is that each hidden state of the former is followed by the fully connected layer and a softmax layer. This means that once a word is input, the system can only make a ‘decision’ (prediction) once. Note that, both of them are trained with the CLDC corpus.

As shown in Table 2, the regular LSTM-based model (denoted as R-LSTM) achieves a perplexity of 62.80. By introducing the iterative attention decoder, our model (denoted as IA-LSTM) achieves a much lower perplexity of 47.46. We can conclude that adding iterative attention mechanism can lead to a better performance.

Table 2. Perplexity and test loss on the CLDC corpus

Method	Perplexity	Test loss
R-LSTM [5]	62.80	4.14
IA-LSTM	47.46	3.86

Additionally, Fig. 3 shows several examples on how the proposed iterative attention decoder iteratively updates the attentions and refines the current predictions. As we can see, although the model may make an inexact prediction at the beginning, it can update the attentions to focus on the last few words and make a more reasonable prediction. This is also corresponds to common sense that the associated words are more related with their adjacent words [24].

Input	他 接 碗 中 间 吃 劲 就 摸 了	Association (pr)	
$n = 1$		隔 (0.21)	
$n = 2$		她 (0.15)	
$n = 3$		她 (0.33)	GT: 她
Input	学 四 年 级 时 的 唐 山 大 地	Association (pr)	
$n = 1$		提 (0.18)	
$n = 2$		方 (0.24)	
$n = 3$		震 (0.47)	GT: 震

Fig. 3. Examples on how the proposed iterative attention decoder iteratively updates attentions and refines current predictions. At each time-step n , the current association word is listed. Each result is followed by the corresponding probability. Words in red are the most appropriate ones. Note that we use red squares to display the attention weight of each word, the deeper the color is, the greater the weight is.

3.5 Output Visualization of Word Association System

Our word association system generates an arbitrary length string of associated words. The more information is provided to the system, the more meaningful words will be generated. As shown in Fig. 4(a), given different numbers of words as beginning, our system associates sentences with completely different meanings. When only less information is available, the system randomly generates the sentences. However, when given more detailed information, the system associates a sentence that is quite relevant to the given words. In Fig. 4(b), the words in the first line are the input to the word association system and the subsequent lines are the associated sentences of different lengths. Note that regardless of the length of the associated sentences, they are reasonable and meaningful.

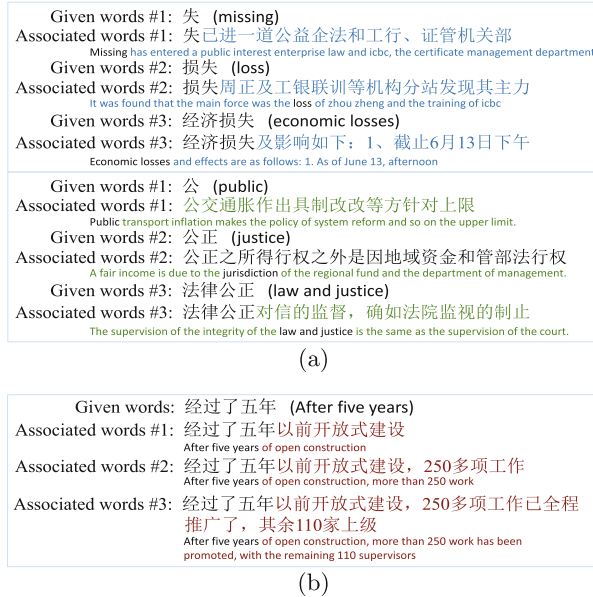


Fig. 4. Output of word association system. In (a), there are three kinds of inputs to the system, ordered by the amount of information in Chinese. In (b), there are three different lengths of output for the same input to the system. The associated sentence is syntactically reasonable for any arbitrary length. The tiny English sentence right below the Chinese sentence is the corresponding translation.



Fig. 5. Result of the model trained with the THTP corpus (shown in poetry format). Given arbitrary words, our system associates a meaningful poem with the Tang poem style.

3.6 Generating Poems

To verify the significance of our word association system, an poetry generating experiment is conducted using the THTP corpus. In the testing phase, a contiguous piece of a sentence is input to the word association system, and the system attempts to associate a poem accordingly.

To generate a poem, as shown in Fig. 5, arbitrary words are given to the association system. Staring with the given words, the system produces a meaningful poem of the Tang poem style. Furthermore, the associated poem is incredibly ‘real’ that it is difficult to distinguish whether it is one of the original poems in the dataset.

4 Conclusion

In this paper, we presented a flexible Chinese word association method which consists of a multi-layer LSTM encoder and an iterative attention decoder. Experiments show that the attention mechanism can improve the performance of Chinese word association system. Besides, the iterative attention decoder implemented in our system can iteratively uses its previous prediction to update attentions and to refine current predictions. Moreover, by adopting the DropContext layer in our proposed model, over-fitting can be avoided during the training procedure, which is proved to be better converged. Additionally, we showed that our system can generate syntactically reasonable associated words of arbitrary length and tends to associate more meaningful yet relative words when given more context information. Finally, we verify the significance of our word association system through an interesting poem generating experiment.

Acknowledgement. This research is supported in part by GD-NSF (no. 2017A030312006), the National Key Research and Development Program of China (No. 2016YFB1001405), NSFC (Grant No.: 61673182, 61771199), and GDSTP (Grant No.: 2014A010103012, 2017A010101027), GZSTP(no. 201607010227).

References

1. Swiffin, A.L., Pickering, J.A., Arnott, J.L., Newell A.F.: PAL: an effort efficient portable communication aid and keyboard emulator. In: ACRT, pp. 197–199 (1985)
2. Carlberger, A., Carlberger, J., Magnuson, T., Hunnicutt, M.S., Palazuelos-Cagigas, S.E., Navarro, S.A.: Profet, a new generation of word prediction: an evaluation study. In: Proceedings, ACL Workshop on Natural Language Processing for Communication Aids, pp. 23–28 (1997)
3. Shein, F., Nantais, T., Nishiyama, R., Tam, C., Marshall, P.: Word cueing for persons with writing difficulties: WORDQ. In: Proceedings of CSUN 16th Annual Conference on Technology for Persons with Disabilities (2001)
4. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. *J. Mach. Learn. Res.* **3**(Feb), 1137–1155 (2003)
5. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing LSTM language models. *CoRR*, abs/1708.02182 (2017)
6. Yang, Z., Dai, Z., Salakhutdinov, R., Cohen, W.W.: Breaking the softmax bottleneck: a high-rank RNN language model. In: ICLR (2018)
7. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH, vol. 2, pp. 3 (2010)
8. Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: ICML, pp. 1017–1024 (2011)
9. Zhang, X., Lapata, M.: Chinese poetry generation with recurrent neural networks. In: EMNLP, pp. 670–680 (2014)
10. Hinton, G., Deng, L.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29**(6), 82–97 (2012)
11. Jenckel, M., Bukhari, S.S., Dengel, A.: Training LSTM-RNN with imperfect transcription: limitations and outcomes. In: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, pp. 48–53. ACM (2017)

12. Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. CoRR, abs/1508.04025 (2015)
13. Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: NIPS, pp. 577–585 (2015)
14. Dzmitry B., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
15. Yang, W., Jin, L., Tao, D., Xie, Z., Feng, Z.: DropSample: a new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten chinese character recognition. Pattern Recognit. **58**, 190–203 (2016)
16. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors 3 July 2012. CoRR, abs/1207.0580 (2016)
17. Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R.: Regularization of neural networks using DropConnect. In: ICML (2013)
18. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: ACL, pp. 310–318. ACL (1996)
19. Show, attend and tell: neural image caption generation with visual attention. In: ICML (2015)
20. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR, abs/1406.1078 (2014)
21. Chinese linguistic data consortium. The Contemporary Corpus developed by State Language Commission P. R. China, Institute of Applied Linguistics (2009). <http://www.chineseldc.org>. Accessed 22 Oct 2016
22. Wikipedia. Three Hundred Tang Poems (2018). https://en.wikipedia.org/wiki/Three_Hundred_Tang_Poems
23. Jurafsky, D., James, H.: Speech and language processing an introduction to natural language processing, computational linguistics, and speech (2000)
24. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based N-gram models of natural language. Comput. Linguist. **18**, 467–479 (1992)

Face Recognition and Analysis



Face Recognition Based on Multi-view Ensemble Learning

Wenhui Shi and Mingyan Jiang^(✉)

School of Information Science and Engineering, Shandong University,
Qingdao 266237, China
jiangmingyan@sdu.edu.cn

Abstract. Face recognition is an important research area in human-computer. To solve the problem about the inaccuracy and incompleteness of feature extraction and recognition, an ensemble learning method on face recognition is proposed in this paper. This method is a combination of a variety of feature extraction and classification ensemble technology. In feature extraction, wavelet transform and edge detection are used for extracting features. In classification recognition, the K nearest neighbor (KNN) classifier, wavelet neural network (WNN) and support vector machine (SVM) are used for preliminary identification. Each classifier corresponds to a feature method and then the classification of the three views are constructed. The final output results are integrated by voting strategy. Experimental results show that this method can improve the identification rate compared with the single classifier.

Keywords: Face recognition · Multi-view · Feature extraction
Ensemble learning · Voting

1 Introduction

Biometric authentication is a kind of personal identification, which is performed using the characteristics of the human body by computer [1]. Face recognition is an example of using biometric to authenticate. Compared with the other biological features such as iris and fingerprint, the acquisition of face image is more convenient and the equipment is more hidden. As a method of using effective information for identification, face recognition has been widely used in many aspects in the past few decades [2].

In the past few decades, face recognition technology has become more and more concerned by researchers in the world. Especially since recent years, the research and application of face recognition technology has made great progress and a large number of academic papers have been published every year [3]. Some websites and APP use face login and face registration. In the last year, the iPhone X produced by Apple Inc uses the face recognition function. At the same time, there are many commercial face recognition system into the market, such as law enforcement advanced video surveillance, surveillance portal control and so on.

As a complex pattern recognition problem [4], face recognition involves many disciplines, including image processing, mathematics, physiology, computer vision,

etc. Because of the influence of many factors, face recognition is a technique with high complexity. In order to deal with these complex problems, some good methods are needed in feature extraction and recognition.

How can we extract features of the face accurately? Feature extraction is a key step in face recognition, which determines the results of recognition directly. It is affected by many aspects, including posture, expression, age, etc. [5]. The extracted features should reflect the identity as much as possible. It is inaccurate if we just use a single method to extract feature, then the recognition results are unsatisfactory. We can obtain more complete features by combining a variety of methods to extract features and lay the foundation for the recognition of the back. There are many methods to extract features. Reference [6] has proposed a method based on Canny operator to detect edges. The wavelet transform has a good time-frequency localization properties, so it is suitable for image processing. Reference [7] used stationary wavelet transform (SWT) to extract features from MR brain images.

In addition to feature extraction, the design of classifier also has great influence on the performance of face recognition algorithm. Different classification can make different results. In general, feature recognition usually adopts single classifier such as SVM [8], neural network and so on. However, it is unable to ensure the accuracy and stability of the results only relying on a single classifier for recognition. Thus, multiple classifiers are combined by the integration technology [9] to improve the generalization ability and reliability of the classification system. When designing an integrated system, the selection of a single classifier is critical, which is the first factor affecting the performance. The selected single classifier need to be stable and diverse. Secondly, the strategy of ensemble method is the second influencing factor. Reference [10] has used weighted majority voting classifier combination for relation extraction from biomedical sentences.

We proposed a method of ensemble learning for face recognition in this paper. Canny operator, wavelet transform were used to extract features of the images itself and transformation domain in this method [11]. Then we utilized three simple and common classifiers the KNN, SVM and WNN to identify. A classifier combined a feature extraction method and the classification of the three views were constructed subsequently. The voting strategy was adopted to integrate decision finally.

2 Classifier

The classifier can affect the final result, and we will introduce several classifiers used in this chapter.

2.1 KNN (k Nearest Neighbor Classifier)

The K nearest neighbor classifier is an effective classifier in pattern recognition [12].

It uses the known categories of the nearest neighbor samples to judge the unknown sample, which is suitable for dealing with overlapping or crossover samples. Specific steps are as follow: Calculate the distance of the sample (also as known similarity) to be sorted and the known samples in the feature space. This is the key to the method. Then

find the k samples that are closest to the unknown sample. Count the category of k samples, and find the category which has the largest number. Finally classify the unknown sample into this category.

2.2 SVM (Support Vector Machine)

Support Vector Machine [8] has great advantages in solving nonlinear classification. The basic principle is to transform the input space into the high-dimensional space by non-linear mapping. The samples can be divided linearly, in which case the optimal interface can be obtained.

Suppose the known training sets are $C = \{(x_i, y_i)\}$, where $x_i \in R^n$, $y_i \in \{-1, 1\}$, ($i = 1, 2, \dots, l$). For linear transformation of x , the equation of linear separation is $w^T x_i + b = 0$. The surface that satisfies $y_i(w^T x_i + b) - 1 \geq 0$. The surface that satisfies $y_i(w^T x_i + b) - 1 \geq 0$ and $\|w\|^2$ is the optimal classification surface.

Under this condition, it can be transformed into an optimization problem:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^j \sum_{j=1}^l y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^l \alpha_j \quad (1)$$

Then the discriminant function can be determined according to the optimal solution α and the threshold b determined from the training samples:

$$f(x) = \text{sgn}\left(\sum_{x_i \in \mathcal{S}_i}^n \alpha_i y_i K(x_i, x) + b\right) \quad (2)$$

Where α is Lagrange multiplier, $K(x_i, x)$ is the kernel function.

We can construct multiple classifiers to solve the multiple class problems. On the one hand, the SVM multi-class classifier can be realized by combining multiple two-class classifiers. On the other hand, the objective function can be modified to merge the problem of multiple classification surfaces into an optimization problem.

2.3 WNN (Wavelet Neural Network)

Wavelet neural network is the combination of wavelet transform and artificial neural network. It not only includes the local time-frequency characteristics and multi-scale decomposition characteristics of wavelet transform, but also contains the self-learning, adaptive and fault-tolerant ability of neural network [13]. Simply speaking, the wavelet function is used to replace the function in hidden layer on the basic of the BP neural network. The signal of wavelet neural network is transmitted forward, and error is transmitted backward at the same time.

The output of WNN is given by:

$$y(k) = \sum_{j=1}^l \omega_{jk} * h_j\left(\left(\sum_{i=1}^k \omega_{ij} x_i - b_j\right) / a_j\right) \quad (3)$$

Where h_j is the mother wavelet function; a_j is the scaling factor and b_j is the translation factor.

The error function is used as the fitness function to verify the degree of parameters correction:

$$Error = \sum_{k=1}^m (y(k) - D(k))^2 / 2 \quad (4)$$

Where $D(k)$ is the expected output of the network.

We need to adjust the parameters according to the error. There are many methods for parameter revision and the gradient descent method is the most common in the wavelet neural network. However, it converges slowly and is easy to fall into the minimum. In this paper, we use the method of adding momentum item to modify the local parameters:

$$\omega_{ij}(i+1) = \omega_{ij}(i) + \Delta\omega_{ij}(i+1) + k(\omega_{ij}(i) - \omega_{ij}(i-1)) \quad (5)$$

$$a_j(i+1) = a_j(i) + \Delta a_j(i+1) + l * (a_j(i) - a_j(i-1)) \quad (6)$$

$$b_j(i+1) = b_j(i) + \Delta b_j(i+1) + l * (b_j(i) - b_j(i-1)) \quad (7)$$

3 Multi-view Ensemble Learning

3.1 The Multi-view Ensemble Learning Model

The classification technique of ensemble learning is a combination of multiple classifiers to enhance the reliability and generalization of system. In order to identify face images better, different feature extraction methods and identification classifiers are adopted in this study. The recognition model is shown in Fig. 1.

View 1 (LDA + KNN)

In this view, LDA is used to obtain the features with fewer dimensions and then we use KNN to identify the features.

LDA [11] also called Fisher Linear Discriminant, is a supervised algorithm that reduces the dimension. The principle is: The data with label can be projected to a lower dimension by mapping, the projecting points in the same class are as close as possible, and the distance between different classes are as large as possible. Thus the data after projection can be distinguish by category.

View 2 (Edge detection + SVM)

As an edge detection method, Canny operator has good anti-noise performance and detection accuracy [14]. In this view, we use Canny operator to obtain the edge information of the image. The gradient amplitude and direction of images can be calculated after the Gaussian smoothing. We can use non-maximal suppression and double threshold processing to get the final edge.

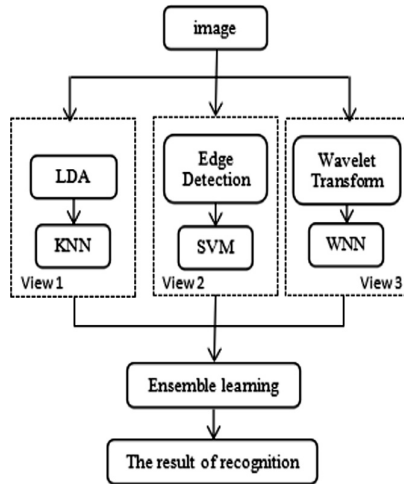


Fig. 1. The multi-view ensemble learning model.

After obtaining the edge features, we utilize SVM to classify to get the results. Face recognition is a typical multi-class identification problem. The support vector machine has strong generalization ability and good recognition rate for face recognition on pattern classification.

View 3 (Wavelet Transform + WNN)

Firstly, we used the wavelet transform to deal the image. It is well known that the wavelet transform has the ability of multi-scale expression. We use the two-dimensional discrete wavelet transform in this model and it can be realized by one-dimensional wavelet transform. The transformed image is divided into four parts: The LL part is an image with approximate coefficient that contains the major feature of the image. LH, HL and HH are images with detail coefficient that contain the details of the image. Among them, HH has high frequency both in horizontal direction and vertical direction, LH has low frequency in horizontal direction and high frequency in vertical direction, HL has high frequency in horizontal direction and low frequency in vertical direction.

In WNN, we adopt the three-layer feed-forward neural network shown in Fig. 2. This kind of wavelet neural network has one hidden layer.

3.2 Ensemble Learning Method

When designing an integrated system, multiple classifiers need to be integrated to achieve good integration [15]. And the selection of ensemble method affects the final results. There are many methods to integrate. Among them, the bagging as the most intuitive method has a surprisingly good performance. Table 1 shows the voting method.

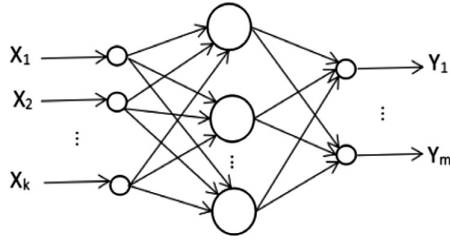


Fig. 2. The structure of WNN for MIMO system.

Table 1. Algorithm: voting method.

Algorithm: voting method

Input: Given N unlabeled data
 Use the T classifiers to identify the data and the classification results are:
 $C = \{C_1, C_2, \dots, C_M\}$
 where M is the total number of categories.
 Suppose

$$\varphi_{i,j} = \begin{cases} 1 & \text{if } T_i \text{ is recognized as } C_j \\ 0 & \text{otherwise} \end{cases}$$
 represents the vote result of C_j by the classifier T_i .
 The total voting result for each classifier is obtained by:

$$\phi_j = \sum_{i=1}^T \varphi_{i,j} \quad j = 1, 2, \dots, M$$
 Select the class with the highest total voting result as the final category:

$$C_j = \max \phi_j$$

The voting results can be divided into three categories:

The Unanimous Voting: the result of ensemble learning is the class on which all classifiers are consistent. In other words, if KNN, SVM and WNN are identified as the same output, the final result will be this output.

The Plurality Voting: the ensemble result is the class on which more than one half of the classifiers are consistent. For example, if KNN and WNN are identified as the same output A, the SVM is identified as another output B, then the final result is output A.

The Weighted Voting: If the outputs of the three classifiers are different, the output of the classifier which has the highest recognition rate will be the final result. In this experiment, the recognition rate of WNN is higher than KNN and SVM, so the final result is derived from the WNN.

4 Experiments

In order to verify the feasibility of this algorithm, the experiment is carried out in ORL face database. In this paper, we select 320 images of human face in ORL face database consisting of 32 people. The size of each image is 92×112 pixels with a grayscale of 256. Some of the face images are shown in Fig. 3. We select 5 images of each person that are 160 images as the training samples, and the rest of the images are used as test samples.



Fig. 3. Some sample images in ORL database.

In view 1, we obtain 160 dimensional features after the process of LDA. Then we classify the features according to the K nearest neighbor classifier. There are many methods to calculate the distance between the sorted samples and the known samples, such as the Euclidean distance, the Minkowski distance, the Manhattan distance, and so on. Here, we use the Euclidean distance. And we choose 5 neighbors through the experiment finally.

In view 2, the two-dimensional Gaussian function is served as the noise filter in Canny operator. Then we use the LIBSVM-FarutoUltimate toolbox to construct SVM classification after obtain the edge features. This toolbox provides a series of auxiliary functions for parameter searching, processing and result visualization, which are more convenient to use. Different inner product kernel functions in SVM will form different algorithms. In this model, we use sigmoid kernel function.

In view 3, in order to improve the speed, we adopt the wavefast function in wavelet toolbox. The Fig. 4 is the original image and its wavelet transform. The left image is the original image, and the right image is a 1-scale wavelet transform. As can be seen from the figure, the low frequency part retains the approximate information, and the high frequency part retains some edge information and noise. In the wavelet neural network, the morlet wavelet shown in Fig. 5 is exploited as the activation function in hidden layer.



Fig. 4. The original image and its wavelet transform.

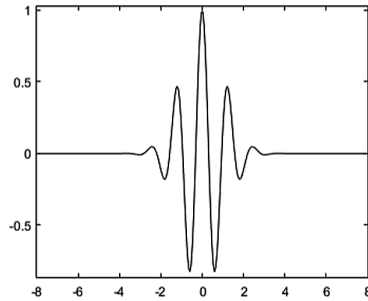


Fig. 5. Morlet wavelet function.

Table 2 shows the recognition rate of the ensemble learning method proposed in this paper. The single classifier is also used to identify the same data set. It can be seen that the recognition rate is lowest when using KNN. Compared with the single classifier, the recognition rate of ensemble learning method has been improved obviously.

Table 2. Average accuracy rates on ORL.

Methods	Rates (%)
LDA + KNN	86.88
Edge detection + SVM	90
Wavelet transform + WNN	91.88
Ensemble learning	96.88

In order to increase the contrast, we select some images from the FERET database randomly. Each person has 7 different images. In the experiment, four images of each person are used for training set randomly and the remaining 3 images of each person are used for testing set. Some of the face images are shown in Fig. 6. Table 3 shows the recognition rate on this small data set. We can see that the ensemble learning method has the highest recognition rate.

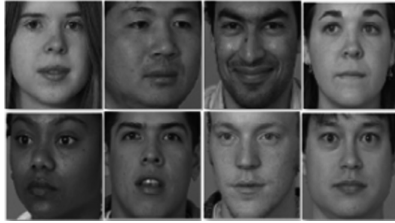


Fig. 6. Some sample images in FERET database.

Table 3. Average accuracy rates on FERET.

Methods	Rates (%)
LDA + KNN	88.54
Edge detection + SVM	92.71
Wavelet transform + WNN	93.75
Ensemble learning	97.91

5 Conclusions

In this paper, combining multiple feature extraction and classification techniques, we propose a method of multi-view ensemble learning in face recognition. A variety of methods are used to extract features, which avoids the incompleteness of information and represents the feature more fully. The classification uses SVM, KNN, WNN as the base classifier to identify respectively. Multi-view results are integrated with voting strategy to ensure the accuracy of identification results. The experimental results show that our method has impressive recognition accuracy on face database.

Future work includes implementing the parallelism of the algorithm to compensate the complexity. In addition, there is a need for further reduction in running time. I believe that face recognition technology will be more perfect, stable and powerful in the near future.

Acknowledgement. This research was financially supported by the National Science Foundation of China (Grant No. 61771293).

References

1. Murillo-Escobar, M.A., Cruz-Hernández, C., Abundiz-Pérez, F., López-Gutiérrez, R.M.: A robust embedded biometric authentication system based on fingerprint and chaotic encryption. *Expert Syst. Appl.* **42**(21), 8198–8211 (2015)
2. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823. IEEE Press, Boston (2015)

3. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: a survey. *IEEE Trans. Affect. Comput.* **4**(1), 15–33 (2013)
4. Zipfel, C.: Plant pattern-recognition receptors. *Trends Immunol.* **35**(7), 345–351 (2014)
5. Drira, H., Ben Amor, B., Srivastava, A., Daoudi, M., Slama, R.: 3D face recognition under expressions, occlusions, and pose variations. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(9), 2270–2283 (2013)
6. Zhang, X., Zhang, Y., Zheng, R.: image edge detection method of combining wavelet lift with Canny operator. *Procedia Eng.* **1**(15), 1335–1339 (2011)
7. Zhang, Y., Dong, Z., Liu, A., Wang, S., Ji, G., Zhang, Z., Yang, J.: Magnetic resonance brain image classification via stationary wavelet transform and generalized eigenvalue proximal support vector machine. *J. Med. Imaging Health Inform.* **5**(7), 1395–1403 (2015)
8. Gu, B., Sheng, V.S., Tay, K.Y., Romano, W., Li, S.: Incremental support vector learning for ordinal regression. *IEEE Trans. Neural Netw. Learn. Syst.* **26**(7), 1403–1416 (2014)
9. He, Y., Wu, H., Zhong, R.: Face recognition based on ensemble learning with multiple LBP features. *Appl. Res. Comput.* **35**(1), 292–295 (2018)
10. Remya, K.R., Ramya, J.S.: Using weighted majority voting classifier combination for relation classification in biomedical texts. In: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICT), pp. 1205–1209. IEEE Press, Kanyakumari (2014)
11. Martis, R.J., Acharya, U.R., Min, L.C.: ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. *Biomed. Signal Process. Control* **8**(5), 437–448 (2013)
12. Zhao, Y., You, X., Yu, S.: Multi-view manifold learning with locality alignment. *Pattern Recognit.* **78**, 154–166 (2018)
13. Ardestani, M., Zhang, X., Wang, L., Lian, Q., Liu, Y.: Human lower extremity joint moment prediction: a wavelet neural network approach. *Expert Syst. Appl.* **41**(9), 4422–4433 (2014)
14. Guiming, S., Jidong, S.: Remote sensing image edge-detection based on improved Canny operator. In: 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN), pp. 652–656. IEEE Press, Beijing (2016)
15. Xu, W., Shen, Y., Bergmann, N.: Sensor-assisted multi-view face recognition system on smart glass. *IEEE Trans. Mob. Comput.* **17**(1), 197–210 (2018)



Conditional Face Synthesis for Data Augmentation

Rui Huang^{1,2,3}, Xiaohua Xie^{1,2,3}(✉), Jianhuang Lai^{1,2,3},
and Zhanxiang Feng^{1,2,3}

¹ School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
xiexiaoh6@mail.sysu.edu.cn

² Guangdong Key Laboratory of Information Security Technology, Guangzhou, China

³ Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China

Abstract. Conditional face synthesis has been an appealing yet challenging problem in computer vision. It has a wide range of applications. However, few works attempt to leverage the synthesized face images for data augmentation and improve performance of recognition model. In this paper, we propose a conditional face synthesis framework that combines a variational auto-encoder with a conditional generative adversarial network, for synthesizing face images with specific identity. Our approach has three novel aspects. First, we propose to leverage the synthesized face images to do data augmentation and train a better recognition model. Second, we adopt multi-scale discriminators to enable high-quality image generation. Third, we adopt identity-preserving loss and classification loss to ensure identity invariance of synthesized images, and use feature matching loss to stabilize the GAN training. With extensive qualitative and quantitative evaluation, we demonstrate that face images generated by our approach are realistic, discriminative and diverse. We further show that our approach can be used for data augmentation and train superior face recognition models.

Keywords: Conditional face synthesis · Data augmentation
Generative adversarial network

1 Introduction

Since deep learning is data-driven methods, ample data have been utilized to train high performance models in various computer vision tasks, such as image classification [14], face recognition [20] and so on. However, There are many realistic scenarios which limit data are available. The deep neural networks is prone to overfit in the training set and yield poor generalization ability.

X. Xie—This project is supported by the Natural Science Foundation of China (61702566, 61672544) and Top-top Scientific and Technical Innovative Youth Talents of Guangdong special support program (No. 2016TQ03X263).



Fig. 1. Synthesized faces. Given an identity label and a randomly sampled latent vector, generating diverse face images with specific identity.

As a generative problem in computer vision, image synthesis is appealing yet challenging. In the past few years, it has received great research interests and has a wide range of applications, such as image generation [3], face attribute editing [5], image translation [19], face completion [4], image super-resolution [15] among others. However, exist works seldom utilize the synthesized images for further recognition or detection tasks, like face recognition. In this work, we propose to leverage the synthesized face images for data augmentation and improve performance of recognition model.

Traditional data augmentation techniques [14], like translation, rotation, horizontal flip and random crop, can introduce some known intra-class variance. These techniques are proved to be valid, but the transformations are limit and constant. We argue that we can learn a generative model to do data augmentation. Through a trained model, we can generate images with more abundant intra-class variance.

This work mainly focuses on conditional face synthesis, i.e., given an identity label and a randomly sampled latent vector, generating face images with specific identity, as illustrated in Fig. 1. We hope that synthesized face image have following characteristics: (1) Images are photo-realistic, diverse and rich in intra-class variance, such as pose, illumination and expression. (2) Images must preserve identity so that they can be used for face recognition.

Inspired by CVAE-GAN [3], we propose a conditional face synthesis framework that combines a variational auto-encoder with a conditional generative adversarial network, for synthesizing face images with specific identity. However, we find that using traditional discriminator structure and adversarial loss function will lead to many problems. First, the GAN training is unstable because of the gradient vanishing problem. Then the quality of synthesized face images are poor. Moreover, synthesized images are easy to loss identity information which is the key for recognition task. To tackle these problems, we first adopt multi-scale discriminators [19] to enable high-quality image generation. Specifically, we use multiple discriminators that have the same network structure but handle different image scales to improve image quality. Second, we adopt identity-preserving loss and classification loss to ensure identity invariance of synthesized images. Third, we use feature matching loss to stabilize the GAN training.

In summary, This paper makes the following contributions.

1. We propose a conditional face synthesis framework that combines a variational auto-encoder with a conditional generative adversarial network, for synthesizing face images with specific identity. Furthermore, we leverage the synthesized face images to do data augmentation and train a better recognition model.
2. We adopt multi-scale discriminators to enable high-quality image generation, adopt identity-preserving loss and classification loss to ensure identity invariance of synthesized images, and use feature matching loss to stabilize the GAN training.
3. With extensive qualitative and quantitative evaluation, we demonstrate that face images generated by our approach are realistic, discriminative and diverse. Furthermore, we show that our approach can be used for data augmentation and train superior face recognition models.

2 Related Work

In the last few years, deep generative models have made significant breakthroughs in face synthesis. Since deep neural network is able to learn powerful feature representations, These methods can capture complex data distributions and generate more realistic images than traditional methods. The mainstream face generative models can be roughly divided into two categories: Variational Auto-encoder (VAE) [6] and Generative Adversarial Network (GAN) [2, 3, 7, 11, 19].

Variational Auto-encoder (VAE) [6] is one of the most popular approaches to unsupervised learning of complicated distributions. It is actually a pair of connected networks: an encoder and a decoder/generator. The encoder maps an input image to a latent representation, and the decoder/generator converts it back to the original input. With the reparameterization trick [6], VAE is able to be optimized using stochastic gradient descent. However, since VAE uses l2 loss or l1 loss as reconstruction loss, the images generated by VAE often suffer from fuzzy effect.

Generative Adversarial Network (GAN) has attracted significant attention on the research of deep generative models [2, 3, 7, 11, 19]. GAN consists of a discriminator D and a generator G that D and G compete in a minimax two-player game. Huang et al. [11] proposed a Two-Pathway Generative Adversarial Network (TP-GAN) for synthesising photorealistic frontal view face from profile. This work perceives global structures and local details simultaneously. To improve the quality of generated images, Wang et al. [19] adopted multi-scale generator and discriminator architectures, as well as improved adversarial loss. Arjovsky et al. [2] adopted Earth Mover Distance to measure the similarity between two distributions, which stabilize the GAN training and alleviate mode-collapse phenomenon to a certain extent.

Bao et al. [3] presents variational generative adversarial networks (CVAE-GAN) for synthesizing images in fine-grained categories. Their work is related to our work. But compared with their method, our method has the following differences: (1) We introduce identity-preserving loss to ensure identity invariance

of synthesized images. (2) We adopt multi-scale and multi-task discriminators to enable high-quality image generation.

3 Approach

In this section, we first review the vanilla generative adversarial network (Sect. 3.1). Then we introduce the overall of our conditional face synthesis framework (Sect. 3.2). Next, we describe the detailed network architecture of our method (Sect. 3.3). Finally, we introduce the object functions of the proposed method and the training pipeline (Sect. 3.4).

3.1 Generative Adversarial Network

Generative Adversarial Network (GAN) consists of a discriminator D and a generator G that D and G compete in a minimax two-player game. Specifically, a discriminator D tries to distinguish a real image from a synthesized one, while a generator G tries to capture the data distribution and generate images that can fool D . Specifically, D and G play the following two-player minimax game with value function $V(D, G)$:

$$\min_G \max_D V(D, G) = E_{x \sim p_d(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

3.2 Problem Formulation

In this section, we elaborate the proposed conditional face synthesis framework. Given an identity label c and a randomly sampled latent vector z , our goal is to generate face images with specific identity. The overall framework is visualized in Fig. 2. Our method consists of four components: (1) encoder network E , (2) generative network G , (3) discriminative network D , (4) identity-preserving network FR . Next, we introduce the function of each component.

The encoder network E is similar to the encoder of VAE. By learning a distribution $P(z|x)$, E first maps the image x to the mean and covariance, and then obtains the latent representation z by reparameterization trick [6]. The generative network G is similar to the generator of conditional GAN [16]. By learning a distribution $P(x|z, c)$, G generates a image $G(z, c)$ given a identity label c and a randomly sampled latent vector z . Specifically, The latent representation z_{encode} is obtained from E and the latent representation z_{random} is sampled from normal gaussian distribution. The generated images are x_{encode} and x_{random} , respectively. Different from the traditional discriminator, we adopt multi-task learning for discriminative network D . D distinguishes real/fake faces and performs identity classification, i.e., estimate the posterior $P(c|x)$, simultaneously. In order to leverage synthesized face images for face recognition task, it is crucial to keep the identity invariance of synthesized images. We thus introduce an identity-preserving network FR to ensure identity invariance through feature matching manner.

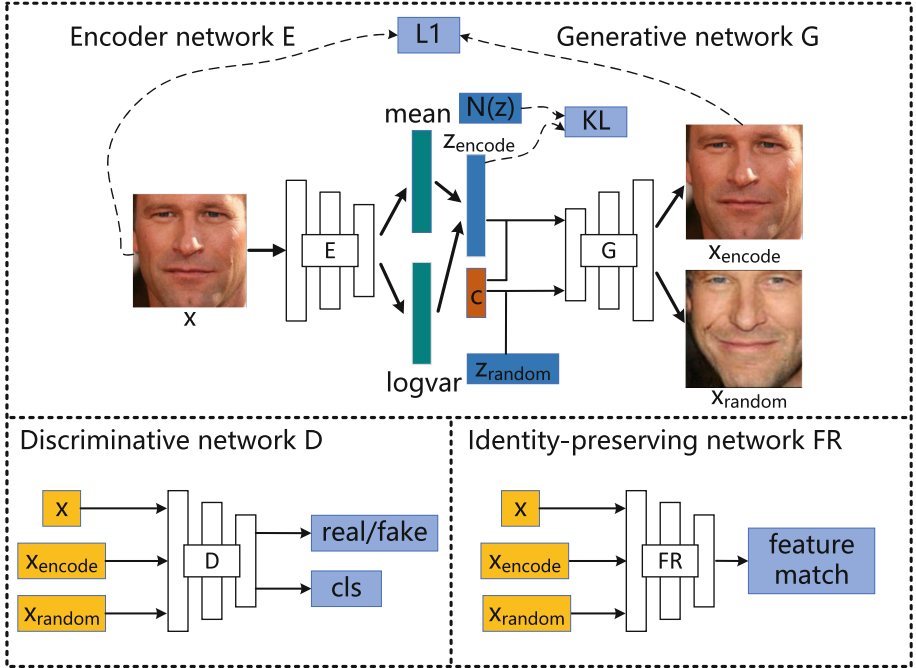


Fig. 2. The overall framework of our conditional face synthesis method. Our method consists of four components: (1) encoder network E, (2) generative network G, (3) discriminative network D, (4) identity-preserving network FR.

3.3 Network Architecture

The encoder network E consists of four residual blocks with 2x downsampling. The architecture of residual block is shown in Fig. 3. The generative network G consists of 6 deconvolution layers with 2x upsampling.

The discriminative network D consists of six convolution layers with 2x downsampling. Different from traditional GAN that the discriminator only distinguishes real/fake images, we adopt multi-task learning for D. D distinguishes real/fake faces and performs identity classification simultaneously. Specifically, our discriminator produces two probability distributions, i.e., $D : x \rightarrow \{D_{src}(x), D_{cls}(x)\}$, where $D_{src}(x)$ is the probability that discriminator regards the input as true, and $D_{cls}(x)$ is the posterior for identity classification.

Recent work [19] shows that the discriminator needs a large receptive field to produce a high-quality image. Inspired by [19], we introduce multi-scale discriminators to distinguish real/fake images from different scales. As illustrated in Fig. 4, we use two discriminators D^1 and D^2 . Each has the same network structure but handle images from different scales. The discriminator with coarse scale has large receptive field, which helps to keel global structure information. The discriminator with fine scale has small receptive field, which helps to produce details.

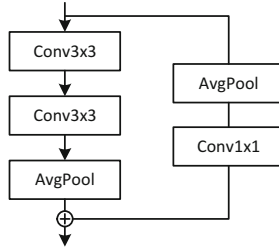


Fig. 3. The architecture of residual block [23].

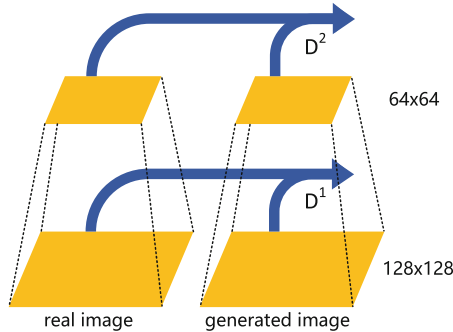


Fig. 4. Illustration of multi-scale discriminators.

3.4 Object Function

The object function used in our approach is a weighted sum of five individual loss functions. Next, we will describe each loss function, respectively.

Adversarial Loss. Traditional GAN uses cross entropy as adversarial loss. Actually at the early stage of training, the distributions of real/fake images may not overlap with each other. So it is easy for D to distinguish real/fake images. This leads to gradient vanishing problem [1]. To stabilize the training process, we use Wasserstein GAN with gradient penalty [2, 8] as adversarial loss. It takes the form:

$$\mathcal{L}_{adv}(G, D^k) = E_x[D_{src}^k(x)] - E_{z,c}[D_{src}^k(G(z_{encode}, c))] - E_{z,c}[D_{src}^k(G(z_{random}, c))] - \lambda_{gp} E_{\hat{x}}[(\|\nabla_{\hat{x}} D_{src}^k(\hat{x})\|_2 - 1)^2] \quad (2)$$

where $D_{src}^k(\cdot)$ denotes the output probability from k -th discriminator. \hat{x} is the linear interpolation between real and fake samples. λ_{gp} is the weight of gradient penalty and we use $\lambda_{gp} = 1.0$ for all experiments.

Feature Match Loss. To stabilize the GAN training, we adopt feature match loss to train generator. Specifically, the feature match loss tries to minimize the

distance of intermediate features from multi-scale discriminators between real and fake images. We denote the i -th layer feature of k -th discriminator as $D_{(i)}^k$. The feature match loss is defined as follows:

$$\mathcal{L}_{FM}(G, D^k) = \sum_{i=1}^T \frac{1}{N_i} [\|D_{(i)}^k(x) - D_{(i)}^k(G(z_{encode}, c))\|_1 + \|D_{(i)}^k(x) - D_{(i)}^k(G(z_{random}, c))\|_1] \quad (3)$$

where T is the number of layers used for feature matching. N_i is the number of elements in i -th layer. Here we use features of the last three convolution layers.

Pixel Reconstruction Loss. When passing an input image x through E and G, we can get a generated img $G(z_{encode}, c)$. We hope that $G(z_{encode}, c)$ can reconstruct the input x as far as possible. Hence, we adopt pixel-wise L1 loss to maintain structure information:

$$\mathcal{L}_{pixel} = \|x - G(z_{encode}, c)\|_1 \quad (4)$$

In addition, the encoder network E maps input x to the mean(μ) and covariance(ϵ). We apply KL loss to ensure that the latent representation obeys normal gaussian distribution:

$$\mathcal{L}_{KL} = \frac{1}{2}(\mu^T \mu + \text{sum}(\exp(\epsilon) - \epsilon - 1)) \quad (5)$$

Classification Loss. For an arbitrary face image, we hope D can not only distinguish real/fake, but also predict the identity. In detail, the classification loss is defined as

$$\mathcal{L}_{cls}^r(D^k) = E_{x,c}[-\log D_{cls}^k(c|x)]$$

$$\mathcal{L}_{cls}^f(G, D^k) = E_{z,c}[-(\log D_{cls}^k(c|G(z_{encode}, c)) + \log D_{cls}^k(c|G(z_{random}, c)))] \quad (6)$$

where $D_{cls}^k(c|x)$ represents the posterior for identity classification from k -th discriminator. By minimizing this objective, D tries to classify a real image to its corresponding identity, and G tries to generate a image with specific identity.

Identity-Preserving Loss. In order to leverage synthesized face images for face recognition task, it is crucial to keep the identity invariance of synthesized images. We imitate the perceptual loss [12] widely used in image style transfer. Specifically, with a pre-trained face recognition model Light CNN9 [20], we learn to match the intermediate features between real and fake images that have the same identity. The identity-preserving loss is calculated as follows:

$$\mathcal{L}_{id} = \|FR(x) - FR(G(z_{encode}, c))\|_1 + \|FR(x) - FR(G(z_{random}, c))\|_1 \quad (7)$$

where $FR(\cdot)$ is the output of penultimate fc layer of Light CNN9. Since Light CNN9 is dedicated to face recognition, its intermediate features contain rich

Algorithm 1. The training pipeline

Require: initial network parameters $\{\theta_E, \theta_G, \theta_D, \theta_{FR}\}$, hyper-parameters $\lambda_{cls} = 1, \lambda_{FM} = 1, \lambda_{pixel} = 10, \lambda_{KL} = 0.01, \lambda_{id} = 1$

Ensure: optimal network parameters $\{\theta_E, \theta_G, \theta_D\}$

- 1: Sample a batch from real data $\{x, c\} \sim P_r$
 - 2: $z_{encode} \leftarrow E(x)$
 - 3: Sample random noise from normal gaussian distribution $z_{random} \sim P_z$
 - 4: $\mathcal{L}_D = \sum_k [-\mathcal{L}_{adv}(G, D^k) + \lambda_{cls} \mathcal{L}_{cls}^r(D^k)]$
 - 5: $\mathcal{L}_{E,G} = \sum_k [\mathcal{L}_{adv}(G, D^k) + \lambda_{cls} \mathcal{L}_{cls}^f(G, D^k) + \lambda_{FM} \mathcal{L}_{FM}(G, D^k)] + \lambda_{pixel} \mathcal{L}_{pixel} + \lambda_{KL} \mathcal{L}_{KL} + \lambda_{id} \mathcal{L}_{id}$
 - 6: $\theta_D \leftarrow \theta_D - \nabla_{\theta_D}(\mathcal{L}_D)$
 - 7: $\theta_{E,G} \leftarrow \theta_{E,G} - \nabla_{\theta_{E,G}}(\mathcal{L}_{E,G})$
 - 8: If not converge, back to step 1; else stop iteration.
-

identity information. So it’s reasonable to keep identity invariance by such feature matching manner. During the training process, we freeze the parameters of Light CNN9 and only propagate the gradients back to E and G. We note that a similar loss is used in [11].

Overall Object Function. Finally, the overall object function is a weighted sum of loss functions defined above:

$$\begin{aligned} \mathcal{L}_D &= \sum_k [-\mathcal{L}_{adv}(G, D^k) + \lambda_{cls} \mathcal{L}_{cls}^r(D^k)] \\ \mathcal{L}_{E,G} &= \sum_k [\mathcal{L}_{adv}(G, D^k) + \lambda_{cls} \mathcal{L}_{cls}^f(G, D^k) + \lambda_{FM} \mathcal{L}_{FM}(G, D^k)] \\ &\quad + \lambda_{pixel} \mathcal{L}_{pixel} + \lambda_{KL} \mathcal{L}_{KL} + \lambda_{id} \mathcal{L}_{id} \end{aligned} \quad (8)$$

where $\lambda_{cls}, \lambda_{FM}, \lambda_{pixel}, \lambda_{KL}, \lambda_{id}$ are hyper-parameters to control the importance of each loss. We use $\lambda_{cls} = 1, \lambda_{FM} = 1, \lambda_{pixel} = 10, \lambda_{KL} = 0.01, \lambda_{id} = 1$ for all experiments. The whole training pipeline is shown in Algorithm 1.

4 Experiments

To validate the effectiveness of our approach, we evaluate our model qualitatively and quantitatively on FaceScrub [17] and LFW [9] datasets. We train our model on FaceScrub and test the model on LFW.

At preprocess stage, we perform face detection and get the facial landmarks using the multi-task cascaded CNN [22]. Then we align the faces by similarity transformation based on facial landmarks. The sizes of real and synthesized images are 128×128 . All the input are horizontal flip randomly.

For E and G, we use ReLU as activation function. The instance normalization is applied after each convolution layer. For multi-scale discriminators, we use two discriminators D^1 and D^2 , where the input of D^1 is 128×128 , and the input of D^2 is 64×64 . We use Leaky ReLU ($\lambda = 0.01$) as activation function.

In our experiments, the dimension of latent representation is 256. Our model is implemented using deep learning framework pytorch. The models are optimized using Adam [13] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We train all models with a learning rate of 0.0002 for the first 100 epochs and linearly decay the learning rate to 0 over the next 100 epochs. Training takes about 36 h on four NVIDIA 1080Ti GPU.

4.1 Qualitative Evaluation

Visualization Comparison. In this section, we compare the proposed method with CVAE and CGAN qualitatively. For CVAE, we remove the discriminative network D and only keep the pixel reconstruction loss \mathcal{L}_{pixel} and KL loss \mathcal{L}_{KL} . For CGAN, we remove the encoder network E as well as the pixel reconstruction loss and KL loss, i.e., set λ_{pixel} and λ_{KL} as 0. For fair comparison, we use the same network structure and training data. All methods use G to generate images.

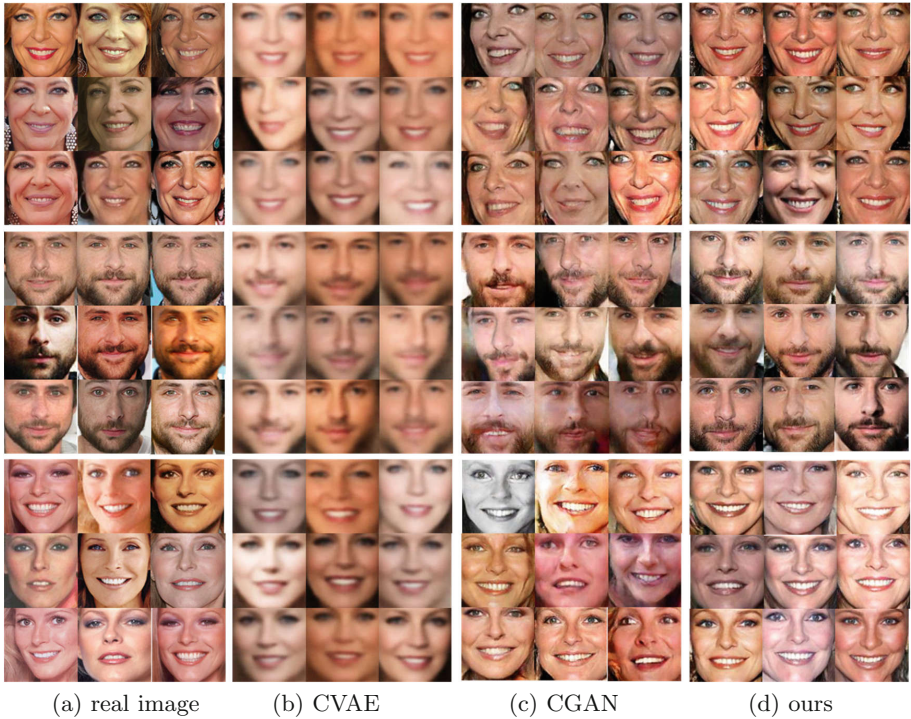


Fig. 5. Visualization results of each method. (a) Real images of 3 different identities. (b) Results of CVAE. It’s blur and lack of identity information. (c) Results of CGAN, which loses some structure information in some regions. (d) Our results, which is realistic, diverse and identity-preserving.

At test stage, we first randomly sample a identity c and a latent vector $z \sim N(0, I)$, and then pass them through G to generate a image with identity c . The visualization results of each method are shown in Fig. 5. We can see that images generated by CVAE are very blur. The reason is that CVAE merely uses the l1 reconstruction loss. Then images generated by CGAN often loss structure information in some regions, which is because of the absence of encoder. On the contrary, images generated by our approach are realistic and contain abundant intra-class variance, such as pose and expression. Furthermore, our method can keep the identity well. This shows the effectiveness of our approach.

Latent Representation Interpolation. To validate that our method can learn continuous and general latent space, we perform interpolation for latent representation. Specifically, we first randomly choose two faces of the same identity x_1 and x_2 , and then get latent vectors z_1 and z_2 through encoder network E . Next, we obtain a series of latent vectors by linear interpolation, i.e., $z = \alpha z_1 + (1 - \alpha)z_2, \alpha \in [0, 1]$. Finally, we generate samples using these interpolated vectors, as shown in Fig. 6. At each row, the left and right side are x_1 and x_2 , respectively. The interpolation results are in the middle. It can be seen that the facial pose, expression and skin color change gradually from left to right, which shows that the latent space learnt by our model is continuous.

4.2 Quantitative Evaluation

Evaluating the performance of generative model is a challenging problem. Many existing methods in face synthesis evaluate images by human, which is a laborious work and lack of objectivity. Following [3], we evaluate the model on image discriminability, realism and diversity.



Fig. 6. The result of latent representation interpolation

We first randomly generate 53k face images (100 images for each identity) using our method, CVAE and CGAN, respectively. To validate the discriminability of generated images, we train a face classification model using real data. Here we choose Light CNN29 [20] as a basic model, whose structure is similar to Light CNN9 but deeper than it. With the pre-trained classification model, we calculate the top-1 accuracy of images generated by each method. Table 1 shows the

results. Since CVAE merely uses the l1 reconstruction loss. It can't ensure identity invariance. So the accuracy is very poor. Our method achieve the best top-1 accuracy, showing significant margin than CVAE and is closing to real data (99.56% vs 99.69%). This suggests that images generated by our method is discriminative. It can be noted that CGAN also achieve high accuracy. We guess it's the contribution of identity-preserving loss. To validate this assumption, We remove the identity-preserving loss (set λ_{id} as 0), and retrain the model. We find that the accuracy drops dramatically (from 99.56% to 79.50%), which demonstrates that the identity-preserving loss plays a crucial role in keeping identity information.

We adopt inception score [18] to evaluate the realism and diversity of generated images. Specifically, we first train a face recognition model on CASIA-Webface [21] dataset, and then use $exp(E_x KL(p(y|x)||p(y)))$ as metric. If the model can generate more photo-realistic and diverse images, the inception score will be higher. From Table 1 we can see that our method achieve the highest score and is closing to the real data.

Table 1. Quantitative evaluation of image discriminability, realism and diversity.

-	Real image	CVAE	CGAN	Ours	Ours (w/o \mathcal{L}_{id})
Top-1 accuracy	99.69%	29.13%	98.35%	99.56%	79.50%
Inception score	48.86 \pm .79	20.81 \pm .21	43.95 \pm .46	45.17 \pm .59	44.14 \pm .45

4.3 Data Augmentation

The ultimate goal of this paper is to utilize generated images to train better face recognition models. In this section, we further demonstrate that our method can be used for data augmentation. We use FaceScrub as training set and LFW as testing set.

Following [3], we exploit two data augmentation strategies: (1) Generating more faces of existing identities. (2) Generating faces of new identities by mixing existing identity label. For strategy 1, we generate 200 images for each person in training set and get totally 100k images. For strategy 2, we first randomly sample 5k new identities by linearly interpolating three existing identity label, and then generate 100 images for each new identity, getting totally 500k images. The generated images are combined with original FaceScrub dataset to train face recognition model. The models used in this experiment are Light CNN29 [20] and Concentrate Loss [10].

At the testing stage, we use the output of penultimate fc layer as face feature. We adopt cosine similarity as metric for Light CNN29 and euclidean distance for Concentrate Loss. We compare the LFW accuracy with and without data augmentation, as shown in Table 2. We can observe that, Light CNN29 gets 1.30% improvement (from 92.23% to 93.53%) with existing ID augmentation and 0.90%

improvement (from 92.23% to 93.13%) with new ID augmentation. Consistently, Concentrate Loss gets 1.10% improvement (from 93.12% to 94.22%) with existing ID augmentation and 1.08% improvement (from 93.12% to 94.20%) with new ID augmentation. This demonstrates that our method can be used for data augmentation effectively and bring improvement for face recognition.

Table 2. Results of data augmentation

-	Data size	Light CNN29	Concentrate loss
Without data augmentation	90K	92.23%	93.12%
Existing ID augmentation	90K + 100K	93.53%	94.22%
New ID augmentation	90K + 500K	93.13%	94.20%

5 Conclusion

In this paper, we propose a conditional face synthesis framework that combines a variational auto-encoder with a conditional generative adversarial network, for synthesizing face images with specific identity. To improve image quality, we adopt multi-scale discriminators. Furthermore, we incorporate identity-preserving loss and classification loss to ensure identity invariance of synthesized images, and use feature matching loss to stabilize the GAN training. Experimental results demonstrate that our approach not only produces realistic, discriminative and diverse images but also is available for data augmentation.

References

1. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. arXiv preprint [arXiv:1701.04862](https://arxiv.org/abs/1701.04862) (2017)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223 (2017)
3. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: CVAE-GAN: fine-grained image generation through asymmetric training. arXiv preprint [arXiv:1703.10155](https://arxiv.org/abs/1703.10155) (2017)
4. Chen, Z., Nie, S., Wu, T., Healey, C.G.: High resolution face completion with multiple controllable attributes via fully end-to-end progressive generative adversarial networks. arXiv preprint [arXiv:1801.07632](https://arxiv.org/abs/1801.07632) (2018)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. arXiv preprint [arXiv:1711.09020](https://arxiv.org/abs/1711.09020) (2017)
6. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint [arXiv:1606.05908](https://arxiv.org/abs/1606.05908) (2016)
7. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
8. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems, pp. 5769–5779 (2017)

9. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07–49, University of Massachusetts, Amherst (2007)
10. Huang, R., Xie, X., Feng, Z., Lai, J.: Face recognition by landmark pooling-based CNN with concentrate loss. In: IEEE International Conference on Image Processing, pp. 1582–1586 (2017)
11. Huang, R., Zhang, S., Li, T., He, R., et al.: Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis. arXiv preprint [arXiv:1704.04086](https://arxiv.org/abs/1704.04086) (2017)
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
15. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network, 2016. arXiv preprint [arXiv:1609.04802](https://arxiv.org/abs/1609.04802) (2017)
16. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint [arXiv:1411.1784](https://arxiv.org/abs/1411.1784) (2014)
17. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 343–347. IEEE (2014)
18. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
19. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. arXiv preprint [arXiv:1711.11585](https://arxiv.org/abs/1711.11585) (2017)
20. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. arXiv preprint [arXiv:1511.02683](https://arxiv.org/abs/1511.02683) (2015)
21. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint [arXiv:1411.7923](https://arxiv.org/abs/1411.7923) (2014)
22. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**(10), 1499–1503 (2016)
23. Zhu, J.Y., et al.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems, pp. 465–476 (2017)



Face Image Illumination Processing Based on GAN with Dual Triplet Loss

Wei Ma¹, Xiaohua Xie^{1,2,3(✉)}, Jianhuang Lai^{1,2,3}, and Junyong Zhu^{1,2,3}

¹ Sun Yat-sen University, Guangzhou, China
mawei23@mail2.sysu.edu.cn

{stsljh, xiexiaoh6, zhujuny5}@mail.sysu.edu.cn

² Guangdong Key Laboratory of Information Security Technology, Guangzhou, China

³ Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Beijing, China

Abstract. It is generally known that the illumination could seriously affect the performance of face analysis algorithms. Moreover, in most practical applications, the illumination is usually uncontrolled. A number of methods have been put forward to tackle the problem of illumination variations in face images, but they always only work on facial region and need to segment faces in advance. Furthermore, many illumination processing methods only demonstrate on grayscale images and require strict alignment of face images, resulting in limited applications in the real world. In this paper, we propose a face image illumination processing method based on the Generative Adversarial Network (GAN) with dual triplet loss. Through considering the inter-domain similarity and intra-domain difference between the generated images and the real images, we put forward the dual triplet loss. At the same time, we introduce the self-similarity constraint of the images in the target illumination field. Experiments on the CMU Multi-PIE face datasets demonstrate that the proposed method preserve the facial details well when relighting. The experiment of 3D face reconstruction also verifies the effectiveness of the proposed method.

Keywords: Face image · Illumination processing
Generative adversarial nets · Dual triplet loss

1 Introduction

Because of the great development of biometric recognition and machine learning, face analysis technologies, such as face detection, face recognition and 3D face reconstruction, have received great attention. Nowadays, in a highly constrained environment, many classical algorithms have been able to achieve nearly perfect performance. However, in the real world, the imaging environment in most applications is uncontrolled. For example, the user's posture or expression are not a neutral state, the illumination condition changes and so on. Compared

with other interference factors, illumination has a greater impact on many face analysis algorithms. Therefore, the normalization of illumination is crucial for exploring the method of illumination invariant.

Over the years, a large number of methods on illumination invariance have been put forward. The invariant feature method is proposed to get the illumination invariant feature of images. Among them, Xie et al. [3] divided face images into large scale and small scale, and processed them separately. Recently, Wang et al. [4] proposed robust principal component analysis to eliminate the shadow produced by high-frequency features based on Xie's work. All these methods have achieved impressive results in the removal of soft shadows, but they are not effective in dealing with problems such as hard edge shadow caused by self occlusion. At the same time, these technologies can not be extended to color space, resulting in limited application in the real world.

With the development of 3D technology and deep learning, many researchers turn to use them to solve the illumination problems. Zhao et al. [5] propose a method for minimizing illumination difference by unlighting a 3D face texture via albedo estimation using lighting maps. Hold-Geoffroy et al. [6] trained a convolutional neural network to deduce the illumination parameters and reconstruct the illumination environment map. These methods are powerful and accurate. However, they are easily limited by data collection and unavoidable highly computing cost. In addition, most of the existing methods only focus on dealing with the carefully segmented face regions, which are not robust to the whole face images.

Inspired by the successful application of the Generative Adversarial Network in transfer learning [8] and domain adaptation [9], we propose to reformulate the face image illumination processing problem as a style translation task with a Generative Adversarial Network (GAN) in [10]. By using the circle reversible iterative scheme and via the multi-scale adversarial learning, we build the mapping from any complex illumination field to a target illumination field and its inverse mapping to effectively achieve the normalization of illumination without affecting any other non-illumination features of the image. In this paper, by analyzing the distance relationship between the generated image and the real image, an improved illumination processing method based on the dual triplet loss is proposed in order to better retain the details of the image and improve the quality of the generated image.

Overall, our contributions are as follows:

- We propose an improved illumination processing method based on Generative Adversarial Nets with dual triplet loss.
- We put forward the dual triplet loss through considering the inter-domain similarity and intra-domain difference between the generated images and the real images.
- We introduce the self-similarity constraint of the images in the target illumination field and add two image similarity indexes, SSIM and PSNR, to supplement the measure of similarity.

- We demonstrate that the proposed method can outperforms the state-of-the-arts realistic visualization results on non-strictly aligned color face images and eliminate the ill effects caused by illumination.

2 The Proposed Approach

2.1 Overall Network Framework

The overall network framework of our generative adversarial nets is shown in Fig. 1. The same as [10], our network consists of one generator and a pair of multi-scale discriminators with the same network structure but different classification constraint. We train G to translate an input image x under any lighting conditions into an expected lighting image \tilde{x}' conditioned on the target illumination label c' , $G(x, c') \rightarrow \tilde{x}'$. And then reconstruct \tilde{x}' to the input image conditioned on the original illumination label c using the same G , $G(\tilde{x}', c) \rightarrow \tilde{x}$. The discriminator $D1$ distinguishes between the synthesized output images \tilde{x}' and the real ones x , and classify the illumination category \tilde{c}' . The classification loss of real images used to optimize $D1$, and the fake images' used to optimize G . Similar but different, $D2$ distinguishes between \tilde{x}' and a randomly selected picture y' of maybe anybody's under target illumination condition and recognizes the identity \tilde{l}' to optimize G and $D2$.

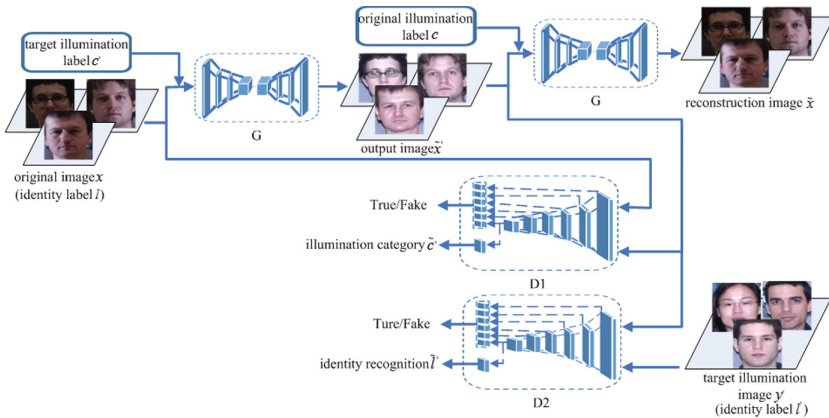


Fig. 1. Basic network architecture for face image illumination processing based on GAN with dual triplet loss.

2.2 Inter-domain Similarity and Intra-domain Difference

According to our research idea, face images under the same illumination conditions are divided into the same domain and our goal is to learn the mapping from any other illumination domain to the target illumination domain, which

refers the positive standard illumination in this paper. As shown in Fig. 2(a), the images before and after illumination normalization belong to different illumination domains, but their non-illumination information are same, which we call “inter-domain similarity”. At the same time, the different images after normalization belong to the same illumination domain, but their non-illumination information are different, which we call “intra-domain difference”.

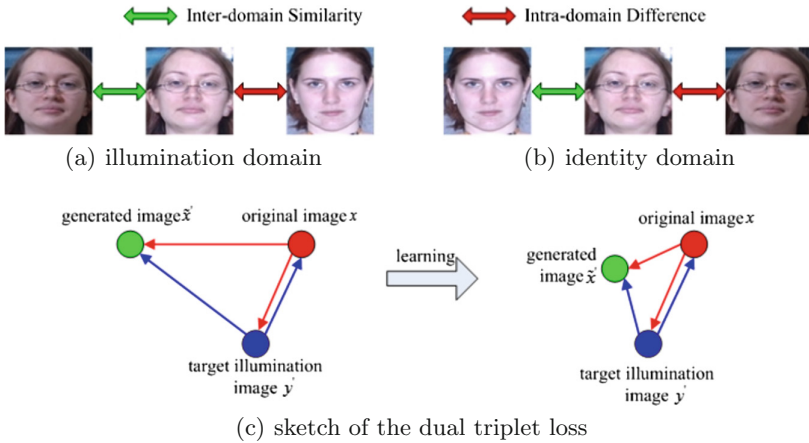


Fig. 2. Sketch of inter-domain similarity, intra-domain difference and the dual triplet loss.

Besides, as shown in Fig. 2(b). If we treat the non-illumination information as a symbol of the domain division, the two images before and after the normalization belong to the same identity domain, but their illumination information are different. That is, the two images have intra-domain difference now. Similarly, for any two different images after illumination normalization, they belong to different identity domains, but their illumination information are consistent. That is, the two images have inter-domain similarity now.

2.3 Dual Triplet Loss

Inspired by the thought of the triplet loss [11], we propose to construct a dual triplet loss based on the intra-domain difference and inter-domain similarity between the generated image and the real image. As is shown in Fig. 2(c).

The dual triplet loss include two triplet loss, each is composed of the original image x , the generated image \tilde{x}' after illumination normalization and the real image y' captured randomly from the target illumination domain. The first triplet loss takes y' as anchor and takes \tilde{x}' and x as positive and negative sample respectively. The second triplet loss takes x as anchor and takes \tilde{x}' and y' as positive and negative sample respectively.

Define $f(x)$, $f(\tilde{x}')$ and $f(y')$ are the features of x , \tilde{x}' and y' extracted from our multi-scale discriminant network. In the illumination domain, x and \tilde{x}' have inter-domain similarity. So the distance between them should be as small as possible and must be shorter than the distance between y' and x . That is:

$$\|f(x) - f(\tilde{x}')\|_2^2 - \|f(x) - f(y')\|_2^2 < 0 \quad (1)$$

Similarly, in the identity domain, \tilde{x}' and y' have inter-domain similarity. So the distance between them should be as small as possible and must be shorter than the distance between y' and x . That is:

$$\|f(y') - f(\tilde{x}')\|_2^2 - \|f(y') - f(x)\|_2^2 < 0 \quad (2)$$

In addition, \tilde{x}' and y' belong to the same illumination domain, but their non-illumination information are different. So, the distance between them should be larger than a minimum distance interval Δ_1 . That is:

$$\Delta_1 - \|f(y') - f(\tilde{x}')\|_2^2 < 0 \quad (3)$$

Similarly, in the identity domain, the distance between \tilde{x}' and x should be larger than a minimum distance interval Δ_2 . That is:

$$\Delta_2 - \|f(x) - f(\tilde{x}')\|_2^2 < 0 \quad (4)$$

In summary, the formula for calculating the loss function of dual triplet constraints is:

$$\begin{aligned} L_{dual-tri} = & \mathbb{E}[\|f(x) - f(\tilde{x}')\|_2^2 - \|f(x) - f(y')\|_2^2]_+ \\ & + \mathbb{E}[\|f(y') - f(\tilde{x}')\|_2^2 - \|f(y') - f(x)\|_2^2]_+ \\ & + \mathbb{E}[\Delta_1 - \|f(y') - f(\tilde{x}')\|_2^2]_+ + \mathbb{E}[\Delta_2 - \|f(x) - f(\tilde{x}')\|_2^2]_+ \end{aligned} \quad (5)$$

where $[\bullet]_+$ is a brief description of $\max[\bullet, 0]$, which indicates that the loss is valid only when the result value of $[\]$ is greater than 0, otherwise it is recorded as 0. The threshold distance Δ_1 is set as the minimum value of the feature distance between any two face images in the target illumination domain of the current training batch. Similarity, Δ_2 is set to the minimum value of the distance between any two face images in the original identity domain.

2.4 Self-similarity Constraint and Reconstruction Loss

The ideal function of the generate network is transferring the input image to the target illumination and keeping the non-illumination information unchanged. Therefore, if we use any real image of target illumination domain as input, the generated image should be the same as the original, namely “self-similarity. Because the illumination scene of them are already the target illumination and don’t need to be transferred.

Similar to the definition of the reconstruction loss in the previous article, we use the L1 distance to measure the error between the input and output image at first. The self-similarity constraint can be defined as

$$L_{rec-y'} = \mathbb{E}\|y' - G(y', c)\|_1 \quad (6)$$

L1 distance calculation is the sum of the absolute values of the corresponding pixel difference of all pixels between two images. The advantage is that it is convenient to calculate and can ignore the influence of the abnormal value in the image data, which is relatively stable and robust. But its disadvantage is also obvious, that is, the space between the pixels and their neighborhood is omitted, which may lead to the loss of high frequency information such as texture and detail. Based on the confirmation in [10], we use SSIM [12] and PSNR [13] to supplement the L1 distance in the image reconstruction constraint. Define:

$$\begin{aligned} L_{SSIM}(x_1, x_2) &= 1 - SSIM(x_1, x_2) \\ &= 1 - \frac{(2\mu_{x_1}\mu_{x_2} + c_1)(2\sigma_{x_1x_2} + c_2)}{(\mu_{x_1}^2 + \mu_{x_2}^2 + c_1)(\sigma_{x_1}^2 + \sigma_{x_2}^2 + c_2)} \end{aligned} \quad (7)$$

$$\begin{aligned} L_{PSNR}(x_1, x_2) &= 1 - \frac{PSNR(x_1, x_2)}{30} \\ &= 1 - \frac{1}{3} \log \frac{MAX_x^2}{MSE(x_1, x_2)} \end{aligned} \quad (8)$$

where MAX_x is the maximum possible pixel value of the image. $MSE(x_1, x_2)$ is the mean squared error of x_1 and x_2 . μ_{x_1} , μ_{x_2} , and σ_{x_1} , σ_{x_2} are the average and variance of x_1 and x_2 respectively. $\sigma_{x_1x_2}$ is the covariance of x_1 and x_2 . $c_1 = (0.01L)^2$ and $c_2 = (0.03L)^2$ are two variables to stabilize the division with weak denominator, in which L is the dynamic range of the pixel-values (1 in this paper). Special to note is that we use an empirical value of 30 to normalize the PSNR value.

Then the final cycle consistency loss of the generator can be written as

$$\begin{aligned} L_{rec-all} &= L_{rec-new} + \alpha_1 L_{rec-y'-new} \\ &= \mathbb{E}\|x - x_{rec}\|_1 + \alpha_2 (L_{SSIM}(x, x_{rec}) + L_{PSNR}(x, x_{rec})) \\ &\quad + \alpha_1 (L_{rec-y'} + \alpha_3 (L_{SSIM}(y', G(y', c)) + L_{PSNR}(y', G(y', c)))) \end{aligned} \quad (9)$$

We use $\alpha_2 = 0.5$, $\alpha_3 = 0.5$ and $\alpha_1 = 2$ in all of our experiments.

2.5 Loss Function

Base Loss. To stabilize the training process and generate higher quality images, we use Wasserstein GAN objective with gradient penalty as [8, 10, 14, 15]. Define \check{x}_1 and \check{x}_2 are sampled uniformly along a straight line between a pair of real image and generated image, as well as a pair of target illumination image and

generated image. The discriminator network $D1$ and $D2$ update their parameters by minimizing the following loss:

$$L_{adv1} = \mathbb{E}[D1_{src}(x)] - \mathbb{E}[D1_{src}(G(x, c'))] - \lambda_{gp} \mathbb{E}[(\|\nabla_{\check{x}_2} D1_{src}(\check{x}_1)\|_2 - 1)^2] \quad (10)$$

$$L_{adv2} = \mathbb{E}[D2_{src}(y')] - \mathbb{E}[D2_{src}(G(x, c'))] - \lambda_{gp} \mathbb{E}[(\|\nabla_{\check{x}_2} D2_{src}(\check{x}_2)\|_2 - 1)^2] \quad (11)$$

where we use $\lambda_{gp} = 10$ for all experiments.

For an input image x whose identity label is l and a target illumination label c' , our goal is to translate x into an output image \tilde{x}' , which is properly classified by $D1$ to c' and recognized by $D2$ to l . The classification loss for illumination and identity classification task can be defined uniformly as

$$L_{cls1} = \mathbb{E}[\log D1_{cls}(\hat{c}|\hat{x})] \quad (12)$$

$$L_{cls2} = \mathbb{E}[\log D2_{cls}(\hat{c}|\hat{x})] \quad (13)$$

where \hat{x} represents the image to be classified and the item \hat{c} represents the proper label \hat{x} should be in this classification task.

Loss Function for Generator. Define the illumination label and identity label of the synthesized output image as \tilde{c}' and \tilde{l}' . So, the base objective functions to optimize G can be written as

$$L_{G-base} = L_{adv1}(x, G(x, c')) + L_{adv2}(y', G(x, c')) + \alpha_4 L_{cls1}(\tilde{c}', c) + \alpha_5 L_{cls2}(\tilde{l}', l) \quad (14)$$

where α_4 and α_5 are hyper-parameters that control the relative importance of illumination classification and identity recognition losses respectively, compared to the adversarial loss. We set $\alpha_4 = 1$ and $\alpha_5 = 1$. According to Eqs. (14, 9, 5), the overall objective functions to optimize G can be written as

$$L_G = L_{G-base} + \alpha_6 L_{rec-all} + \alpha_7 L_{dual-tri} \quad (15)$$

The detailed description of all the individual loss functions was postpone above. We use $a_6 = 10$ and $a_7 = 10$ in all of our experiments.

Loss Function for Discriminator. The networks parameters of $D1$ and $D2$ can be optimized by minimizing a specifically designed adversarial loss L_{adv1} , L_{adv2} and the aforementioned classification loss L_{cls1} , L_{cls2} of the real one's respectively:

$$L_{D1} = -L_{adv1}(x, G(x, c')) + \alpha_8 L_{cls1}(\tilde{c}', c) \quad (16)$$

$$L_{D2} = -L_{adv2}(y', G(x, c')) + \alpha_9 L_{cls2}(\tilde{l}', l') \quad (17)$$

we set a_8 and a_9 as 1 in our experiments.

2.6 Model Training

We summarize the details of our algorithm training procedure in Algorithm 1. And we use the same history updating strategy as [10]. Moreover, we set $K_d = 5$, $K_g = 1$, $T = 1000$ and $lr_G = lr_D = 0.0001$ in the first 500 iterations, which both decay to 0 linearly in the following iterations.

Algorithm 1. Face Image Illumination Processing Based on the Dual Triplet Loss

Input: Real images x , identity label l , illumination label c and target illumination label c' . Images with target illumination y' , identity label l' . Max number of steps T , number of the two discriminator network update per step k_d , number of generative network updates per step K_g , the learning rate of lr_G and lr_D .

Output: The network parameters

```

1  for  $i = 1 : T$  do
2    for  $k = 1 : k_d$  do
3      Sample a batch of real images  $x$  and target illumination images  $y'$ ;
4      Get  $G(x, c')$  with current network;
5      If the history buffer is not null, update the batch content with half a
        batch images sampling from the buffer;
6      Update network parameters of  $D1$  by taking a Adam step on batch loss
         $L_{D1}$  in Eq. (16);
7      Update network parameters of  $D2$  by taking a Adam step on batch loss
         $L_{D2}$  in Eq. (17);
8      Sample half a batch images from the original  $G(x, c')$  and add to the
        history buffer.
9    end
10   for  $k = 1 : k_g$  do
11     Sample a batch of real images  $x$  and target illumination images  $y'$ ;
12     Get  $G(x, c')$  and  $G(y', c)$  with current network;
13     Reconstruct  $G(G(x, c'), c)$  and update network parameters of  $G$  by
        taking a Adam step on batch loss  $L_G$  in Eq. (15)
14   end
15 end

```

3 Experimental Results and Analysis

Experiments were conducted on the CMU Multi-PIE Face Database [1] to verify the effectiveness of the proposed methods. Notably, all the images in this dataset are color images, which is always a challenge on illumination normalization for traditional methods. In our experiments, we restrict our attention merely to the frontal face images with neutral expression. All images are simply aligned and resized to 128×128 pixels, among which the first 2000 pictures were used for test and the others used for training.

3.1 Comparisons of the Visual Quality with Other Methods

For convenience, we denote our previous base method in [10] as GAN-base and denote this paper's method as GAN-DTL. In Fig. 3, we compare the visual results of normalized images between the proposed GAN-DTL method, GAN-base method and two baseline algorithms: NPL-QI [17] and ITI [18]. Same as other traditional methods, these two baseline algorithms can only process gray images and require strict alignment of face images. However, even on gray images,

they don't work well. For example, the NPL-QI method can't handle the extreme illumination conditions such as the first group and the third group. There is a general loss of detail in face after processing of the ITI method. And these two methods are not effective in dealing with the self occlusion of nose in the second groups. In contrast, our GAN-DTL method and GAN-base method achieve the best normalization performance and preserve more facial details and almost all appearance information, such as the hairstyle and hair color. At the same time, our GAN-DTL method provides a higher visual quality of normalization results on all kinds of test images. Different skin colors were preserved closer to the original ones, especially obvious on the first group image. And the details of eyeglass frame and whiskers in the third group are preserved more perfect. The result indicated that the proposed GAN-DTL method can preserve the details of generated images better and improve the quality of generated images.

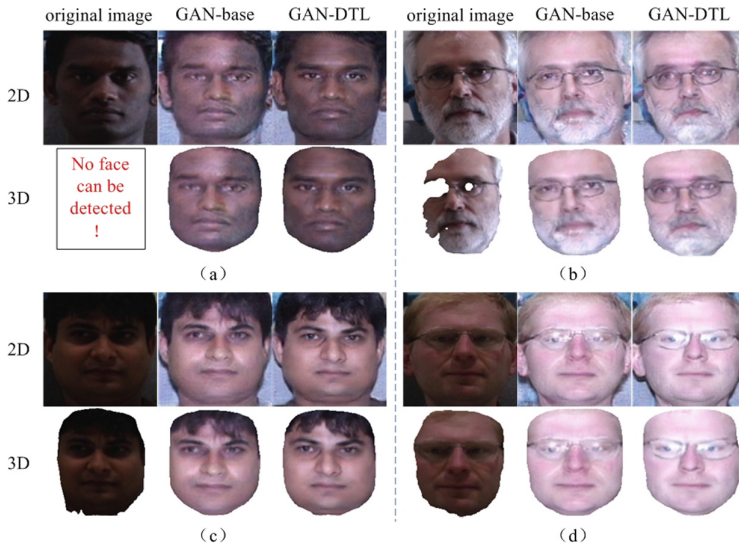


Fig. 3. Quantitative evaluation results comparison between the proposed GAN-DTL method, GAN-base method and two baseline algorithms.

3.2 Comparisons of the Ablation Study

We conduct ablation studies to show the superiority of our GAN-DTL method. We carry out the experiment on our 2000 test images. Take the face image of the same face under the target illumination as benchmark, we calculate the SSIM value and PSNR value of the original image, the generated image of GAN-base and the generated image of GAN-DTL respectively. And take the mean value according to the original illumination category then, which are drawn in black, blue and red curves in Fig. 4 respectively. As we can see, our GAN-DTL method

improves the evaluation results to a new height. The total average value of the SSIM is raised from 0.550 of the GAN-base method to 0.736 and the total average of the PSNR is raised from 16.048 to 21.324, which is consistent with the evaluation of the visual effect.

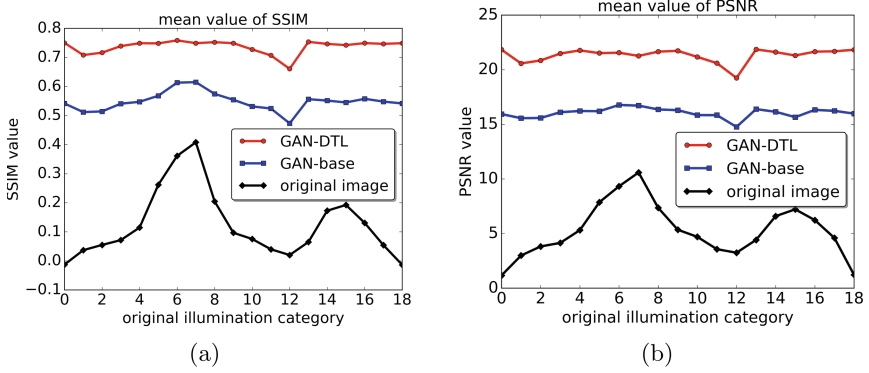


Fig. 4. Comparisons of the ablation study SSIM and PSNR. (Color figure online)

3.3 Test of Face Algorithm Application

We use the online 3D face reconstruction from a single image algorithm [19] which is put forward by the team of nottingham university in 2017. As is shown in Fig. 5. As the initial 3D reconstructed image is not a positive angle of view, the angle and size of the pictures are slightly deviated when they are manually rotated to the front view. But it obviously does not affect the experimental comparison. In group (a), as the original image is in the dark light condition and the skin color of the face is black, the face can not be detected in the 3D reconstruction. In group (b), due to the uneven illumination of original images, the location of facial landmarking is not allowed, resulting in partial deletion of reconstructed 3D models. Similarly, in group (c) and group (d), the face region segmentation of the original image is inaccurate due to the influence of illumination on the location of facial landmarking, and the rough edge produced by the shadow in the chin area. However, in the four sets of images, the 3D model can be built very well and smoothly for the generated images after our GAN-DTL and GAN-base method illumination normalization. And our GAN-DTL method achieve the best results and illustrate the effectiveness of the proposed method in real-world applications.

4 Conclusion

In this paper, we propose a face image illumination processing method based on Generative Adversarial Nets with dual triplet loss. Through considering

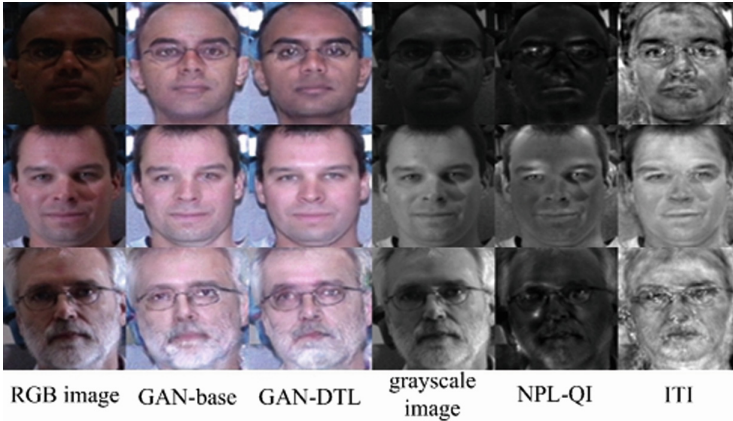


Fig. 5. 3D face reconstruction from a single image.

the inter-domain similarity and intra-domain difference between the generated images and the real images, we put forward the dual triplet loss. At the same time, we introduce the self-similarity constraint of the target illumination images and add two image similarity indexes, SSIM and PSNR, to supplement the measure of similarity. Experiments on the CMU Multi-PIE face datasets demonstrate that the proposed method preserve the details of generated images and improve the quality of generated images. The 3D face reconstruction experiment shows that the face images after our methods processing can eliminate the ill effects caused by illumination, and illustrates the effectiveness of the proposed methods in real-world applications.

Acknowledgment. This project is supported by the Natural Science Foundation of China (61672544, 61702566), Fundamental Research Funds for the Central Universities (No. 161gpy41), and the Tip-top Scientific and Technical Innovative Youth Talents of Guangdong special support program (No. 2016TQ03X263).

References

1. Gross, R., et al.: Multi-pie. *Image Vis. Comput.* **28**(5), 807–813 (2010)
2. Adini, Y., Moses, Y., Ullman, S.: Face recognition: the problem of compensating for changes in illumination direction. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 721–732 (1997)
3. Xie, X., et al.: Normalization of face illumination based on large-and small-scale features. *IEEE Trans. Image Process.* **20**(7), 1807–1821 (2011)
4. Wang, H., Ye, M., Yang, S.: Shadow compensation and illumination normalization of face image. *Mach. Vis. Appl.* **24**(6), 1121–1131 (2013)
5. Zhao, X., et al.: Minimizing illumination differences for 3D to 2D face recognition using lighting maps. *IEEE Trans. Cybern.* **44**(5), 725–736 (2014)
6. Hold-Geoffroy, Y., et al.: Deep outdoor illumination estimation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2 (2017)

7. Wu, Z., Deng, W.: One-shot deep neural network for pose and illumination normalization face recognition. In: 2016 IEEE International Conference on Multimedia and Expo (ICME). IEEE (2016)
8. Choi, Y., Choi, M., Kim, M., et al.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation (2017)
9. Patel, V.M., et al.: Visual domain adaptation: a survey of recent advances. IEEE Sig. Process. Mag. **32**(3), 53–69 (2015)
10. Anonymous: Face image illumination processing based on generative adversarial nets. In: 24th International Conference on Pattern Recognition (ICPR) (2018)
11. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
12. Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
13. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern recognition (ICPR). IEEE (2010)
14. Martin, A., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning (2017)
15. Gulrajani, I., et al.: Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems (2017)
16. Phillips, P.J., et al.: Overview of the face recognition grand challenge. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005. CVPR 2005, vol. 1. IEEE (2005)
17. Xie, X., et al.: Non-ideal class non-point light source quotient image for face relighting. Signal Process. **91**(4), 1048–1053 (2011)
18. Liu, J., et al.: Illumination transition image: parameter-based illumination estimation and re-rendering. In: 19th International Conference on Pattern Recognition, 2008. ICPR 2008. IEEE (2008)
19. Jackson, A.S., et al.: Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (2017)



Face Detection and Encryption for Privacy Preserving in Surveillance Video

Suolan Liu^(✉), Lizhi Kong, and Hongyuan Wang^(✉)

Changzhou University, Changzhou 213164, Jiangsu, China
lan-liu@163.com, hywang@cczu.edu.cn

Abstract. A number of techniques have recently been proposed for privacy preserving in video surveillance. Most of them are irreversible or have interference effect to the observation and recognition of human activities. In this paper, we address these issues by developing an effective method including face detection and encryption. In face detection, skin-color based approach fusing with fuzzy clustering is produced to detect facial candidates coarsely, and then we refine face by using SVM classifier. In face encryption, a reversible hybrid encryption (decryption) scheme based on spatial and value scrambling models is proposed. Simulation results verify the proposed mechanism can effectively detect and obscure faces while leaving the activities comprehensible and has high key sensibility for reducing the probability of attacking.

Keywords: Privacy preserving · Video surveillance · Face detection
Face encryption · Reversible

1 Introduction

Nowadays, video surveillance has become one of the most important auxiliary means in the field of public security monitoring. Video surveillance systems are widely deployed in many public places such as banks, supermarkets, airports, roads and residential areas [1–3]. Everyone is constantly being watched no matter whether you feel like it or not. However, in [4] a report about government surveillance revelations by NAS contractor Edward have raised new concerns about how best to preserve American’s privacy in the digital age. What is personal privacy? One approach defines it in property terms as any information which the individual has certain decisional right [5]. Thus one’s facial image, actions, location or copyrighted material are personal-partly, because they “belong” to the individual. Among these privacies, facial image is crucial and has highly close relationship with the others, because it can be directly used in face recognition technology to identify the monitored person’s identity [2, 3]. In general, privacy preserving measures based on video surveillance can be taken from two aspects [6, 7]. On the one hand, we should enhance law making and law enforcement to regulate videos collection, storage and usage to avoid malicious infringement and disclosure of individual information. On the other hand, it is necessary to take effectively technical measures to protect the data and information, such as using cryptography theories and computer vision algorithms [2, 3, 7]. Cryptography methods mainly focus on encrypting the whole frame images into an unreadable form

so that every unauthorized person cannot recover the original video [8]. The traditional encryption algorithms mainly include symmetric cryptographic algorithm and asymmetric cryptographic algorithm. This kind of methods is fit to processing videos for secure transmission over a communication line instead of real-time security monitoring and alarming for some particular activities recognition (e.g. fall and fights, etc.). Therefore encryption of the image as a whole may not be the most fixable method for this application. Recently, privacy preserving method based on computer vision has been a hot topic in the research field. Most of the preserving mechanisms are focused on partly modify the moving targets in the surveillance scenes [8, 9]. Target detection algorithms are used to localize the sensitive regions (e.g. face, eyes) and other methods are applied to obscure or conceal the selected regions, such as video masking, black boxes and replacing techniques. However, these methods are usually irreversible. Objective to recover the original video whenever needed for authorized person, we should apply reversible image processing methods with low computational complexity to meet the requirement of fast and real-time processing and preserving.

In this paper, we address the above-mentioned issues of privacy preserving in surveillance video by fusing image-processing method with encryption and decryption techniques. In particular, the proposed scheme consists of two steps including face detection and scrambling with the purpose of obscuring human face and monitoring his activities without revealing his identity at the same time.

The remaining of this paper is organized as follows. In Sect. 2, we review previous work related to pedestrian face detection, image encryption and decryption algorithms. Section 3 describes the proposed framework. The overview of the scheme is presented. Face detection approach and image encryption based on pixels spatial and value features scrambling models are given. In Sect. 4, simulations and experimental results are reported. Furthermore, we discuss the security of our proposed scheme. Finally, we conclude our work in Sect. 5.

2 Related Works

At present, video surveillances are widely set up for the purpose of ensuring security and smart life. From this point of view, one may like surveillance to be carried out with not be willing to reveal any individual information. As the most informative part of human, face is usually used for identification. Therefore, obscuring or concealing face technique becomes an urgent demand for video surveillance with privacy preserving. Face detection is the first step of this application. Many of the current face detection techniques contain two major modules including face localization and verifying by extracting ‘facial’ features. To accurately localize face region, some prior information of human face are required. Skin color and face geometry make explicit use as apparent properties. Human skin color is one of the most robust face features and can be efficiently applied to find the pixels belonged to human skin in a scene. Roughly, physical-based methods and statistical-based methods are two basic kinds of skin color-based face localization approaches. Furthermore, statistical-based approaches can be grouped into parametric approaches and non-parametric approaches. In parametric approaches, mean values, covariance matrices, Gaussian or mixtures of

Gaussians are used to build parametric face skin distribution models. For instances, in [10] Pujol et al. developed a fuzzy system to detect facial region by computing and fusing image variances from three color spaces of RGB, HSV and YCbCr. For the considering of error detections, a method of detecting where truly face locates is further proposed to eliminate these similar regions, such as the neck and hands. Experiments showed about 93% correct face detection rate in brief backgrounds and stable light conditions. In RGB space, Zhen et al. [11] built a maximum entropy model called the first order model (FOM) for parameter estimating human face. And then belief propagation algorithm was used to obtain fast selection and exact location for facial skin region. But the output of detection was in a gray scale skin map and the special region was not exactly located and marked. In non-parametric approaches, histogram, Bayesian approach and neural networks are usually developed to distinguish “face” or “non-face”. In [12], authors applied the histograms of oriented gradients (HOG) as skin feature extraction clue and a feed-forward neural network was trained to classify the face from candidates. They tested the performance of their proposed scheme in sequences of color images and achieved an accuracy of 91.4%. The recent research of convolutional neural networks (CNNs) as the hottest algorithm in application of videos has proposed different solutions for incorporating the face detection and human recognition. Lu et al. [13] proposed using Clarifai net [14] and VGG-D model [15] to extract features and fuse them before fine-tuning. A binary classification by support vector machine (SVM) was conducted to realize face detection. Experimental results on three public datasets verify its state-of-the-art performance. Although great progress has made in recent years, face detection is still confronted with many challenges and cannot handle the large variations in different poses, occlusion, illumination condition and face in poor-quality video sequences.

As reported in [8], Boulton proposed to protect privacy by using and adapting encryption techniques and combining them with intelligent video processing methods. The main contribution showed as cryptographically invertible obscuration only for authorized users in possession of the decryption key. Image encryption methods have been increasingly applied to meet security demands in video surveillance. The traditional encryption algorithms mainly include symmetric cryptographic algorithm and asymmetric cryptographic algorithm. Data encryption standard (DES), Triple data encryption standard (TDEA), Rivest Cipher5 (RC5) and International data encryption algorithm (IDEA) are typically symmetric cryptographic algorithms, while RSA (proposed by Ron Rivest, Adi Shamir and Leonard Adleman in 1977), ELGAMAL, RABIN, Diffie-Hellman and Elliptic curve cryptography (ECC) are asymmetric cryptographic algorithms. Video processing requires meeting its need such as fast and high-level efficiency. Therefore these traditional encryption algorithms may not be the most desirable algorithms for encrypting video frames with large size. By analyzing recent reports and publications, encryption schemes for image application may be grouped into three categories including pixel-position permutation, value permutation and hybrid scrambling methods. Arnold transform, Fibonacci transform and Hilbert transform are position permutation approaches with the disadvantage of not being able to change the original histogram. They only rearrange the positions of the image pixels rather than the pixel values. Once the histogram is revealed, exhaustion method can be used to find the original image. Value permutation-based algorithms such as Virginia

encryption [16], chaotic map [17] and gravitational transform [18] aim at changing value by setting some parameters in advance. However, contour of original image can always be found in the encrypted image, which may cause security issues. Hybrid scrambling methods are produced by combing the advantages of the two former methods. In [19], to property compromise between imperceptibility and robustness of logo image encryption, Roy et al. proposed to fuse redundant discrete wavelet transform (RDWT) with Arnold scrambling and furtherly reshape it. Qin et al. [20] presented a novel image hash securely generated scheme by diving the image into several quantizes and scrambling the variances of pixel values. Testing results showed good performances with respect to perceptual robustness and discrimination. In [21], a hybrid encryption scheme based on quaternion hartley transform (QHT) and two-dimensional logistic map are suggested to enhance the security level. Simulation results verified that the novel scheme not only had satisfied security level but also had certain robustness against cropping and noise disturbance.

3 The Proposed Method

In this section, we describe the proposed privacy preserving method based on two steps: face detection and face encryption. The framework is shown in Fig. 1. In our scheme, first, we develop cascaded classifiers to extract face from coarse-to-fine. Then, a hybrid encryption approach based on spatial and value scrambling models are used to change and rearrange pixels in facial region. The following subsections will discuss the procedure detailedly.

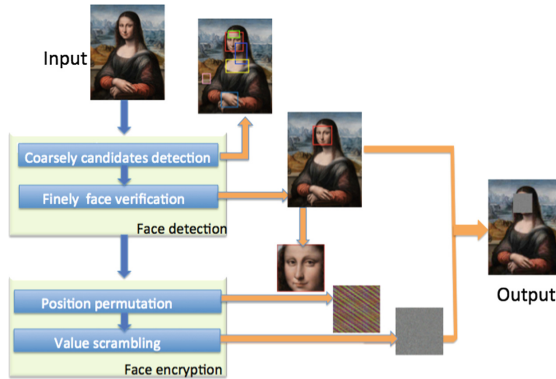


Fig. 1. Framework of our proposed scheme

3.1 Face Detection

Since most of the monitoring devices provide RGB video streams, approaches developed in this paper are based on RGB applications. Skin color model can be used to coarsely search facial candidates. Obviously, RGB has the negative property of each

coordinate (red, green and blue) is subject to luminance effects from light changes, which may cause misclassification of skin and non-skin regions. Reported researches show that skin color models work effectively only on the chrominance subspaces such as Cb-Cr [22, 23] and Hue-Saturation (H-S) [24]. Inspired by the work in [25, 26], in our approach skin candidates are produced using fuzzy c-means clustering (FCM) based on pixel local properties termed as LFCM in Cb-Cr subspace. In [27], the standard FCM is used to localize skin-like regions. However, they only consider pixel value instead of other useful information, such as the relationships between pixels, which play important roles in discriminating the category of a pixel. With this in mind, we improve FCM by considering attributions from 8-neighbor pixels of a point. Therefore, the conditional probability of a pixel x_i categorized into the j th class can be expressed as:

$$f(j_i|\eta_i) = \frac{e^{\beta\delta_i(j)}}{\sum_{i=1}^c e^{\beta\delta_i(j)}}, i = 1, 2, \dots, N \quad (1)$$

where j_i means that the pixel x_i is classified into j th class. η_i is the class label from 8 neighbors. $\delta_i(j)$ is the statistical number of 8-neighbor pixels belonged to j th class. β is the weight factor, $\beta \geq 0$. We set $\beta = 0.5$ in all of our tests in Sect. 4. The following criterion can be used to discriminate the pixel's category:

$$j^* = \arg \max u_{ij}, i = 1, \dots, N; j = 1, \dots, c \quad (2)$$

where u_{ij} is the fuzzy membership value and can be calculated by the following formula:

$$u_{ij} = u'_{ij} \times f(j_i|\eta_i) \quad (3)$$

u'_{ij} can be obtained from the standard FCM.

To refine facial region from several candidates, we conduct finely classification by SVM. To reduce the influences from illumination and different sizes, we do preprocessing including light compensation [27] and resizing every candidate to $64 * 64$. Define the block size as $16 * 16$ composed by cells sized $8 * 8$ with moving step $8 * 8$. Next, nine gradient orientation bins are selected to produce HoG features and concatenate them as final feature vector to train SVM model by using polynomial kernel function [28].

3.2 Face Encryption

Once face region is properly detected, the next step is scrambling it for security and privacy protection. Note that the encrypted face should be able to be recovered as needed [29]. Motivated by this requirement, a reversible hybrid encryption (decryption) scheme is proposed in this section, which uses Arnold transform in spatial position permutation [30] combining with gravitational transforms termed as GTs in value permutation [18] to encrypt and decrypt human facial region. In our numerical setting,

to facilitate Arnold transform, facial region is located in a bounding box sized $N \times N$. The facial region image is expressed as $f(x_i, y_i)$. In mathematics, the hybrid encryption operation is described as follows:

$$F(x_o, y_o) = G\{A(f(x_i, y_i))\}(x_o, y_o) \quad (4)$$

where $F(x_o, y_o)$ represents the output. The symbol “A” means Arnold transform (ART), “G” denotes GTs.

Furthermore, the facial image is imported to Arnold transform function [18], which is defined as:

$$A_N : \begin{bmatrix} x'_i \\ y'_j \end{bmatrix} = \text{mod} \left(\begin{bmatrix} 1, 1 \\ 1, 2 \end{bmatrix} \begin{bmatrix} x_i \\ y_j \end{bmatrix}, N \right) \quad (5)$$

where (x_i, y_j) and (x'_i, y'_j) are the coordinates before and after position permutation A_N .

The GTs can be given as:

$$G : \left[\gamma \frac{m_r \times m_{x'_i y'_j}}{(x_r - x'_i)^2 + (y_r - y'_j)^2 + k^2} \right] \text{mod} 256 \oplus V(x'_i, y'_j) \quad (6)$$

γ is gravitational coefficient and assigned a large positive number in experiments. $m_r = 1$ is the quality of unit particle which location is (x_r, y_r) . k is an adjusting parameter to ensure $(x_r - x'_i)^2 + (y_r - y'_j)^2 + k^2 > 0$. $m_{x'_i y'_j}$ is the quality of the pixel point (x'_i, y'_j) with pixel value $V(x'_i, y'_j)$. Note that V can be a three-tuple corresponding to components of color images.

4 Numerical Simulations and Discussion

The main idea of our work is to develop a reversible method for human face obscuring while having no interference to recognizing and monitoring their activities. To verify the performance of the proposed scheme, we do experiments by choosing several video clips with life scenarios. The operations in the processes of face detection, encryption and decryption will be conducted in Matlab running on a laptop.

4.1 Test One

In this test, the original testing image shown in Fig. 2(a) contains two faces with variations in illumination, position, orientation and accessories. As displayed in Fig. 2 (b), our approach can effectively detect faces with a certain range of skin color changes. Even though the left-side person is lowering the head, his facial region is properly localized. For the right-side person, accessories such as sunglasses greatly increase the difficulties of face refining, which may result in partial detection of human face.

However, our algorithm can successfully suppress this kind of influence and detect the whole face region. Obviously, the effectiveness of this part will greatly facilitate the next step of encryption.



Fig. 2. An example of face detection

In Table 1, we list the encrypted results by setting different parameters. For the sake of conducting fair comparisons, in GTs we set the unit particle's position as mean values for each facial position and assign the adjusting parameter $k = 100$. The first list displays the closeup of the detected faces; in the second list the ART results with different numbers of iterations are presented. We show the GTs results based on ART position permutation in the third list. The final encrypted results are displayed on the original images in the last list. From Table 1, one may find that with the changes of iterations from 3 to 80, the position scrambling effects show better from vision. Note that, once the number of iterations increases to a certain extent, it becomes a decryption operation. On the other hand, with the increase of gravitational coefficient the permutation of pixel values show more uniform and indistinguishable.


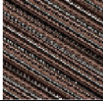













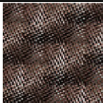
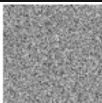



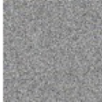
4.2 Test Two

A frame image contains multiple faces from different views coupled with cluttering background is utilized to test the robustness and security of our proposed scheme.

For the purpose of strengthening the security, in this test we set encrypt key as KEY4 as following: the number of iterations is 150 and three different sets of parameters for GTs corresponding to 3 channels [18] are applied. For red, $m_{x_i, y_j}^r = 85 \times x_i^2 + y_j^3 + 230$, $\gamma = 9 \times 10^{14}$; for green, $m_{x_i, y_j}^g = 60 \times x_i^2 + y_j^3 + 175$, $\gamma = 11.8 \times 10^{15}$; for blue, $m_{x_i, y_j}^b = 115 \times x_i^2 + y_j^3 + 70$, $\gamma = 10.5 \times 10^{13}$. The cipher-image is displayed in Fig. 3(c). As can be seen that even though the image shows small scaled faces and one of the actors in his profile, our method still achieves good detection rates and localizes the core areas of all faces.

To verify the key sensibility of the proposed method, we select the face from "Monica" shown in Fig. 4(a) and try to recover the encrypted image in Fig. 4(b) by using different decryption keys. Firstly, we decrypt it by using KEY5 of incorrect iterations as 90 for inverse ART operation, but no change to other parameters. The decrypted result displays in Fig. 4(c). Furthermore, we utilize only incorrect keys for inverse GTs operation with $m_{x_i, y_j}^r = 50 \times x_i^2 + y_j^3 + 60$, $\gamma = 9 \times 10^{13}$ for all 3 channels

Table 1. Scrambling processings and encrypted results

Closeup of face	ART	GTs following ART	Encrypted results
			 KEY1: For ART, the number of iterations is 3; For GTs, $m_{x_i y_j} = 50 \times x_i'^2 + y_j'^3 + 80$, $\gamma = 10^{13}$
			
			 KEY2: For ART, the number of iterations is 50; For GTs, $m_{x_i y_j} = 70 \times x_i'^2 + y_j'^3 + 100$, $\gamma = 10^{14}$
			
			 KEY3: For ART, the number of iterations is 80; For GTs, $m_{x_i y_j} = 90 \times x_i'^2 + y_j'^3 + 150$, $\gamma = 10^{15}$
			



(a) original image



(b) face detection result



(c) encrypted result

Fig. 3. An example of multiple faces detection and encryption

as KEY6 and show the result in Fig. 4(d). Figure 4(e) is decrypted image with correct keys. Conclusively, Figs. 4(c) and (d) indicate that the cipher-image can withstand some potential attacks. Experimental results show the high key sensibility in our scrambling scheme.

4.3 Discussion

Correlation coefficient between plain-image and cipher-image can be used to quantify the performance of an encryption algorithm. The lower correlation coefficient indicates that the encryption algorithm can better hide the feature information of the plain-image.

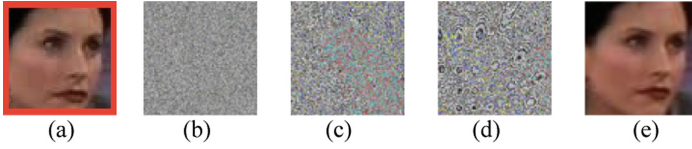


Fig. 4. Test results of key sensibility on a face

In this way, it will become more difficult to be attacked. Here we analyze the performance of the proposed algorithm by calculating the correlation coefficients of red, green and blue color channels respectively. The correlation coefficient between two-dimensional image matrix A and B can be defined as:

$$C_{AB} = \frac{\left| \sum_{i=1}^N \sum_{j=1}^N (A_{i,j} - \text{mean}(A))(B_{i,j} - \text{mean}(B)) \right|}{\sqrt{\sum_{i=1}^N \sum_{j=1}^N (A_{i,j} - \text{mean}(A))^2 \times \sum_{i=1}^N \sum_{j=1}^N (B_{i,j} - \text{mean}(B))^2}} \quad (5)$$

Where $\text{mean}(x)$ is the mean value.

Table 2 displays the correlation coefficient by using different keys. For Test One, we calculate the correlation coefficient between plain-image (Fig. 2(a)) and cipher-images (displayed in the forth list of Table 1). For Test Two, we calculate the correlation coefficient between cipher-image (Fig. 4(b)) and decrypted images (Figs. 4(c) and (d)).

From Table 2 one may find that for Test One most correlation coefficients are low as approximately zero. It indicates that the relevance between plain-image and cipher-image is very weak. From the aspect of encryption sensitivity, it means that the algorithm presented in this paper has superior sensitivity. Conversely, for Test Two while KEY5 is used to decrypt the cipher-image, correlation coefficient varies from 0.0243 to 0.0618, which shows high relevance. The reason for this phenomenon is the incomplete decryption of spatial position. However, we can see that once KEY6 is applied to decrypt, the average correlation coefficient is dramatically reduced from 0.0400 to 0.0061. As expected, localize and encrypt multiple faces in a picture is more challenging, but our proposed scheme is able to perform quite well with satisfied anti-attack property.

Table 2. Correlation coefficient between the red (r), green (g) and blue (b) color channels

Correlation coefficient		C_{rr}	C_{rg}	C_{rb}	C_{gr}	C_{gg}	C_{gb}	C_{br}	C_{bg}	C_{bb}	Average
Test one	KEY1	0.0007	0.0021	0.0015	0.0009	0.0018	0.0011	0.0014	0.0006	0.0023	0.0014
	KEY2	0.0013	0.0007	0.0003	0.0024	0.0007	0.0015	0.0008	0.0002	0.0009	0.0010
	KEY3	0.0004	0.0018	0.0021	0.0026	0.0015	0.0003	0.0020	0.0014	0.0031	0.0017
Test two	KEY5	0.0317	0.0243	0.0430	0.0357	0.0532	0.0618	0.0351	0.0426	0.0322	0.0400
	KEY6	0.0079	0.0050	0.0071	0.0068	0.0082	0.0064	0.0047	0.0038	0.0046	0.0061

5 Conclusion

We have proposed a practical privacy preserving technique for the application of video surveillance. Faces corresponding to privacy sensitive information are detected and encrypted. We aim to conceal faces while not interfere the observation and recognition of human activities using in intelligent monitoring and alarm systems. Our method is reversible for revealing faces whenever needed to the authorized person. Simulation results demonstrate that the proposed scheme can successively detect and obscure faces while leaving the activities comprehensible. Finally, the performance evaluation with key sensibility shows that the developed mechanism can withstand some potential attacks.

Acknowledgment. This work is supported by National Natural Science Foundations of China (No. 61572085, 61502058), Jiangsu Joint Research Project of Industry, Education and Research (No. BY2016029-15) and Changzhou Science and Technology Support Program (Social Development) Project (No. CE20155044).

References

1. Otto, C., Wang, D., Jain, A.: Clustering millions of faces by identity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2**(40), 1–14 (2018)
2. Torre, M., Granger, E., Gorodnichy, D.: Adaptive skew-sensitive ensembles for face recognition in video surveillance. *Pattern Recognit.* **11**(48), 3385–3406 (2015)
3. Radtke, P., Granger, E., Sabourin, R.: Skew-sensitive boolean combination for adaptive ensembles: an application to face recognition in video surveillance. *Inf. Fusion* **15**(20), 31–48 (2014)
4. Maddern, M., Rainie, L.: Americans' attitudes about privacy, security and surveillance. <http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/>
5. Haggerty, K., Ericson, R.: Varieties of personal information as influences on attitudes toward surveillance. <http://web.mit.edu/gtmarx/www/vancouver.html>
6. Bonetto, M., Korshunov, P., Ramponi, G.: Privacy in mini-drone based video surveillance. In: *Workshop on De-Identification for Privacy Protection in Multimedia*, vol. 4, pp. 2464–2469 (2015)
7. Dufaux, F., Ebrahimi, T.: Scrambling for privacy protection in video surveillance systems. *IEEE Trans. Circuits Syst. Video Technol.* **8**(18), 1168–1174 (2008)
8. Boulton, T.: PICO: privacy through invertible cryptographic obscuration. In: *Proceedings of the Computer Vision for Interactive and Intelligent Environment*, pp. 27–38, October, 2005
9. Carrillo, P., Kalva, H., Magliveras, S.: Compression independent reversible encryption for privacy in video surveillance. *J. Inf. Secur.* **1**, 1–13 (2009)
10. Pujol, F., Pujol, M.: Face detection based on skin color segmentation using fuzzy entropy. *Entropy* **26**(10), 1–22 (2017)
11. Zhen, H., Daoudi, M., Jedynak, B.: Blocking adult images based on statistical skin detection. *Electron. Lett. Comput. Vis. Image Anal.* **2**(4), 1–14 (2004)

12. Aulestia, P.S., Talahua, J.S., Andaluz, V.H., Benalcázar, M.E.: Real-time face detection using artificial neural networks. In: Lintas, Alessandra, Rovetta, S., Verschure, P.F.M.J., Villa, A.E.P. (eds.) ICANN 2017. LNCS, vol. 10614, pp. 590–599. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68612-7_67
13. Lu, X., Duan, X.: Feature extraction and fusion using deep convolutional neural networks for face detection. *Math. Probl. Eng.* **3**(2), 1–9 (2017)
14. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. <https://arxiv.org/abs/1409.1556>
16. He, M., Qiang, S.: Novel image scrambling algorithm based on changing pixel values. *Appl. Res. Comput.* **12**(29), 4635–4638 (2012)
17. Belazi, A., Hermassi, H., Rhouma, R., Belghith, S.: Algebraic analysis of a RGB image encryption algorithm based on DNA encoding and chaotic map. *Nonlinear Dyn.* **4**(76), 1989–2004 (2014)
18. Liu, S., Yue, C., Wang, H.: An improved hybrid encryption scheme for RGB images. *Int. J. Adv. Sci. Technol.* **4**(95), 37–44 (2016)
19. Roy, S., Pal, A.: A robust blind hybrid image watermarking scheme in RDWT-DCT domain using Arnold scrambling. *Multimed. Tools Appl.* **2**(76), 1–40 (2017)
20. Qin, C., Sun, M., Chang, C.: Perceptual hashing for color image based of hybrid extracting of structural features. *Signal Process.* **142**, 194–205 (2017)
21. Li, J.: Hybrid color and grayscale images encryption scheme based on quaternion hartley transform and logistic map in gyration domain. *J. Opt. Soc. Korea* **3**(20), 42–54 (2016)
22. Gundimada, S., Tao, L., Asari, V.: Face detection technique based on intensity and skin color distribution. In: International Conference on Image Processing, pp. 1413–1416, November 2004
23. Qing, L., Min, L.: Face detection using skin color and location relation. *Comput. Eng. Des.* **13**, 3396–3398 (2008)
24. Sabottka, K., Pitas, I.: Segmentation and tracking of faces in color images. In: International Conference on Automatic Face & Gesture Recognition, Vermont, pp. 236–241 (1996)
25. Anwar, N., Rahman, A.: RGB-H-CbCr skin colour model for human face detection. http://pesona.mmu.edu.my/~johnsee/research/papers/files/rgbhcbcr_m2usic06.pdf
26. Lu, J., Yuan, X., Yahagi, T.: A method of face recognition based on fuzzy c-means clustering and associated sub-NNs. *IEEE Trans. Neural Netw.* **1**(18), 150–160 (2007)
27. Hsu, R., Mottaleb, M.: Face detection in color image. *IEEE Trans. Pattern Anal. Mach. Intell.* **5**(24), 696–706 (2012)
28. Patilkulkarni, S., Lakshmi, H.: Vanishing moments of a wavelet system and feature set in face detection problem for color images. *J. Comput. Appl.* **16**(66), 36–42 (2013)
29. Liu, Z., Li, Q.: Image encryption based on random scrambling of the amplitude and phase in the frequency domain. *Opt. Eng.* **8**(48), 1–6 (2009)
30. Li, C., Lin, D., Lu, J.: Cryptanalyzing an image-scrambling encryption algorithm of pixel bits. *IEEE Trans. Multimed.* **3**(24), 64–71 (2017)



Content-Aware Face Blending by Label Propagation

Lingyu Liang and Xinglin Zhang(✉)

South China University of Technology, Guangzhou, China
lianglysky@gmail.com, zhxlinse@gmail.com

Abstract. Facial blending is critical for various facial editing applications, whose goal is to transfer the facial appearance of the reference to the target in seamless manners. However, when there are significant illumination or color differences between the reference and the target, visual artifacts may be probably introduced into the result. To tackle this problem, we propose content-aware masks that adaptively adjust the facial lighting and blended region to achieve seamless face blending. To generate the content-aware masks with good visual consistency, we formulate it as a label propagation process from a semi-supervised learning perspective, where the intensity of the initialized masks are propagated to the whole masks based on the local visual similarity of the images. Then, we construct a content-aware face blending framework that consists of three stages. Firstly, the facial region of the reference and the target are aligned according to the detected facial landmarks. Secondly, a facial quotient image and a binary mask are obtained as the initialized masks, and the content-aware masks for illumination and region adjustment are generated using the label propagation model with different guided feature. Finally, we combine the reference to the target using the generated masks to produce the face blending effects. Experimental results show the effectiveness and robustness of our methods for different image-based facial rendering tasks.

Keywords: Image-based rendering · Label propagation · Face transfer

1 Introduction

Facial image photo-realistic rendering is a novel computational photographic technique [12] to achieve facial effects that can be used for many applications, such as advertisement, movie production, digital entertainment, personalized photo editing and identity protection. Among the various current rendering

This research was supported in part by the National Natural Science Foundation of China under Grant No. 61502176, 61872151, the Natural Science Foundation of Guangdong Province under Grant No. 2016A030313480, the Pearl River S&T Nova Program of Guangzhou under Grant No. 201806010088 and Fundamental Research Funds for the Central Universities (No. 2017BQ058).

techniques [14], this paper specifically focuses on the facial appearance transfer problem of the image-based portrait rendering.

Facial appearance transfer is the critical component of various facial editing tasks, including face replacement [4, 13], face swapping [1, 7, 18, 19], face reenactment [5, 16] and age progression [6]. It aims to transfer the facial appearance of the reference to the target with good visual consistency.

It is challenging to achieve seamless face transfer. Most previous methods are based on the facial mask, and the facial property matching (like lighting or color) between a target and a reference. Pérez et al. [13] proposed the Poisson seamless cloning by the guided interpolation in the gradient domain. Dale et al. [4] used a novel graph-cut method that estimates the optimal seam on the face mesh to obtain video face replacement. Bitouk et al. [1] used the shading model based on a linear combination of spherical harmonics to adjust facial color and lighting for face swapping. Recently, Garrido et al. [5] proposed the automatic face reenactment system that replaces the face of an actor with the face of a user using a color adjustment with the Poisson blending [13].

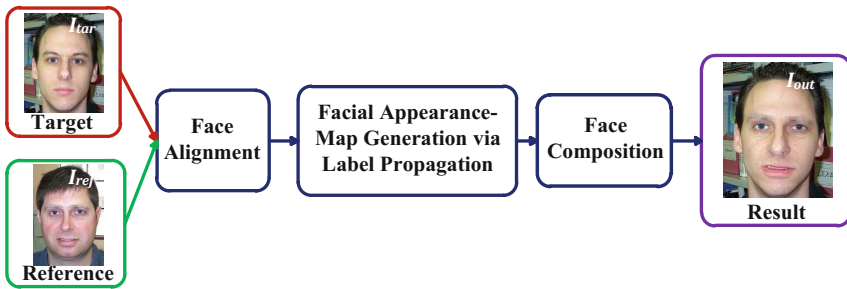


Fig. 1. The framework of the face transfer to blend the facial region of the reference to the target, which is based on facial appearance-map generated by adaptive label propagation.

The framework of the facial appearance transfer is shown in Fig. 1, which aims to transfer the facial region of the aligned reference to the target to produce the blended portrait. It consists of three stages: face alignment, facial appearance-map generation, and face composition.

Due to the complex appearance differences between the faces, however, a simple facial mask with Gaussian feathering may cause visual artifacts on the boundary of the transferred region; even the Poisson image editing [13] may fail to perform well when there is a large lighting or color difference between the target and reference, as shown in Fig. 4. To tackle the problem of illumination and region variances, this paper proposes a facial appearance map with good illumination-aware and region-aware properties for seamless facial appearance transfer. Inspired by Liang’s work on face enhancement [11], we formulate the facial appearance map generation as a label propagation process [20] from a

semisupervised learning perspective [2]. Since the face blending problem is different from the face enhancement problem in [11], we propose an adaptive label propagation model with a new regularization structure and guided features to achieve seamless face transfer.

Based on the adaptive facial appearance map, we construct the facial appearance transfer framework containing three stages. Firstly, the facial region of the reference is aligned to the target according to the detected facial landmarks. Secondly, a facial quotient image [10, 15] and a binary mask are generated, and then the guided label propagation model is used to diffuse the initial features of the quotient image and the binary mask to obtain the adaptive facial appearance-map for illumination and mask adjustment, respectively. Finally, we use the appearance-maps to seamlessly transfer the reference to the target. Experimental results show the effectiveness and robustness of our methods compared with the previous methods for various image based facial rendering tasks, such as face replacement and face dubbing in [1, 4, 13].

The main contributions of the paper are summarized as follows: (1) An adaptive label propagation model with guided features to generate the illumination-aware and region-aware facial appearance map for seamless face transfer; (2) A facial appearance transfer framework based on the adaptive facial appearance map, which achieves various image-based face blending effects, such as face replacement and face dubbing.

2 Facial Appearance Transfer Framework

2.1 Face Alignment

In face alignment, we aim to match the reference I_{ref} and the target I_{tar} to obtain the transformed reference I'_{ref} and the wrapped target I'_{tar} for appearance-map generation and face composition.

Firstly, we use the Viola-Jones face detector [17] and the active shape model (ASM) [3] to locate the 86 landmarks in the facial components of the reference S_{ref} and the target S_{tar} , respectively. Secondly, the transformed appearance I'_{ref} and shape S'_{ref} of the reference are obtained by matching the reference I_{ref} to the target I_{tar} using the affine transformation with the landmarks. Finally, we wrapped the target by the multilevel B-splines approximation (MBA) [9] according to the transformed shape of the reference S'_{ref} , i.e. the appearance of the wrapped target $I'_{tar} = f_{MBA}(I_{tar}, S_{tar}, S'_{ref})$. For more technical detail of MBA, we refer the readers to the article [9].

2.2 Facial Appearance-Map Generation

In face blending, directly pasting the face region of the reference to the target probably fail to perform well. According to our observation, apparent visual artifacts may be introduced to the results even through the gradient-domain Poisson cloning [13] is used when the reference and the target have large lighting

or color variances. To tackle these problems, we construct two different types of facial appearance-maps (T_{quot} and T_{mask}) that perform adaptive illumination and region adjustments of the reference for seamless face transfer.

Inspired by Liang’s recent work [11] for face enhancement, we formulate the facial appearance-map generation as a label propagation process, which diffuses the features within the initialized facial map to obtain the whole map. Since the two appearance-maps require different diffusion processes, we integrate different regularization structures with different guided features to the label propagation model for the corresponding map diffusion.

Specifically, the appearance-map T_{quot} aims to relight the reference so that the illumination and the color of the reference appearance is consistent to the target, and it uses the quotient image [15] as the initialization. Unlike the original quotient image that only handles the region within the faces, the diffused the quotient appearance-map T_{quot} facilitates to relight the face with consistent background illumination.

The appearance-map T_{mask} is to adaptively select the facial region of the relighted reference for seamless face transfer with smooth region transition, which uses the binary mask of the facial landmarks as the initial map.

The benefit of the diffusion-based map generation is twofold. Firstly, it is fault-tolerant to the small inaccurate landmark detection, since the final map value is determined by the label propagation process instead of the initialized value. Secondly, the map is adapted to the complex facial boundary and texture variance of the region by using different regularization structures and guided features. More detail of the structure and initialization of the label propagation model for T_{quot} and T_{mask} will be presented in Sect. 3.

2.3 Face Composition

To produce the output I_{out} of face transfer, we replace the facial region of the target I_{tar} with the facial region of I'_{ref} using the generated facial map T_{quot} , T_{mask} as follows:

$$I_{out} = I'_{ref} \circ T_{quot} \circ T_{mask} + I_{tar} \circ (J - T_{mask}), \quad (1)$$

where \circ denotes the element wise product operation, and J is the all-ones matrix with the same dimension of T_{mask} . The results of face blending are shown in Fig. 1, where the corresponding generated masks T_{quot} and T_{mask} are shown in Fig. 2.

3 Facial Appearance-Map Generation

3.1 Adaptive Label Propagation Model

The appearance-map for face transfer is formulated as a label propagation model with an adaptive regularization structure and guided features, which generates

the whole map by propagating the value of the initial map to the others according to the pixel similarity.

Specifically, the facial appearance-map T with n pixels is mapped into a graph $\mathbf{g} = (\mathcal{V}, \mathcal{E})$ of n nodes, where the node v_p corresponds to the p^{th} map location, and the edge e_{pq} links the node pair (p, q) with the pixel similarity W_{pq} . We initialize the node value by R (more details of R are in Sect.3.2), and obtain the appearance-map T by propagating the initial value of R through the graph according to the pixel-wise edge similarity given by the affinity matrix W .

The label propagation for appearance-map can be formulated as the minimization of the following quadratic cost functional:

$$Z(T) = \sum_p S_{pp}(T_p - R_p)^2 + \frac{\lambda}{2} \sum_{p,q} W_{pq}(T_p - T_q)^2 + \lambda\epsilon \sum_p T_p^2$$

The first term is the data term to constrain the diffusion region, where S is an $n \times n$ diagonal matrix given by $S_{pp} = 1$ in the constraint region, otherwise $S_{pp} = 0$. The third term is a small added regularization term that prevents degeneration.

The second term is the smoothing term to determine the local smoothness property of the generated map T , where λ is used to balance the relative weights of the data term and the smoothness term; the weight matrix W_{pq} is non-zero iff v_p and v_q are ‘‘neighbors’’, and its value measures the similarity between the nodes (pixels). In this paper, we use the typical value $\lambda = 1$ and $\epsilon = 0.0001$ for all the experiments.

The smoothness term has a closely relationship with graph Laplacian L_g . Specifically, D is a diagonal matrix with $D_{pp} = \sum_q W_{pq}$, and $L_g = D - W$ is the un-normalized graph Laplacian. A more compact form of the cost function can be obtained as following:

$$Z(T) = \|S(T - R)\|^2 + \lambda T^\top (L_g + \epsilon I) T. \quad (2)$$

The derivative of the cost is

$$\begin{aligned} \frac{1}{2} \frac{\partial Z(T)}{\partial T} &= S(T - R) + \lambda(L_g + \epsilon I)T \\ &= (S + \lambda L_g + \lambda\epsilon I)T - SR, \end{aligned} \quad (3)$$

T can be obtained when the derivative is set to 0:

$$T = (S + \lambda L_g + \lambda\epsilon I)^{-1} SR = L^{-1} SR, \quad (4)$$

which is a linear equation about a symmetric, positive-definite Laplacian matrix L . It can be solved efficiently by the conjugate gradient descent with the multi-level preconditioning [8].

Also, Eq. 4 can be solved using a Jacobi iteration, which is similar to the iterative label propagation proposed by Zhu and Ghahramani [20] and Liang’s mask propagation model [11], except for the weight matrix that controls the diffusion property.

To obtain the appearance map of face transfer, we construct a new kernel structure with guided features for the weight matrix W of appearance-map diffusion.

3.2 Diffusions of Facial Appearance-Map

The edge-aware property of the optimization-based label propagation model is mostly controlled by the smoothness term, specifically the similarity metric of the weight matrix W_{pq} . To produce the appearance-map for face transfer, we design a new kernel structure with guided features:

$$W_{pq} = \frac{c_{pq}G'_{pq}}{\|G'_p - G'_q\|^\alpha + \varepsilon}, \quad (5)$$

where G and G' are the guided features to control the local property of the map diffusion, c and α are the parameters to adjust the sensitivity of the guided features, ε is a small constant to avoid division by zero (typically $\varepsilon = 0.0001$). The appearance-map T_{quot} and T_{mask} can be generated by different initialization $R_{\{quot,mask\}}$ and weight matrix $W_{\{quot,mask\}}$ with the corresponding guided features and parameters.

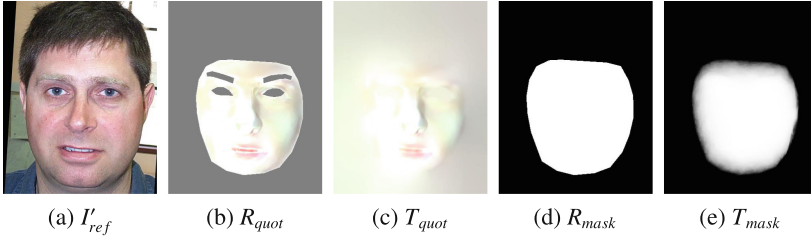


Fig. 2. Diffusions of facial appearance-map T_{quot} and T_{mask} based on adaptive label propagation with different guided features.

The appearance-map T_{quot} aims to adjust the illumination of the reference according to the target based on the facial shading model of the quotient image [15], as shown in Fig. 2(c). To produce T_{quot} , we set $R_{quot} = \frac{f_{aWLS}(I'_{tar})}{f_{aWLS}(I'_{ref})}$, where R_{quot} is the quotient image of the matched target I'_{tar} and reference I'_{ref} using Liang’s adaptive weighted least squares filter f_{aWLS} [10] for edge-aware smoothing, as shown in Fig. 2(b). For the weight matrix W_{quot} , we set $\alpha = 1$ and $G = \log L'_{ref}$, where L'_{ref} is the luminance of I'_{ref} (Fig. 2(a)); the value of cG' is small within the facial region and large in the background so that makes the features of the quotient image diffuse across the significant edges within the facial region to the whole image.

The appearance-map T_{mask} is responsible to paste the facial region of the reference to the target with smooth transition between different regions. For

T_{mask} , we set R_{quot} as a binary mask according to the facial landmarks, as shown in Fig. 2(d). To produce T_{mask} with adaptive region boundary, we set $\alpha = 1.2$, $G = \log L'_{ref}$ and $c = 0.5$ with $G' = J$, where L'_{ref} is the luminance of I'_{ref} and J is the all-one matrix. The map diffusion is controlled by the gradient of the guided feature G , which assures the smooth transition of the blended region between I'_{ref} and I_{tar} , as shown in Fig. 2(e).

4 Experiments

4.1 Basic Evaluation

The evaluations for facial appearance-map are shown in Figs. 2 and 3. The results show that the generated T_{quot} efficiently propagates the quotient value from the constrained regions of R_{quot} to the other regions, like eyes, eyebrows and background, and preserves the illumination consistence in the blended face. The appearance-map T_{mask} is generated with smooth transition, which is adapted to the region boundary between the face regions of the faces. The illumination-aware and region-aware diffusion of T_{quot} and T_{mask} ensure the robustness of the appearance transfer for faces with different properties, as shown in the experiments of face transfer.

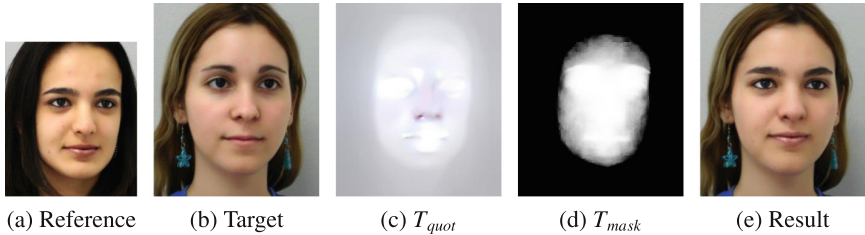


Fig. 3. Facial appearance-map for quotient-based illumination diffusion (T_{quot}) and blending mask diffusion (T_{mask}) using the proposed label propagation with corresponding guided feature.

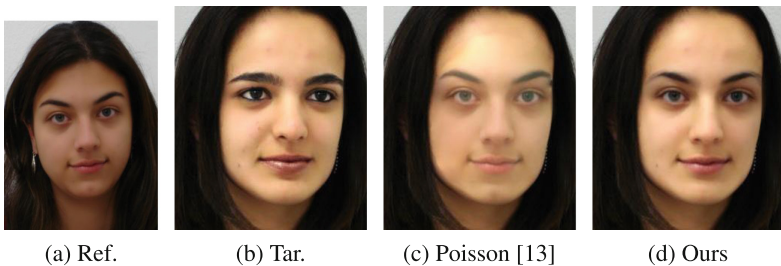


Fig. 4. Comparison with Poisson image cloning [13] for faces with large differences in age, color and lighting.

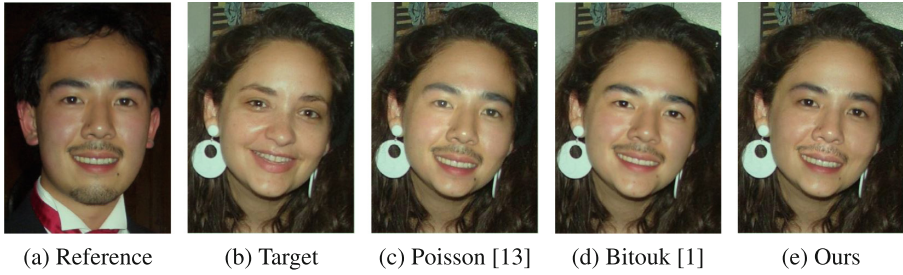


Fig. 5. Comparison with Bitouk et al. [1] for face replacement using reference target pair with different gender and roll rotation.

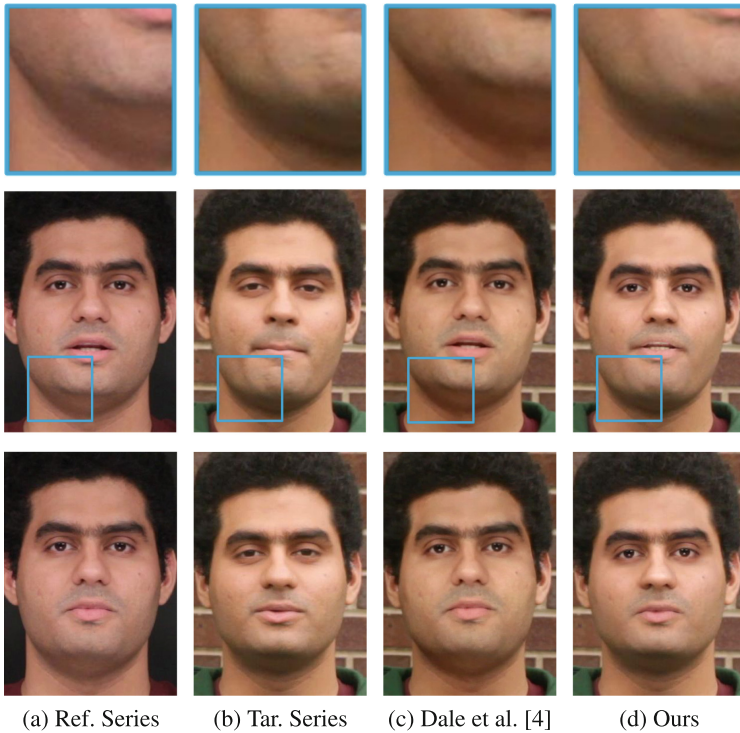


Fig. 6. Comparison with Dale et al. [4] for face dubbing, aims to transfer the series of the face appearance of the reference to the target. Comparison of the close-up images in the top rows illustrate that our method obtain better illumination consistency to the target than Dale's [4]

The basic experimental evaluations of face blending with the appearance-map were performed for the face pairs with significant different appearance properties, such as lighting, color, age and gender, as shown in Figs. 4, 5 and 6. The test

images were taken from the FEI face database or the internet. The good visual consistency of the results indicate the effectiveness and robustness of our method.

4.2 Comparison with Related Methods

We also made comparison with the related methods for face replacement [1, 13]. Figure 4 shows the comparison with the Poisson cloning [13]. Due to the dependency of the gradient and boundary of the blended region, the results of [13] are sensitive to the lighting and color differences of the faces. In contrast, our method obtains natural face blending effects. Comparison with Bitouk’s method in Fig. 5 further validates the effectiveness of our diffusion-based model.

We made the comparison between Dale’s [4] and our method for face dubbing, which aims to transfer the series of the face appearance of the reference to the target. The results indicate that both the methods can achieve good visual consistency in a global manner, as shown in Fig. 6. The close-up images of the local region in the first rows of Fig. 6, however, show the subtle differences. Dale’s method [4] tends to transfer the lighting property of the reference to the target, while ours tends to preserve the original appearance property of the target, which is complementary to [4].

5 Conclusion

This paper proposes a label propagation model with adaptive regularization to achieve facial blending with good visual consistency. Specifically, the illumination-aware and region-aware facial appearance maps are generated by diffusion with different guided features. Experiments illustrate the effectiveness and robustness of our methods for face replacement and face dubbing.

References

1. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. *ACM Trans. Graph.* **27**(3), 39 (2008)
2. Chapelle, O., Schölkopf, B., Zien, A.: *Semi-Supervised Learning*. MIT Press, Cambridge (2006)
3. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. *Comput. Vis. Image Underst.* **61**(1), 38–59 (1995)
4. Dale, K., Sunkavalli, K., Johnson, M.K., Vlasic, D., Matusik, W., Pfister, H.: Video face replacement. *ACM Trans. Graph.* **30**(6), 130 (2011)
5. Garrido, P., Valgaerts, L., Rehmsen, O., Thormahlen, T., Perez, P., Theobalt, C.: Automatic face reenactment. In: *Proceedings of CVPR*, pp. 4217–4224 (2014)
6. Kemelmacher-Shlizerman, I., Suwajanakorn, S., Seitz, S.: Illumination-aware age progression. In: *Proceedings of CVPR*, pp. 3334–3341 (2014)
7. Korshunova, I., Shi, W., Dambre, J., Theis, L.: Fast face-swap using convolutional neural networks. In: *Proceedings of ICCV*, pp. 3677–3685 (2017)
8. Krishnan, D., Fattal, R., Szeliski, R.: Efficient preconditioning of laplacian matrices for computer graphics. *ACM Trans. Graph.* **32**(4), 142 (2013)

9. Lee, S., Wolberg, G., Shin, S.Y.: Scattered data interpolation with multilevel B-splines. *IEEE Trans. Vis. Comput. Graph.* **3**(3), 228–244 (1997)
10. Liang, L., Jin, L.: A new face relighting method based on edge-preserving filter. *IEICE Trans. Inf. Syst.* **E96–D**(12), 2904–2907 (2013)
11. Liang, L., Jin, L., Liu, D.: Edge-aware label propagation for mobile facial enhancement on the cloud. *IEEE Trans. Circuits Syst. Video Technol.* **27**(1), 125–138 (2017)
12. Lukac, R.: *Computational Photography: Methods and Applications*. CRC Press, Boca Raton (2010)
13. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. *ACM Trans. Graph.* **22**, 313–318 (2003)
14. Reinhard, E., Efros, A.A., Kautz, J., Seidel, H.P.: On visual realism of synthesized imagery. *Proceedings of IEEE* **101**(9), 1998–2007 (2013)
15. Shashua, A., Riklin-Raviv, T.: The quotient image: class-based re-rendering and recognition with varying illuminations. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(2), 129–139 (2001)
16. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. *ACM Trans. Graph.* **34**(6), 183 (2015)
17. Viola, P., Jones, M.: Robust real-time face detection. *Int. J. Comput. Vis.* **57**(2), 137–154 (2004)
18. Zhang, Y., Zheng, L., Thing, V.L.: Automated face swapping and its detection. In: *IEEE International Conference on Signal and Image Processing*, pp. 15–19 (2017)
19. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: *Proceedings of CVPR Workshops*, pp. 1831–1839 (2017)
20. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation (2002)



Facial Expression Recognition Based on Region-Wise Attention and Geometry Difference

Heran Du^{1,2,3}, Huicheng Zheng^{1,2,3(✉)}, and Mingjing Yu^{1,2,3}

¹ School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

zhenghch@mail.sysu.edu.cn

² Key Laboratory of Machine Intelligence and Advanced Computing,
Ministry of Education, Guangzhou, China

³ Guangdong Key Laboratory of Information Security Technology,
135 West Xingang Road, Guangzhou 510275, China

Abstract. Facial expression is usually considered as a face movement process. People can easily distinguish facial expressions via subtle facial changes. Inspired by this, we design two models that are expected to better recognize facial expressions by capturing subtle changes in the face. First, we consider to re-calibrate the response of different facial regions to highlight several special facial areas. According to this idea, we constructed cross-channel region-wise attention network (CCRAN), which can underline the important information and mine the correlations between different facial regions effectively. Moreover, we use the feature subtraction method to obtain geographical facial difference information. Based on this idea, we constructed temporal geometric frame difference network (TGFDN), which accepts the facial landmark points as input. These points are extracted from the facial expression frames. This network can effectively extract the slight changes of geographical information on the expression sequences. Through properly fusing these two networks, we have achieved competitive results on the CK+ and Oulu-CASIA databases.

Keywords: Facial expression recognition · Attention mechanisms
Temporal difference

1 Introduction

Facial expressions are part of the human body's language. It is a physical and psychological response commonly used to convey feelings. Therefore facial expression recognition (FER) in the human-computer interaction is very important. In order to conduct the interaction, the machine needs to recognize the human facial expression to perceive their feeling. Considering that the expression often contains rich emotional information, the application of this task is very extensive.

FER is generally considered as a classification problem. Many people have done a lot of research in this field before. Overall, these studies can be divided into two categories: frame-based methods and sequence-based methods [1, 7, 15, 20, 24, 28]. Because facial expressions are generally considered as a movement process, extracting useful temporal and spatial features is very helpful for facial expression recognition. Therefore, the recognition methods based on the image sequence are generally considered to be superior to the methods based on a still single frame [7, 15].

However, the above methods are mainly based on the entire human face. In facial expression recognition tasks, the major changes in expression are often concentrated in several subtle facial regions. Humans can accurately recognize the category of expression through several key areas of the face, such as forehead, mouth, and brow. Therefore, the weights in different areas of the feature maps should be different.

In this paper, we first propose cross-channel region-wise attention network (CCRAN), trying to find the relationship between the different regions of the feature map. We hope to improve the network's ability to express specific image regions by introducing the cross-channel region-wise squeeze and excitation (CCSE) branch. Through this branch, we expect to re-calibrate features and enhance the image regional sensitivity of the network without introducing additional information.

Furthermore, we also propose temporal geometric frame difference network (TGFDN) to extract the temporal features from the facial landmarks. This network can effectively capture facial morphological changes and accurately describe facial movement characteristics. By performing feature extraction and frame difference for the landmarks of each frame separately, the network can extract low-level facial expression movement information from the landmarks. The result of the landmark difference is concatenated along the time axis and then input into the subsequent layers to further extract the high-level expression features. At the end of that, we can obtain the geometry information and movement characteristics of facial expressions.

The main contributions of this paper are divided into three parts.

- We propose CCRAN model, which accepts continuous frames as input, enhancing the network ability to recognize facial expressions by adding cross-channel region-wise attention mechanisms to the network.
- We propose TGFDN model, which can extract the inter-frame difference information from the facial landmarks points and can describe the motion process of expressions accurately.
- Finally, we fuse these two networks. The integrated deep spatial-temporal network takes into account geometry-appearance, regional-global, intra-frame and inter-frame information synthetically, improving the accuracy of expression recognition effectively.

2 Related Work

2.1 FER Based on Traditional Methods

Before the large-scale use of the deep learning-based method, it is a common practice to use hand-crafted features for facial expression recognition. These methods can be further divided into three kinds of methods based on local features extraction, facial action units (FAUs), and spatio-temporal information, respectively. Traditional methods based on local features, such as HOG, SIFT, LBP, and BoW have been extended to video. These methods also have their 3D cases [11, 15, 23, 25, 31]. In FAU based methods [12, 13], facial action coding system (FACS) is used to detect and analyze FAUs to classify facial expressions. The methods based on spatio-temporal information are represented by the work of Liu et al. [15]. They have proposed an expressionlet-based spatio-temporal manifold descriptor.

2.2 FER Based on Deep Methods

In recent years, deep convolutional neural networks have achieved great success in image classification [4, 5, 27], object detection and localization [3, 16, 21, 22], semantic segmentation [3, 17], and other computer vision fields. Corresponding to these tasks, in the field of facial expression recognition, Liu et al. propose 3DCNN-DAP [14], which is based on 3D-CNN, constructing a deformable parts learning component to capture the expression features. Further, Jung et al. [8] trained two small deep networks with facial landmarks and image sequences separately. To achieve the better result, they performed joint fine tuning method to fuse these two networks. Based on this structure, Zhang et al. [29] introduce recurrent neural network to further analyze the facial landmarks. Ding et al. [2] use a large pre-trained face recognition network to help train a simple facial expression recognition network through a regularization mechanism. Based on this, Ofodile et al. [19] further improved the accuracy by introducing the motion trajectory of the landmark points into the network. In addition, Kim et al. [10] attempted to use a small deep encoder-decoder network pre-trained on a face database to obtain a contrastive representation between expression face and neutral face, which helps to distinguish expressions.

3 Approach

In summary, the proposed method uses a combination of two simple networks. First, we construct the TGFND to capture the geographical inter-frame motion information. Then we use CCRAN to extract local appearance information in consecutive frames of the expression. Finally, these two networks are properly combined to improve the performance of facial expression recognition.

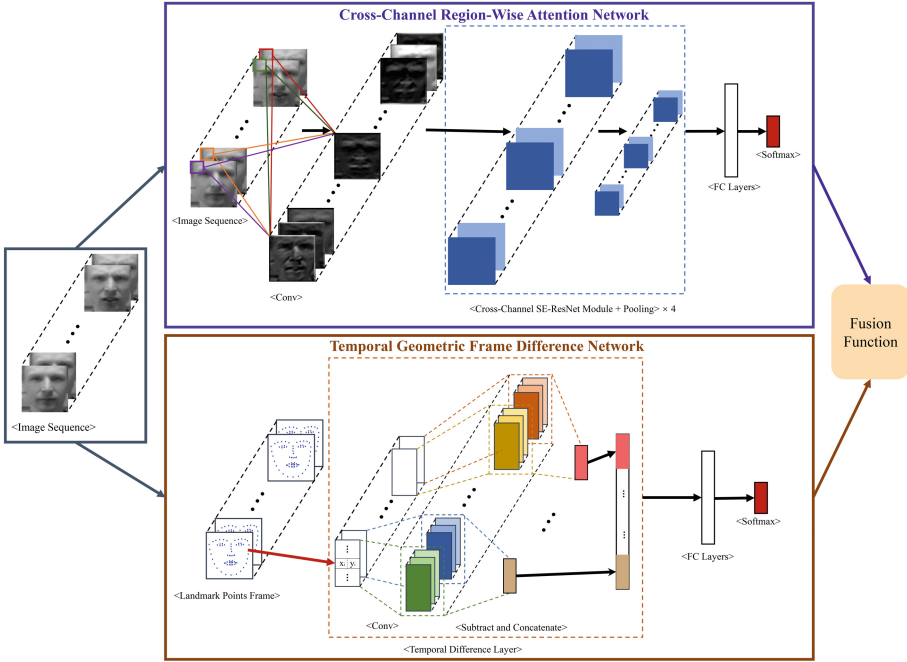


Fig. 1. Overview of our proposed architecture. The upper part of the figure shows the structure of CCRAN. The image sequence is fed into the network directly. Using a simple bottleneck (a convolution layer, a ReLU activation layer, and a batch normalization layer), the channels are increased to 64. After that, four cross-channel region-wise attention (CCRA) blocks are interleaved with four pooling layers and then followed by a fully connected layer to get logits. The lower part of the figure shows the structure of TGFDN. Facial landmark points are extracted from the frame sequence, reshaped into a matrix in which each row stores the coordinates of a point. Then the landmark matrices are fed into convolution layers separately. After the feature subtraction and difference concatenation, a fully connected layer is used to obtain logit values.

3.1 Cross-Channel Region-Wise Attention Network

In recent years, adding short connections to the network has proven to be an effective way to increase the efficiency of network information propagation [4, 6, 26]. So we use a simple CNN-Resnet structure as our backbone, which receives t frames of expression as input. The network includes four residual blocks interleaved with four pooling layers, and a fully connected layer at the end. Each residual block contains two convolutional layers. A batch normalization layer and a ReLU activation layer are between them, as shown in Fig. 3(a).

The whole Resnet block shown in Fig. 3(a) can be regarded as a unit that does not change the size and channels. The main problem with the backbone is that the convolutional operation takes equal considerations for the entire feature map and are less sensitive to subtle local changes. So we have joined the

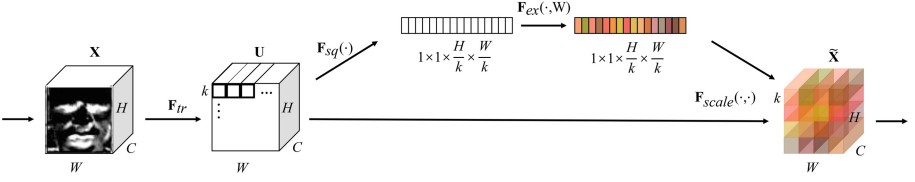


Fig. 2. Overview of the cross channel squeeze and excitation process.

cross-channel region-wise attention branch on the basis Resnet block of this network. This branch draws on the squeeze and excitation network [5] and can be trained end-to-end, including a cross-channel squeeze and a cross-channel excitation operation as shown in Fig. 2.

The purpose of the squeeze operation is to compress the information of all feature maps within a layer into a one-dimensional vector. Specifically, we first compress all feature maps into a single feature map using average pooling. Then we use a $k \times k$ filter to do average pooling again on this entire compressed feature map. Each region of the compressed feature map is compressed to one value. We then flatten these values into a one-dimensional vector. The vector obtained in this way takes into account the context between the channels and the facial regions. Formally, a two-dimensional matrix $z \in \mathbb{R}^{\frac{H}{k} \times \frac{W}{k}}$ is generated by squeezing U through cross-channel $k \times k \times C$ sized average pooling window, where the z_{ij} is calculated by:

$$z_{ij} = F_{sq}(U) = \frac{1}{k \times k \times C} \sum_{c=1}^C \sum_{h=i \cdot k}^{i \cdot k + k - 1} \sum_{w=j \cdot k}^{j \cdot k + k - 1} u_c(i, j) \quad (1)$$

We further extract the contextual relationships between the regions contained in the vector through the excitation operation. Like SE-net [5], in order to reduce the complexity of the model while reducing over-fitting, we use two fully-connected layers as a bottleneck. One layer is the dimension-reduction layer, and the other is the dimension-restoring layer. Between these two layers, we use a ReLU as the activation layer to get more nonlinearity, so as to better fit and mine the complex correlations between different regions. We will use this branch to integrate with the original Resnet block. As we have shown in Fig. 3.

We obtain CCRAN by using the block in Fig. 3(b) to replace the block in Fig. 3(a). It can be seen from Fig. 3(b) that the cross-channel SE branch we proposed can be added flexibly to the original network structure. Here, we join the cross-channel SE branch before the identity addition operation.

3.2 Temporal Geometric Frame Difference Network

The entire network includes a temporal difference layer and two fully connected layers as shown in the upper part of Fig. 1. The TGFND network receives the sequence of facial landmarks as input. We select t -frame facial landmarks to

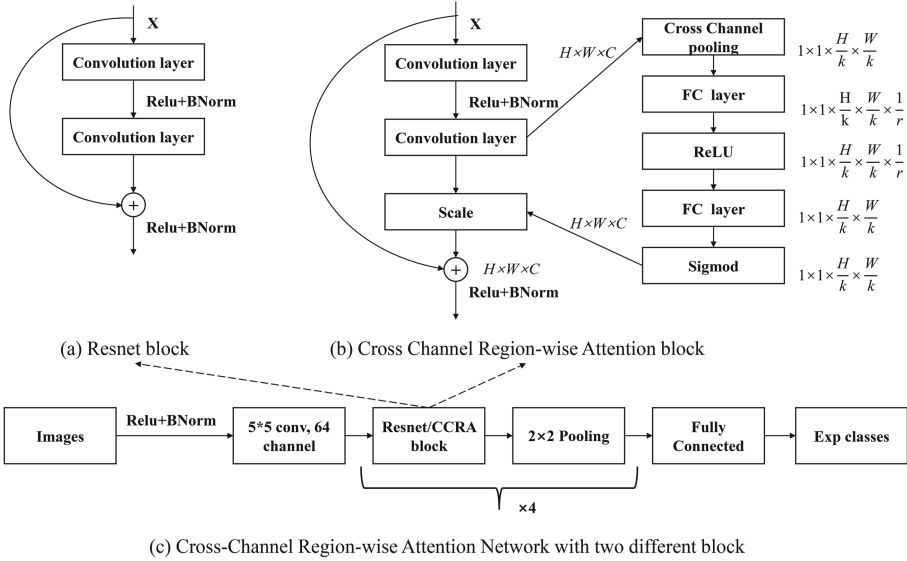


Fig. 3. Overview of the CCRAN architecture: (a) resnet block, (b) cross channel region-wise attention block, (c) the backbone network with two kinds of block.

describe the expression features. In Fig. 1, the landmarks selected for each frame are arranged in a matrix where each row stores the xy -coordinates of a point. Then t matrices are stacked and input into the network at the same time.

In the temporal difference layer, we use a convolutional operation to extract features frame-by-frame. The kernel size is $n \times 1$. Let $X = [x_1, x_2, \dots, x_t]$ denote the input facial landmarks, where x_t refers to the landmark points extracted from the t -th facial expression frame. The set $U = [u_1, u_2, \dots, u_t]$ represents a set of convolution kernels and $V = [v_1, v_2, \dots, v_t]$ denotes the features extracted via convolution operation. Features v_t are extracted from x_t using its corresponding convolution kernels u_t ,

$$v_t^s = u_t^s * x_t \tag{2}$$

where $*$ denotes convolution, while v_t^s denotes the s -th feature map of v_t and u_t^s represents the s -th kernel of u_t . The convolution operation is followed by a batch normalization layer and a ReLU activation layer. Then, we use the feature obtained in this frame minus the features obtained in the previous frame to obtain frame difference. After that, we concatenate all the differences and flatten them into the one-dimensional vector. Formally, Z represents the concatenation output, and C is the concatenation operation. Here we have:

$$Z = C(u_2 - u_1, u_3 - u_2, \dots, u_t - u_{t-1}) \tag{3}$$

Then, the difference layers are passed through the two fully connected layers and finally classified using softmax function. The discussion on convolution kernel size and the hyper-parameter t is detailed in Sect. 4.4.

3.3 Model Fusion

We fuse the two networks together through a fusion function referring to the fusion method of Zhang et al. [29].

$$O(x) = \sum_{i=0}^1 a_i(\beta A_i(x) + P_i(x)) \quad (4)$$

$P_i(x)$ ($0 < P_i(x) < 1$) is the output of the softmax layer in the CCRAN and TGF DN. $P_0(x)$ comes from CCRAN and $P_1(x)$ comes from TGF DN. $A_i(x)$ is sorted according to the predicted value of each expression in $P_i(x)$. In addition, β ($0 \leq \beta \leq 1$) acts as a weight parameter. When the value of β is close to 1, the fusion function will give priority to the sorting result of different expressions. When the value of β is close to 0, the fusion function will be a simple weighted-sum function. Finally, a_i is the balance factor between different models. We empirically set a_i to 0.5 and β to 0.1. This function considers the sorting results of the softmax output and actual value of the softmax output simultaneously.

4 Experiments

We evaluated the performance of our model on two widely used databases, including CK+ [18] and Oulu-CASIA [30]. The process and details of the experiments are shown in this section.

4.1 Implementation Details

The structure of CCRAN is I64-[B(5,64)+P2] \times 4-FC1024-S7. I64 means that the size of input frames is 64×64 , and B(5,64) refers to a cross-channel SE block with 64 channels and filters of size 5×5 . Moreover, P2 refers to a 2×2 max pooling layer and FC1024 means a fully connected layer with 1024 nodes. The structure of TGF DN is L(68,2)-C((1,3),16)-FD-FC600-S7. L(68,2) means that landmarks of a frame are reshaped to 68×2 for input, and C((1,3),16) means a convolution operation with 16 output channels and filters of size 1×3 . Moreover, FD means a frame subtraction layer and FC600 means a fully connected layer with 600 nodes. At last S7 is the softmax layer with seven outputs (in CK+ database).

4.2 Databases and Protocols

The CK+ Database. The CK+ database [18] is a representative database of facial expression recognition tasks. This database has a total of 539 sequences of facial expressions, corresponding to 123 subjects with different ages and genders. Among them, 327 expression sequences are marked and correspond to seven types: anger, contempt, disgust, fear, happiness, sadness, and surprise. Each expression sequence begins with a plain frame (neutral expression) and ends with the peak frame of expression. We follow the usual protocol of using 10-fold cross validation [8, 15] for testing.

The Oulu-CASIA VIS Database. There are 80 individuals in this database. Each individual has six expressions, including anger, disgust, fear, happiness, sadness, and surprise. So the database has a total of 480 expression sequences. Like the CK+ database, we use 10-fold cross validation as our experimental method.

4.3 Data Preprocessing and Augmentation

The duration of the expression is not the same, but our network needs to accept a fixed-length image sequence as input. Therefore, we use the average sampling method to regularize the expression sequence along the time axis. From these sampled frames, the faces are detected, cropped and reshaped into 64×64 . What's more, we use dlib [9] to further extract 68 facial landmarks. Then we regularize all the facial landmark points using the method described in [8]. We also follow the method of Jung et al [8], making data augmentation to the training data to alleviate the overfitting problem.

4.4 Experiment Results

Comparison with Other Methods. On the CK+, we can see that our method is very close to state-of-the-art [29] and better than three pre-trained models. The method with * in Table 1 indicates that these methods use the face recognition database for pre-training and the facial expression database for fine-tuning, which introduces additional information to improve the result. On the Oulu-CASIA database, our method has also achieved very good results. The recognition ability of the fused network is higher than VGG-16 pre-trained network. Moreover, the recognition result obtained by CCRAN, which only uses the image frame as input, is surprisingly higher than the DTAGN, which uses both image frames and landmark points as input for recognition on the Oulu-CASIA database. It should be noted that there is no contradiction between our approach and the state-of-the-art [29]. It is very likely to further improve the performance by simply integrating the CCRA mechanism and the frame difference mechanism into the network to form a complementary relationship with our method.

Analysis and Discussion

Region-Wise Squeeze-and-Excitation Blocks. As we can see in Table 2, by adding the cross-channel region-wise attention (CCRA) mechanism to the Resnet block, the network performs better on two databases. This result shows that recalibration of the different region in feature maps can effectively help the network to learn facial expression features.

Facial Landmark Selection. The coordinates of facial landmarks extracted using the dlib [9] can only be integers, which are not accurate and can cause noise in the result. If the sampling frequency of expression frames is too high, the noise

Table 1. Comparisons of different methods on the CK+ and Oulu-CASIA database (where * indicates that the model use face recognition database for pre-training).

Method	Accuracy(CK+)	Accuracy(Oulu)
3DCNN [14]	85.9%	-
3DCNN-DAP [14]	92.4%	-
DTAN [8]	91.44%	74.38%
DTGN [8]	92.35%	74.17%
DTAGN(Weighted Sum) [8]	96.94%	80.62%
DTAGN(Joint) [8]	97.25%	81.46%
PHRNN-MSCNN [29]	98.50%	86.25%
VGG-16 Fine-Tune* [2]	89.9%	83.26%
FN2EN* [2]	96.8%	87.71%
GCNet* [10]	97.93%	86.39%
CCRAN	95.48%	81.58%
TGFDN	94.55%	77.38%
CCRAN-TGFDN	98.11%	83.54%

Table 2. Comparisons between resnet block and cross-channel region-wise attention block on the CK+ and Oulu-CASIA database.

Method	Explanation	Accuracy(CK+)	Accuracy(Oulu)
Baseline	Resnet block	94.39%	79.91%
CCRAN	CCRA block	95.48%	81.58%

Table 3. Comparisons between different input number and filter size of TGFDN on the CK+ and Oulu-CASIA database.

Input Size	Filter size	Accuracy(CK+)	Accuracy(Oulu)
7-frames	1×3	93.68%	74.12%
3-frames	2×2	92.99%	75.54%
3-frames	1×1	93.61%	77.13%
3-frames	1×3	94.55%	77.38%

will be large after frame difference operation. As shown in Table 3, we can see that using landmarks with only three frames ($t = 3$) for recognition has achieved better result than that with 7 frames. In addition, we also tried different filter sizes in the network. Through the display in Table 3, we can see that the results using 2×2 size filters on CK+ and Oulu-CASIA are significantly lower than the other two convolution kernels. We think the reason is that the correlation between the x -coordinate and the y -coordinate of the face landmark points is relatively small. So a single-column-size filter performs better.

Table 4. Confusion matrix of CK+ database.

	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	97.78	0	1.69	0	0	0	0
Contempt	2.22	94.44	0	0	0	3.57	0
Disgust	0	0	98.31	0	0	0	0
Fear	0	0	0	92	0	0	0
Happy	0	0	0	4	100	0	0
Sadness	0	5.56	0	4	0	96.43	0
Surprise	0	0	0	0	0	0	100

Table 5. Confusion matrix of Oulu-CASIA database.

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	78.75	20	0	0	8.75	0
Disgust	12.50	70	12.5	0	2.5	0
Fear	0	0	80	2.5	2.5	8.75
Happy	1.25	0	6.25	97.5	1.25	0
Sadness	7.50	8.75	6.25	0	85	1.25
Surprise	0	1.25	6.25	0	0	90

Confusion Matrix. Tables 4 and 5 show the confusion matrices for our algorithm on the CK+ and Oulu-CASIA databases, respectively. The abscissa of the table represents prediction results and the ordinate represents labels. We can see that in the CK+ and Oulu-CASIA databases, the performance of our model for the fear is relatively poor, but the performance for happy and surprise is good.

5 Conclusion

In this paper, we try to improve the accuracy of expression recognition by capturing subtle facial movements. We propose CCRAN to extract the continuous, region-based, spatial appearance expression information and construct TGFDN to obtain temporal, global-based geographic expression features. After we fused these two networks, our model achieved better results on two different databases. In addition, other popular network structure may also explore the relationship between different areas of the feature map by simply adding the cross-channel region-wise attention mechanism. Therefore, our method is novel, effective, and general.

Acknowledgments. This work was supported by National Natural Science Foundation of China (U1611461), Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase, No. U1501501), and Science and Technology Program of Guangzhou (No. 201803030029).

References

1. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Real time face detection and facial expression recognition: development and applications to human computer interaction. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop, vol. 5, pp. 53–53. IEEE (2003)
2. Ding, H., Zhou, S.K., Chellappa, R.: FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 118–126. IEEE (2017)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint [arXiv:1709.01507](https://arxiv.org/abs/1709.01507) (2017)
6. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261–2269 (2017)
7. Jeni, L.A., Lórinicz, A., Szabó, Z., Cohn, J.F., Kanade, T.: Spatio-temporal event classification using time-series kernel based structured sparsity. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 135–150. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_10
8. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: IEEE International Conference on Computer Vision, pp. 2983–2991. IEEE (2015)
9. Kazemi, V., Josephine, S.: One millisecond face alignment with an ensemble of regression trees. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874. IEEE (2014)
10. Kim, Y., Yoo, B., Kwak, Y., Choi, C., Kim, J.: Deep generative-contrastive networks for facial expression recognition. arXiv preprint [arXiv:1703.07140](https://arxiv.org/abs/1703.07140) (2017)
11. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: British Machine Vision Conference, p. 275-1. British Machine Vision Association (2008)
12. Liu, M., Li, S., Shan, S., Chen, X.: AU-aware deep networks for facial expression recognition. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–6. IEEE (2013)
13. Liu, M., Li, S., Shan, S., Chen, X.: AU-inspired deep networks for facial expression feature learning. *Neurocomputing* **159**, 126–136 (2015)
14. Liu, M., Li, S., Shan, S., Wang, R., Chen, X.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9006, pp. 143–157. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16817-3_10
15. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756. IEEE (2014)
16. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
18. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 94–101. IEEE (2010)
19. Ofodile, I., et al.: Automatic recognition of deceptive facial expressions of emotion. arXiv preprint [arXiv:1707.04061](https://arxiv.org/abs/1707.04061) (2017)
20. Pantic, M., Rothkrantz, L.J.: Facial action recognition for facial expression analysis from static face images. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **34**(3), 1449–1461 (2004)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
23. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: ACM International Conference on Multimedia, pp. 357–360. ACM (2007)
24. Shan, C., Gong, S., McOwan, P.W.: Conditional mutual information based boosting for facial expression recognition. In: British Machine Vision Conference (2005)
25. Sikka, K., Wu, T., Susskind, J., Bartlett, M.: Exploring bag of words architectures in the facial expression domain. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7584, pp. 250–259. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33868-7_25
26. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. arXiv preprint [arXiv:1505.00387](https://arxiv.org/abs/1505.00387) (2015)
27. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2015)
28. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(1), 39–58 (2009)
29. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutionary spatial-temporal networks. *IEEE Trans. Image Process.* **26**(9), 4193–4203 (2017)
30. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikäinen, M.: Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **29**(9), 607–619 (2011)
31. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 915–928 (2007)



Score-Guided Face Alignment Network Under Occlusions

Xiang Yan, Huabin Wang^(✉), Qi Wang, Jinjie Song, and Liang Tao

Key Laboratory of Intelligent Computing and Signal Processing of Ministry of
Education, Anhui University, Hefei 230031, China
xiang199286@gmail.com, {wanghuabin,taoliang}@ahu.edu.cn,
1054588756@qq.com, 3470755438@qq.com

Abstract. Recent state-of-the-art landmark localization task are dominated by heatmap regression and fully convolutional network. In spite of its superior performance in face alignment, heatmap regression method has a few drawbacks in nature, such as do not follow shape constraint and sensitivity to partial occlusions. In this paper, we proposed a score-guided face alignment network that simultaneously outputs a heatmap and corresponding score map for each landmark. Rather than treating all predicted landmarks equally, a weight is assigned to each landmark based on the two relational maps. In this way, more reliable landmarks with strong local information are assigned large weights and the landmarks with small weights that may stay with occlusions can be inferred with the help of the reliable landmarks. Meanwhile, an exemplar-based shape dictionary is designed to take advantage of these landmarks with high score to infer the landmark with small score. The shape constraint is implicitly applied in this way. Thus our method demonstrates superior performance in detecting landmarks with extreme occlusions and improving overall performance. Experiment results on 300W and COFW dataset show the effectiveness of the proposed method.

Keywords: Face alignment · Fully convolutional network · Occlusion

1 Introduction

Face alignment [5, 25, 40], also known as facial landmark detection, which aims to find the locations of a set of predefined facial landmarks (e.g., mouth, eyes, nose, cheek and so on) in a face image. It is a crucial pre-processing step for face recognition [16, 26, 27], expression recognition [3, 13], face analysis [21] and so on. As a well established problem in computer vision, researchers have proposed many methods and made significant progress in face alignment. Recently, heatmap regression method [4, 6, 10] has shown superior performance on face alignment. However, Face alignment under occlusions still remains unsettled. Especially, when face images suffer from heavy occlusions, the performance of face alignment drops severely.

To address face alignment under occlusions, several methods are proposed to tackle face alignment under partial occlusions. The method of [7] divides face into a 3×3 grid and only draw features from the 1/9 of facial region to several separate regressors. The work in [29] proposes a robust cascaded regression framework to handle large facial pose and occlusion. The landmark locations and the landmark visibility probability are updated stage by stage. The method of [18] treat face alignment as an appearance-shape model problem. They learn two dictionaries which are relational, one for the appearance of human face and one for the facial shape. By the two relational dictionaries, the face appearance is employed to infer occlusion and suppress the influence of occluded landmarks. The work in [33] cascades several Deep Regression networks (DR) and De-corrup Auto-encoders (DA) to explicitly handle partial occlusion problem. In contrast with previous methods that only predict occlusion, the proposed De-corrup Auto-encoders can recover the occluded facial appearance. They divide the facial landmarks to seven components, each specific DA is able to recover the occluded appearance. Although these methods have shown superior performance in aligning occluded faces, they have limited scalability and robustness. First is the lack of large-scale ground truth occlusion annotation for images in the wild. The task of providing occlusion annotation is often time-consuming, involving a considerable amount of tedious manual work. Another challenge is in the inherent complex facial appearance. Generally, the performance of appearance-shape dictionary depends on whether the image patterns reside within the variations described by the face appearance dictionary. Therefore, it shows limited robustness in unconstrained environment where appearance variations are too wide and complicated. In addition, recovering the occluded appearance is not without difficulties.

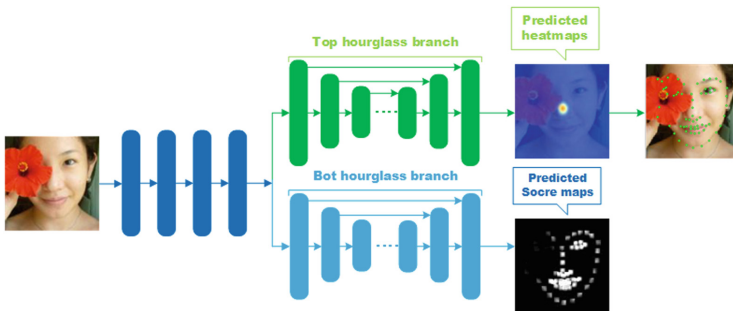


Fig. 1. Papers main idea: Given a face image as input, our network simultaneously outputs heatmaps and score maps. Due to part occlusions, the occluded landmark cannot be located precisely. Observe that the score for the occluded parts is much lower than that of the non-occluded parts in score maps. Based on the two relational maps, the occluded landmarks can be refined with the help of non-occluded landmark by exploiting geometric constraints of face shape.

In this paper, we propose a novel score-guided face alignment network to deal face with large occlusions. The key innovation of our method is score map which is able to dynamically select more reliable landmarks and use these reliable landmarks to refine the landmarks with small score. See Fig. 1 for a graphical representation of our paper’s main idea. The proposed network outputs heatmaps and score maps. The occluded part is obvious in score map and has small score than non-occluded part. Rather than treat all landmarks equally, we assign a weight to each landmark based on heatmaps and score maps and the occluded landmark can be refined with the help of the non-occluded landmarks. More specifically, due to the partial occlusion, the occluded landmark cannot be located precisely. However, the non-occluded landmark can be located precisely. Since the non-occluded landmarks have larger weights than occluded landmarks. An exemplar-based shape dictionary act as shape priors can be utilized to search most similar shapes to reconstruct the face shapes based on the weights of landmarks.

The main contributions of our method can be summarized as follows:

1. We propose a novel face alignment network that simultaneously outputs heatmaps and score maps, which is more robust to occlusions. Note that no occlusion annotations are used.
2. Rather than treating all landmarks equally, we introduce score map to assign weight to each landmark. In this way, more reliable landmarks with large weights can help to refine the occluded landmarks with small weights.

2 Related Work

Prior to deep learning, cascade regression [9, 17, 18, 22, 23, 37] is a popular method in face alignment, it starts with an initial facial shape and refine the shape in a cascaded manner. For each regressor, it learns a mapping function from shape-indexed features to the shape increment. The authors of [31] proposed a method named Supervised Descnet Method (SDM) to learn cascade regressors with strong handcrafted features such as SIFT. The work in [23] proposes learning local binary features by using random forests. Thanks to the sparse binary features, its speed can achieve 3000 FPS. To reduce the influence of inaccurate shape initializations, In [37] a coarse to fine search method is proposed. It begins with a coarse search over a shape pool and employs the coarse solution to finer search of shapes. The authors of [38] reformulates the popular cascaded regression scheme into a cascaded compositional learning (CCL) problem. It divides all training samples into several domains. Each domain-specific cascaded regressor handle one domain. The final shape is a composition of shape estimations across multiple predictions. The method of [11] trains multi-view cascaded regression models using a fuzzy membership weighting strategy, which improving the fault-tolerant of cascade regression. Although cascade regression has achieved good performances on the wild databases, inaccurate shape initializations, independent regressors and handcrafted features still may be sub-optimal for face alignment.

This conventional cascade regression, however, has been greatly reshaped by convolutional neural networks (ConvNets). Recent face alignment methods have universally adopted ConvNets as their main building block, largely replacing hand crafted features. The work in [36] uses multi-stage deep networks to detect facial landmarks in a coarse to fine manner. The authors of [35] formulates a novel tasks-constrained deep model to jointly optimize landmark detection together with the recognition of heterogeneous but subtly correlated facial attributes which improves the performance of landmark detection. The work in [34] employs Autoencoder networks (CFAN) that combined several stacked auto-encoder networks in a cascaded manner. The authors of [28] proposes a convolutional recurrent neural network architecture. The feature extraction stage is replaced with a convolutional network, the fitting stage is replaced with the Recurrent Neural Network. The work in [30] employs an Attention LSTM (A-LSTM) and an Refinement LSTM (R-LSTM), which sequentially selects the attention center by A-LSTM and refines the landmarks around the attention-center by R-LSTM. The authors of [19] presents a deep regression architecture with two stage reinitialization to explicitly deal with the initialization problem by face detection. FAN [6] employs stacked hourglass Network with a state-of-the-art residual block to solve the 2D&3D Face Alignment problem. The work in [10] formulate a novel Multi-view Hourglass Model which tries to jointly estimate both semi-frontal and profile facial landmarks.

3 Methodology

3.1 Network Architecture

Here, we describe our network architecture based on hourglass [20] backbone. The input is a face image with spatial resolution 128×128 . The network starts a 7×7 convolutional layer with stride 2 and padding 3 to process the image to spatial resolution 64×64 , followed by three residual blocks [14] to increase feature channels. Then the network is split in two sub-branches. The top sub-branch is a hourglass network, which is a symmetric top-down and bottom-up full convolutional network. Then two residual blocks process the feature maps to 128 channels. After that, nearest neighbor upsampling is used to increase the spatial resolution to 128×128 , followed by a residual block and a convolutional layer with 1×1 kernels to produce heatmaps. The bottom sub-branch has the same network structure with the top sub-branch. Batch Normalization is used to before all convolutional layers except the first convolutional layer with kernels 7×7 . ReLU is the activation function. In summary, the input of network is a face image with spatial resolution 128×128 . The network output N heatmaps and N score maps, where N is the number of landmarks. Each landmark corresponds to a heatmap and a score map (Fig. 2).

3.2 Score Map and Heatmap

Heatmaps are extensive used in landmark localization tasks. The model outputs N heatmaps where N is the number of landmarks. The pixel with the high-est

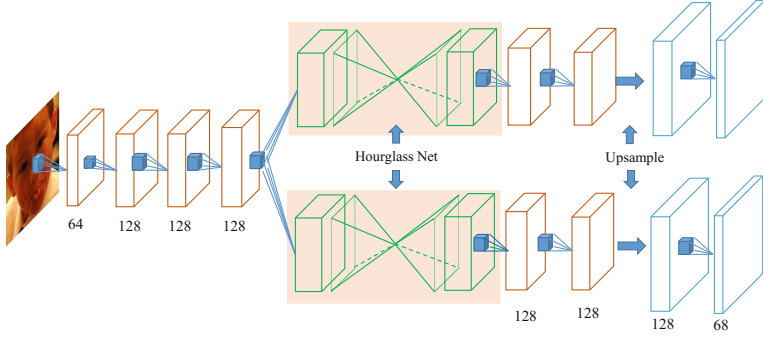


Fig. 2. An illustration of our network architecture.

value is used as the predicted landmark location. Great progress has been made by heatmaps. However, the landmarks with partial occlusion and complex background still cannot be precisely located. To deal with occlusions, we introduce score maps to assign weight to each landmark and suppress the influence of occlusions. During training, Heatmap for one landmark is created by putting a Gaussian peak at ground truth location of the landmark. While the score maps are binary maps, the values within a certain radius around the ground truth locations are set to 1 and the value for the remaining are set to 0. See Fig. 3 for example outputs produced by our network. The non-occluded face part has higher score than the occluded-part in score map. Rather than treating all landmarks equally, we weight each landmark based on their values in score maps. In this way, more reliable landmarks with strong local information are assigned high weights. The landmarks with small weights that may stay with occlusions can be refined with the help of reliable landmarks. Based on the two relational maps, the process of assigning weight can be written via the equation

$$w_i = \frac{\sum_{k=X_i-r}^{X_i+r} \sum_{t=Y_i-r}^{Y_i+r} score_i(k, t)}{(2 * r + 1)^2}. \quad (1)$$

where $score_i(k, t)$ is the value of coordinate (k, t) in i -th score map. X_i and Y_i are the predicted locations of i -th landmark.

3.3 Face Shape Reconstruction

Based on the two relational maps, the weight of each landmark can be determined. For the non-occluded face images, the heatmaps and score maps assign high weights to each landmark. The final predicted face shape is the locations decoded from heatmaps. For the heavy occluded face images, score maps only can check out these inaccurate landmarks with small weights, these landmarks still cannot be accurately located. Intuitively, the predicted face shape should

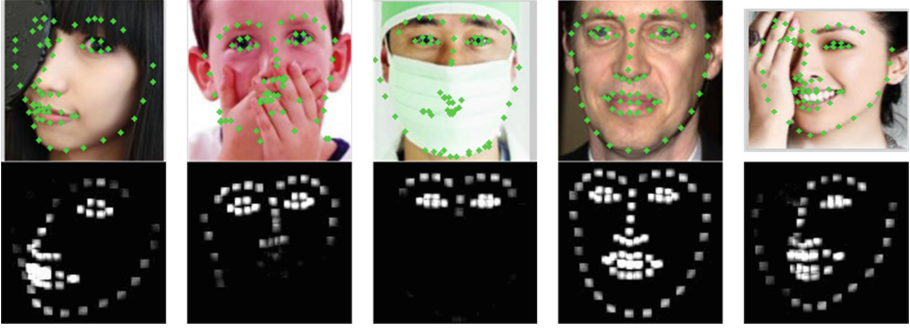


Fig. 3. Example outputs produced by our proposed network. First row shows landmark locations decoded by heatmaps. Second row shows the proposed score maps. Observe that the occluded landmarks cannot be precisely located in most cases. The non-occluded parts in score maps have higher score and are clearer than the occluded parts.

look like a face shape. Human vision has ability to predict good face shape by exploiting geometric constraints. Motivated by this, these inaccurate landmarks caused by occlusions can be refined by searching the most similar face shapes based on non-occluded landmarks, which is feasible and simple.

However, searching from all training samples is time-consuming. There are a lots of similar face shapes which are redundant. Assuming there are M training samples in train set. When M is large, searching from all training samples would be time-consuming. Follow [18], We apply K-SVD [1] on all training shapes to get N representative face shapes and use these face shapes as a shape dictionary D_S . Searching from D_S will be more effective. The searching process is formally written as

$$\min_{s_1 \dots s_k} \|W^S S - (W^S S \odot W^S D_S)\|_2^2 \quad (2)$$

where $W^S = \text{diag}(w_1, \dots, w_N, w_1, \dots, w_N)$ is the weight matrix and the w_i is the weight of the i -th landmark calculated via Eq.1. The goal of W is to force the search process to emphasize on the landmarks with high weights and ignore the landmark with small weights. $s_1 \dots s_k$ are the k nearest exemplar shapes of the non-occluded landmarks. After that, the occlusions landmarks can be reconstructed by the k nearest exemplar shapes and the reconstruction coefficients can be computed by least squares method (Fig.4).

3.4 Training Details

During training, to prevent overfitting, all training samples are augmented by random in-plane rotation (from -30° to $+30^\circ$), translation, scale (from 0.9 to 1.2), flip and adding color jittering. The network input is a RGB image of size 128×128 . The network is optimized by RMSProp with an initial learning rate of 0.0001 and drop to 0.00005 after 20 epochs. All models are trained using

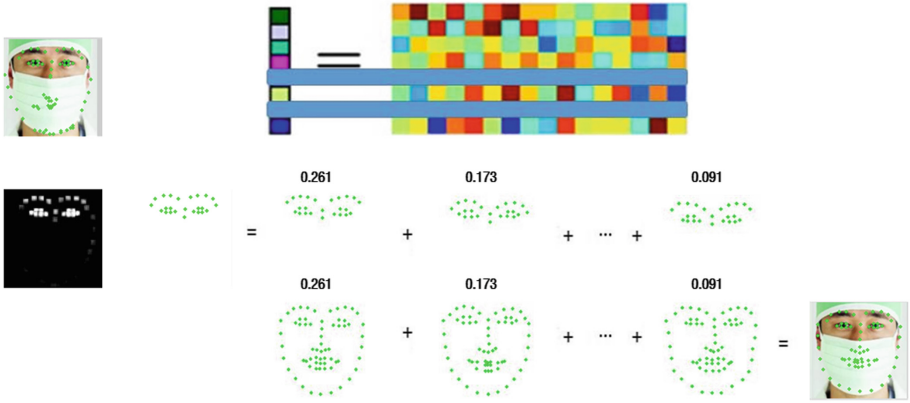


Fig. 4. Face shape reconstruction by the k nearest exemplar shapes.

PyTorch with a Nvidia 1080-Ti GPU card with a mini-batch size of 10 for 80 epochs. The loss function is defined as

$$Loss = \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \|H_k - \hat{H}_k\|_2^2 + \frac{\lambda}{2} \sum_{n=1}^N \sum_{k=1}^K \|S_k - \hat{S}_k\|_2^2 \quad (3)$$

where N is the number of training samples, H_k and \hat{H}_k are the predicted heatmaps and the ground-truth heatmaps. S_k and \hat{S}_k are the predicted score map and ground-truth score map. λ is a hyperparameter to balance the loss functions. During inference, the predicted landmark locations \hat{Z}_k is decoded from the predicted heatmap H_k by taking the locations with the maximum value as follows,

$$\hat{Z}_k = \arg \max_p H_k(p) \quad (4)$$

4 Experiments

4.1 Datasets

For training, 300-W is the most widely-used in-the-wild dataset for 2D face alignment. All face images are labeled by 68 landmarks. The training set consists of AFW [39] dataset, HELEN [15] training set and LFPW [2] training set, there are 3148 face images in total. For testing, we report the results on LFPW testing set, Helen testing set and IBUG dataset. To verify the effectiveness of our method on occluded faces, we evaluate COFW [7, 12] testing set. The COFW dataset is a challenging dataset with severe facial occlusions and large facial pose collected from web. There are 1345 face images in training set and 507 face images in testing set. All face images are labeled by 29 landmarks. Since our model is trained on images with 68 landmarks, Follow [12], we use the COFW with 68 landmarks for testing. Note that we only use COFW testing set for evaluation.

4.2 Metrics

Given the predicted landmark locations and ground-truth landmark locations, the Normalized Mean Error (NME) or cumulative error distribution (CED) curves employed to evaluate the localization performance. The normalization is normalized by inter-pupil distance and the NME is computed as follows:

$$error = \frac{1}{M} \sum_{i=1}^M \frac{\frac{1}{N} \sum_{j=1}^N \|p_{i,j}^{pred} - p_{i,j}^{gt}\|_2}{\|p_{i,l} - p_{i,r}\|_2} \quad (5)$$

where M is the number of testing images, N is the number of landmarks. $p_{i,l}, p_{i,r}$ are the locations of left eye center and right eye center in i -th face image. $p_{i,j}^{pred}$ is the predicted location of landmark location of the j -th landmark in i -th face image. $p_{i,j}^{gt}$ is the ground-truth location of landmark location of the j -th landmark in i -th face image.

4.3 Evaluation Results on 300W

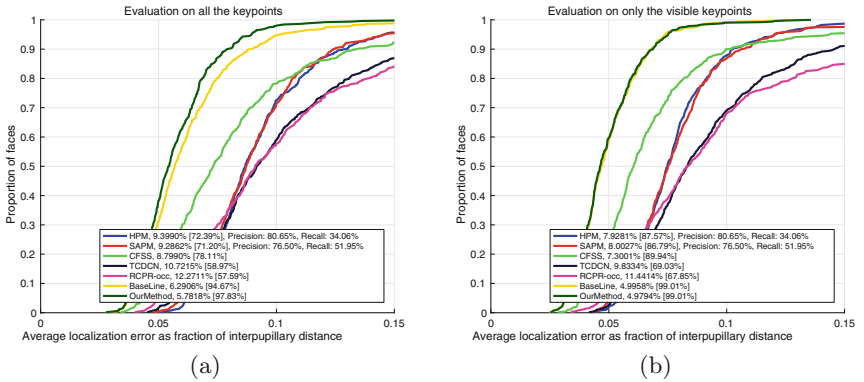
The 300-W [24] testing set consists of common set and challenging set. The common set are Helen testing set and LFPW testing set. The challenging set is the IBUG dataset. Table 1 show the results on 300W dataset. We compare our method with eleven state-of-the-art face alignment methods with RCPR [7], CFAN [34], ESR [8], SDM [31], LBF [22], CFSS [37], TCDCN [35], DNN [32], MD-M [28], RAR [30], TR-DRN [19]. Our method outperform most of these methods except RAR.

4.4 Evaluation Results on COFW

To verify the effectiveness of our method on various occluded face images, we test our method on COFW [7, 12] dataset. The CED curves are shown in Fig. 5. It can be seen our baseline still outperform all other methods by a large margin. That is because our method benefits from heatmap regression and network architecture. By adding occlusion inference and face reconstruction, the NME error decreases from 6.29% to 5.78%. The success rate increases from 94.67% to 97.83%. Moreover, we analyse the evaluation on only the visible landmarks, our method and baseline show similar results on NME error and success rate. It can be concluded that heatmap regression method achieves excellent performance in detecting non-occluded face part. While evaluation on all the landmarks, benefit from score map to assign weight to each landmark and refine the occluded region by face reconstruction, our method show better results than baseline both in NME error and success rate.

Table 1. Landmark detection results on different subsets of the 300-W dataset in terms of the NME averaged over all the test samples.

Method	Common set	Challenging set	Full set
RCPR	6.18	17.26	8.35
SDM	5.57	15.40	7.52
ESR	5.28	17.00	7.58
CFAN	5.50	16.78	7.69
DeepReg	4.51	13.80	6.31
LBF	4.95	11.98	6.32
CFSS	4.73	9.98	5.76
TCDCN	4.80	8.60	5.54
DDN	-	-	5.59
MDM	4.83	10.14	5.88
RAR	4.12	8.35	4.94
TR-DRN	4.36	7.56	4.99
SIR	4.29	8.14	5.04
Ours	4.16	7.54	4.78

**Fig. 5.** Comparison of different models on the COFW dataset: (a) evaluation on all the keypoints, (b) evaluation on only the visible keypoints.

5 Conclusion

In this paper, we propose a score-guided face alignment network which is robust to occlusions. The network simultaneously outputs a heatmap and corresponding score map for each landmark. Based on the two relational maps, more reliable landmark are assigned large weights and landmarks with small weights can be inferred with the help of the reliable landmarks. Experiment results on 300 W

and COFW dataset show the effectiveness of the proposed method and showed significant performance improvements over the state-of-the-arts.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing over-complete dictionaries for sparse representation. *IEEE Trans. Sig. Process.* **54**(11), 4311–4322 (2006)
2. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: *Computer Vision and Pattern Recognition (CVPR)*, pp. 545–552 (2011)
3. Bettadapura, V.: Face expression recognition and analysis: the state of the art (2012). arXiv preprint [arXiv:1203.6722](https://arxiv.org/abs/1203.6722)
4. Bulat, A., Tzimiropoulos, G.: Convolutional aggregation of local evidence for large pose face alignment (2016)
5. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: *The IEEE International Conference on Computer Vision (ICCV)*, vol. 1, p. 4 (2017)
6. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D and 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: *International Conference on Computer Vision (ICCV)*, vol. 1, p. 4 (2017)
7. Burgos-Artizzu, X.P., Perona, P.: Robust face landmark estimation under occlusion. In: *International Conference on Computer Vision (ICCV)*, pp. 1513–1520 (2013)
8. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. *Int. J. Comput. Vis.* **107**(2), 177–190 (2014)
9. Deng, J., Liu, Q., Yang, J., Tao, D.: M3 CSR: Multi-view, multi-scale and multi-component cascade shape regression. *Image Vis. Comput.* **47**, 19–26 (2016)
10. Deng, J., Trigeorgis, G., Zhou, Y., Zafeiriou, S.: Joint multi-view face alignment in the wild (2017). arXiv preprint [arXiv:1708.06023](https://arxiv.org/abs/1708.06023)
11. Feng, Z.H., Kittler, J., Christmas, W., Huber, P., Wu, X.J.: Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3681–3690. IEEE (2017)
12. Ghiasi, G., Fowlkes, C.C.: Occlusion coherence: detecting and localizing occluded faces (2015). arXiv preprint [arXiv:1506.08347](https://arxiv.org/abs/1506.08347)
13. Guo, Y., Zhao, G., Pietikäinen, M.: Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Trans. Image Process.* **25**(5), 1977–1992 (2016)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016)
15. Le, V., Brandt, J., Bourdev, L., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: *European Conference on Computer Vision (ECCV)*, pp. 679–692 (2012)
16. Li, D., Zhou, H., Lam, K.M.: High-resolution face verification using pore-scale facial features. *IEEE Trans. Image Process.* **24**(8), 2317–2327 (2015)
17. Liu, Q., Deng, J., Tao, D.: Dual sparse constrained cascade regression for robust face alignment. *IEEE Trans. Image Process.* **25**(2), 700–712 (2016)

18. Liu, Q., Deng, J., Yang, J., Liu, G., Tao, D.: Adaptive cascade regression model for robust face alignment. *IEEE Trans. Image Process.(TIP)* **26**(2), 797–807 (2017)
19. Lv, J.J., Shao, X., Xing, J., Cheng, C., Zhou, X., et al.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, p. 4 (2017)
20. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
21. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, pp. 17–24. IEEE (2017)
22. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment via regressing local binary features. *IEEE Trans. Image Process.(TIP)* **25**(3), 1233 (2016)
23. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692 (2014)
24. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: *Conference on Computer Vision Workshops (CVPRW)*, pp. 397–403 (2014)
25. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: *Computer Vision and Pattern Recognition (CVPR)*, pp. 3476–3483 (2013)
26. Tai, Y., Yang, J., Zhang, Y., Luo, L., Qian, J., Chen, Y.: Face recognition with pose variations and misalignment via orthogonal procrustes regression. *IEEE Trans. Image Process.* **25**(6), 2673–2683 (2016)
27. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708 (2014)
28. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: a recurrent process applied for end-to-end face alignment. In: *Computer Vision and Pattern Recognition (CVPR)*, pp. 4177–4187 (2016)
29. Wu, Y., Ji, Q.: Robust facial landmark detection under significant head poses and occlusion. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3658–3666 (2015)
30. Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A.: Robust facial landmark detection via recurrent attentive-refinement networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9905, pp. 57–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_4
31. Xiong, X., Torre, F.D.L.: Supervised descent method and its applications to face alignment. In: *Computer Vision and Pattern Recognition (CVPR)*, pp. 532–539 (2013)
32. Yu, X., Zhou, F., Chandraker, M.: Deep deformation network for object landmark localization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9909, pp. 52–70. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_4
33. Zhang, J., Kan, M., Shan, S., Chen, X.: Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders. In: *Computer Vision and Pattern Recognition (CVPR)*, pp. 3428–3437 (2016)

34. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 1–16. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_1
35. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 94–108. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_7
36. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 386–391 (2013)
37. Zhu, S., Li, C., Chen, C.L., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Computer Vision and Pattern Recognition (CVPR), pp. 4998–5006 (2015)
38. Zhu, S., Li, C., Loy, C.C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3409–3417 (2016)
39. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE (2012)
40. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, pp. 146–155 (2016)



Robust Face Detector with Fully Convolutional Networks

Yingcheng Su, Xiaopei Wan, and Zhenhua Guo^(✉)

Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
zhenhua.guo@sz.tsinghua.edu.cn

Abstract. Many of the exist face detection algorithms are based on the generic object detection methods and have achieved desirable results. However, these methods still struggle in solving the problem of partial occluded face detection. In this paper, we introduce a simple and effective face detector which uses a fully convolutional networks (FCN) for face detection in a single stage. The proposed FCN model is used for pixel-wise prediction instead of anchor mechanism. In addition, we also apply a long short term memory (LSTM) architecture to enhance the contextual information of feature maps, making the model more robust to occlusion. Besides, we use a light-weighted neural network PVANet as the backbone, which greatly reduces the computational burden. Experimental results show that the proposed method achieves competitive results with state-of-the-art face detectors on the common face detection benchmarks, including the FDDB, WIDER FACE and MAFA datasets, what's more, it is much more robust to the detection of occluded faces.

Keywords: Face detection · FCN · LSTM · Occlusion

1 Introduction

Face detection has always been a research hotspot as it is a crucial step of many facial applications, such as face alignment, face recognition, etc. Since the pioneering work of Viola-Jones face detector [1], a lot of face detection methods have been proposed. The hand-crafted features [2, 3] usually rely on prior knowledge leading to poor performance in complex scenes, especially faces with occlusion.

In recent years, convolutional neural networks (CNNs) have great success in the field of computer vision, including image classification [4, 5] and object detection [6–9], etc. The Object detection algorithms such as fast [6]/faster [7] R-CNN, SSD [9], YOLO [8] continue to make new breakthroughs in both speed and precision. Face detection is a special case of object detection. Many face detection approaches are based on object detection methods [10–13] and achieve promising results. However, these anchor-based methods are badly rely on the

Z. Guo—The work is partially supported by the Natural Science Foundation for China (NSFC) (No. 61772296) and Shenzhen fundamental research fund (Grant Nos. JCYJ20160531194840025 and JCYJ20170412170438636).



Fig. 1. Our face detector is robust to heavy occlusion and large appearance.

number of matching proposals. If the faces are partial occluded, it's very likely that the models would miss the proposals of occluded faces or be confused by the features of occluded faces. The cascaded network [17, 18] is another type of CNNs-based face detection approach. Several small CNNs are cascaded to detect faces in a coarse-to-fine manner. In spite of very fast speed, these shallow networks failed to represent robust image features to handle faces with occlusion.

Inspired by [20], we consider face detection problem as the combination of binary classification and bounding box regression. In this paper, we propose a fast and efficient face detector that only need two steps for face detection. First, a FCN is used to do the pixel-wise classification and bounding box regression. Then, the produced face predictions are sent to Non-Maximum Suppression (NMS) to yield final results. By making such dense predictions, the model has strong robustness to faces with occlusion. In addition, considering the highly-correlated of adjacent regions of the feature map, we use an in-network recurrent architecture to encode rich context information of the feature map. Even if the face is partial occluded, the model can make the correct predictions from the non-occluded part. An example of our detection results can be found in Fig. 1.

The main contribution of this paper can be summarized as:

- We propose a novel FCN-based face detection method that directly make dense predictions in feature maps. The proposed method is fast, accurate and quite simple, which only consist of two step: a forward propagation of the FCN and a NMS merging.
- We use a recurrent architecture to connect the context information of the feature maps, improving the model's capacity of detecting faces with occlusion.

- The proposed method achieves competitive results in FDDB, WIDER Face datasets, and outperforms state-of-the-art methods in occluded faces datasets like MAFA.

2 Related Work

Before the revolution of deep learning, Face detection has been widely studied. Numerous face detector are based on traditional machine learning methods. The pioneering work of Viola-Jones [1] utilizes Adaboost with Haar-like feature to train a cascade model to detect face and get real-time performance. Since then the studies of face detection focus on designing more efficient features [22, 23] and more powerful classifiers [26, 27]. Deformable pattern models (DPM) [25] are employed for face detection task and achieve promising results. Liao et al. [24] proposed normalized pixel difference (NPD) features and constructed a deep quadratic tree to handle unconstrained face detection. However, these hand-crafted features always require prior assumptions which would be untenable in complex scenarios, leading to low precision in the challenging face datasets, such as WIDER Face and MAFA.

In recent years, the CNN-based face detectors achieved remarkable performance. Li et al. [17] use cascaded CNNs for face detection. Zhang et al. [18] propose Multi-task cascaded CNNs (MTCNN) to detect face and align face, simultaneously. Qin et al. [19] integrate the training of cascaded CNNs into a framework for end-to-end training, which greatly improves the performance of cascaded networks. Faceness [28] generates face parts responses from attribute-aware networks to detect faces under occlusion and unconstrained pose variation. However, this method needs to label facial attributes of different facial parts and generate face proposals according to facial part response maps, which is complicated and time consuming.

There are also a variety of face detection methods that inherit the achievements from generic object detection methods. Face R-CNN [12] is based on Faster R-CNN and adopts center loss [29] to minimize the intra-class distances of the deep features. It also utilizes some training tricks such as online hard example mining and multi-scale training. CMS-RCNN [10] uses contextual information for face detection. DeepIR [13] concatenate features of multiple layers to improve face detection performance. Hu et al. [16] build image pyramids and defines multiple templates to find tiny faces. SSH [14] establishes detection modules on different feature maps to detect face in a single stage. SFD [15] focuses on scale-invariance by using a new anchor matching strategy. Zhu et al. [30] analyze the anchor matching mechanism with the proposed expected max overlap (EMO) score and introduce new designed anchors to find more tiny faces. All these anchor-based methods have obtained promising results. However, we know that the scale of faces is continuous. The anchor mechanism makes the scale discrete, which may lead to the low matching rate of hard samples, especially occluded faces. A naive way to increase the number of matching anchors is to increase the total number of anchors. But this will result in heavily computational burden.

DenseBox [20] is another kind of object detection method. Different from the above anchor-based methods, DenseBox utilizes a FCN to perform pixel-wise predictions. By doing the upsampling operation to keep a high-resolution output, it has great advantages in handling the detection of small objects. The approach of dense prediction can also improve the robustness of detecting heavy occluded objects. UnitBox [21] further presents a new intersection-over-union (IoU) loss for bounding box prediction. Yet there are some drawbacks of UnitBox. On one hand, an up-sample layer is used to perform linear interpolation to resize the feature map to the original image size. Although it can detect smaller faces, the computational cost is unacceptable. On the other hand, the feature maps are upsampled 16 times for pixel-wise classification, which may bring artifacts. In this paper, we propose a novel face detector that utilizes a FCN framework to do the dense prediction on the feature maps whose size is just $1/4$ of the original image size. The FCN architecture consists of a bottom-up path and a top-down path similar to [20, 31]. Inspired by [32], we further employ an in-network recurrence mechanism to explore meaningful information of the convolutional feature maps and improve the robustness of detecting faces with occlusion, leading to state-of-the-art detection performance.

3 Proposed Method

The proposed face detector is trained to directly predict the existence of faces and their locations from full images instead of dividing the detection task into bounding box proposal and classification. A fully convolutional neural network is used to do the pixel-wise dense prediction of faces. The post-processing of our method is quit simple, which only contains thresholding and NMS.

3.1 Base Framework

As we know from [33] that feature maps of different layer represent different semantic information. The shallow layers have high spatial resolution responding to corners and edge/color conjunctions, which is good for spatial localization. The deep layers have lower spatial resolution but more class-specific which is good for classification. Inspired by recent works [20, 31, 34], we adopt a neural network that contains a top-down architecture with lateral connection to fuse features from different layers.

Our network architecture is shown in Fig. 2. We use PVANet [35] as the backbone. The bottom-up pathway is the feed-forward computation of the backbone ConvNet generating four levels of feature maps, whose sizes are $1/4$, $1/8$, $1/16$ and $1/32$ of the original image, respectively. We define that layers producing the output maps of the same size are in the same network stage. Since the deeper layer should have stronger features, the last layer of each stage is chosen to connect with deeper layer with the same output size. It is very difficult to detect tiny object by low resolution features. The top-down pathway increases the resolution by upsampling operations while keeps the semantic information. Each

upsample operation is at a scaling step of 2. The top-down pathway features are enhanced by features from the bottom-top pathway via lateral connections. By doing such lateral connections, the network can maintain both geometrical and semantic information. As shown in Fig. 2, we use a 1×1 conv layer to preprocess the lateral features and merge different features by concat layer. Then a 1×1 conv layer and a 3×3 conv layer are used to further cut down half of the number of channels and produce the output of this merging stage, respectively. The size of the final feature maps is only $1/4$ of the original image, making the network computation-efficient. The network is then split into two branches, one for classification and the other one for bounding box regression.

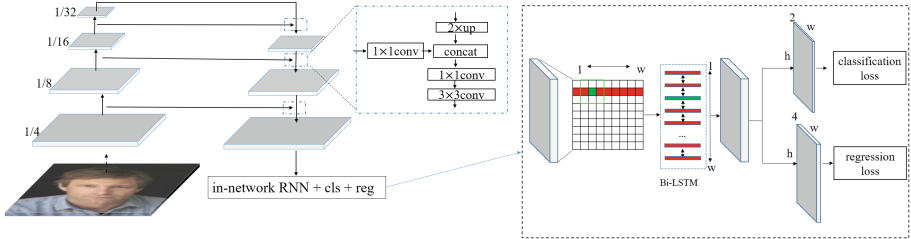


Fig. 2. An overview of our network architecture

3.2 In-Network Recurrence Architecture

Recurrent neural network (RNN) is often applied in scenarios with sequences of inputs such as video, audio, text lines to encode the contextual information. Recent work [32] has shown that the sequential context information is good for text detection. Motivated from this work, we believe that RNN may also benefit for face detection, especially detecting faces with occlusion. We note that features of the face area are highly-correlated, so we can use this correlation via recurrent structure to make correct predictions of the occluded part of face. Besides, the regression task predicts a 4-D distance vector (the distances between the current pixel and the four bounds of the ground truth box), and there is also a strong correlation among the distance vectors of adjacent pixels. RNN can encode these contextual information recurrently using its hidden layers. Formally, The internal state of RNN at t moment is given by

$$H_t = \varphi(H_{t-1}, X_t) \quad (1)$$

where $X_t \in R^{3 \times 3 \times C}$ is the input sequential features from t -th sliding-window (3×3) as shown in Fig. 2. The sliding window slides from left to right at a stride of 1, generating $t = 1, 2, \dots, W$ sequential inputs for each row. W is the width of the input feature map. In this paper, we adopt the bi-directional long short-term memory (Bi-LSTM) architecture for the RNN layer just as [32] do. The Bi-LSTM allows the model to encode the contextual features in both directions. The outputs of the two inverse LSTMs is then merged by a concat layer, followed by a 1×1 conv layer to cut down the number of channels.

3.3 Label Generation

We consider the face area is a rectangle. The classification task is to predict a binary score map $\in \{0, 1\}$ which indicates the negative area and positive area. The positive area of the rectangle on the score map is designed to be roughly a shrunk version of the original rectangle. For each edge, we shrink it by moving its two endpoints inward along by 0.2 of its length, illustrated in Fig. 3(a). The regression task is to predict a 4 channels of distance map as shown in Fig. 3(d). The ground truth distance map is generated by calculating a 4-D distance vector for each pixel with a value of 1 on the score map, illustrated in Fig. 3(c).

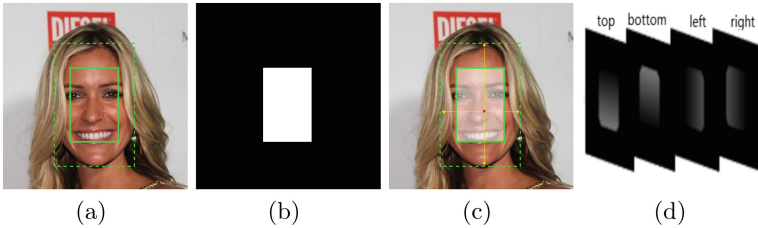


Fig. 3. Label generation. (a) Face bounding box (green dashed) and the shrunk rectangle (green solid); (b) score map; (c) pixel-wise distances generation; (d) 4 channels of distances of each pixel to rectangle boundaries. (Color figure online)

4 Training

In this section, we introduce our training details, including loss function, training dataset, data augmentation and other implementation details.

4.1 Loss Functions

Considering that there is a class imbalance problem, we restrict the number of positive pixels and negative pixels during training, making them numerically equal. This can be done by hard examples mining. We simply use softmax loss for the classification. The regression task is optimized by IoU loss, more details can be found in [21]. These two tasks are joint optimized equally. The multi-task loss is formulated as

$$L = L_{cls} + L_{IoU} \quad (2)$$

We empirical note that model optimized by Eq. 2 has a problem in locating tiny faces, leading to lots of false positives. We solve this problem by employing a focal loss to focus training on locating tiny face. The new loss function can be rewritten as

$$L = L_{cls} + \alpha S^{-\gamma} L_{IoU} \quad (3)$$

where S is the face area, α and γ are two constant. In our experiments, we empirically set $\alpha = 4, \gamma = 0.5$.

4.2 Training Dataset and Data Augmentation

We use the WIDER FACE training set which contains 12,880 images to train our model. In order to get better results, we also apply the following data augmentation techniques: (1) **Scale modification.** Each image is random scaling in a range between $[0.6, 2]$ via bilinear interpolation. (2) **Random crop.** We randomly crop a square patch from the image. And the size of the image patch is 640×640 . For images with shorter side less than 640 pixels, we firstly pad the images with 0, making their shorter side greater than 640. (3) **Horizontal flip.** After random crop, we obtain 640×640 image patch, and then we horizontally flip it with probability of 0.5.

4.3 Other Implementation Details

Online hard examples mining is employed to boost the performance of the model. For the parameter initialization, the parameters of the backbone are initialized from the corresponding pre-trained models. We use PVANet as the backbone in our experiments. Other additional layers are randomly initialized with the “xavier” method. All models are trained by SGD with a single GPU. The mini-batch sizes of models are 6, because of the GPU memory limitation. Weight decay is $1e-5$ and momentum is 0.9. Our networks are trained for 500 K iterations. The initial learning rate is 0.001 and drops by a factor of 5 after 200 K iterations. During inference, the score threshold is set to 0.01 and NMS with a threshold of 0.3 is performed on the predicted bounding boxes.

5 Experiments

5.1 Evaluation on Benchmark

We compare the proposed method with existing methods on two common face detection benchmarks: FDDB, WIDER FACE.

FDDB. It contains 2845 images with 5171 annotated faces. The Evaluation criteria include discrete score and continuous score. We compare our face detector against the state-of-the-art methods. Figure 4 shows the results. Our Face detector achieves competitive results with SFD [15] and outperforms other methods, indicating that our method can robustly detect unconstrained faces.

WIDER FACE. It contains 32203 images with a total of 393703 annotated faces with different scales, poses and occlusions. The data set is divided into training (40%), testing (50%) and validation (10%) set. Faces in the testing and validation set are split into three kinds of difficulty (easy, medium and hard). It is one of the most challenging face data sets. Our face detector is trained on WIDER FACE training set and tested on both validation and test set. We set the long side of the test image to 800, 1120, 1400, 1760 and 1920 for multi-scale

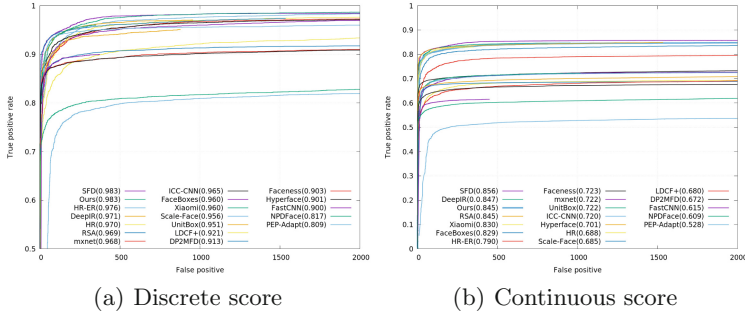


Fig. 4. Evaluation on FDDB

testing. Figure 5 illustrates the precision-recall curves along with AP scores. Our face detector outperforms other recent published methods including Zhu et al. [30], SFD [15], SSH [14] on the validation set and achieves competitive results with Zhu et al.’s [30], which demonstrate that the proposed method has a strong capacity in detecting small and hard faces.

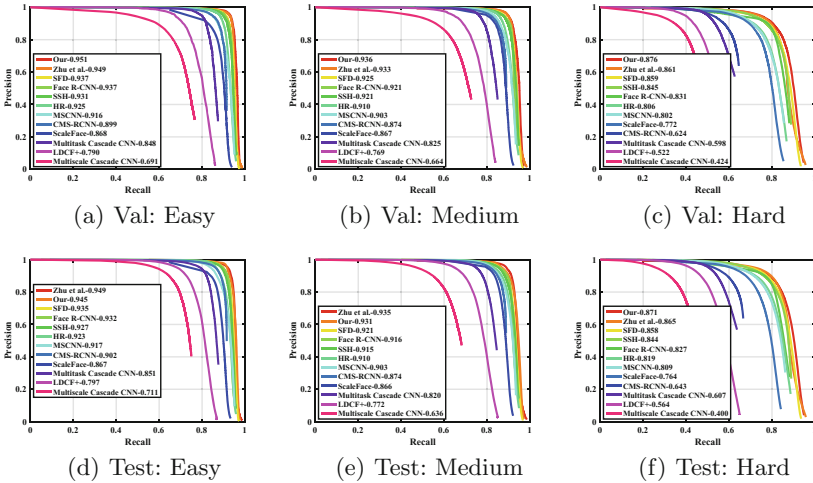


Fig. 5. Precision-recall curves on WIDER FACE validation and test sets.

5.2 Robustness to Occlusion

We further explore the ability of our detector in detecting occluded faces. To demonstrate the effectiveness of LSTM, we carry out comparative experiments with Two models: PVA, PVA+LSTM, where PVA uses PVANet [35] as the

backbone without Bi-LSTM architecture. Two occluded face data sets are used for this purpose, i.e. WIDER FACE validation set with artificial occlusion and MAFA with real occlusion. We also compare our method with other algorithms that release their trained models and testing codes such as MTCNN [18], SFD [15], SSH [14].

Faces with Artificial Occlusion. In this experiment, We generate a new occluded face data set by blacking a rectangle area on every faces of the WIDER FACE validation set. The rectangle black is randomly distributed in the left, right and bottom side of the face, accounting for 40% area of the face annotated box. Examples of occluded images are shown in Fig. 6. Table 1 shows the results of different methods. It’s clear that our two models outperform other methods. We note that adding LSTM or not makes little difference. The main reason is that the WIDER Face contain lots of tiny face, the role of encoding the context information of the RNN structure is weakened after adding the artificial occlusion.

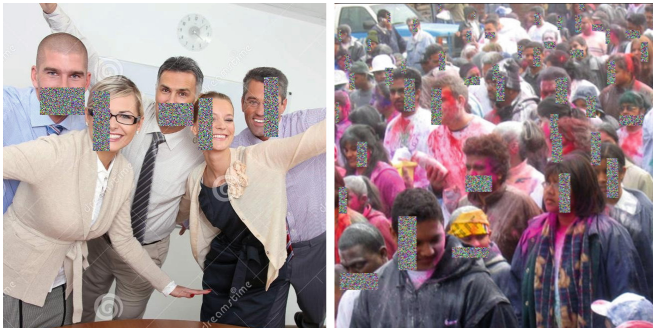


Fig. 6. Examples of WIDER FACE validation set with Occlusion

Table 1. Comparison of different models on the WDIER FACE validation set with artificial occlusion.

Methods	AP (easy)	AP (medium)	AP (hard)
MTCNN [18]	0.565	0.526	0.361
SSH [14]	0.801	0.768	0.625
SFD [15]	0.835	0.798	0.621
PVA	0.881	0.850	0.723
PVA+LSTM	0.881	0.851	0.720

Faces with Real Occlusion. MAFA data set contains 30,811 image with 35,806 faces collected from the Internet. Most of the faces are occluded by mask. We only use the testing set which contains 4,935 images to evaluate our face

detector. The long side of all testing images is set to 1280. Table 2 shows the results of different methods. Our base models without LSTM have already outperform other methods. And the LSTM structure further improves the robustness of our face detectors in detecting faces with real occlusion.

Table 2. Comparison of different models on the MAFA data set.

Methods	MTCNN [18]	SSH [14]	SFD [15]	LLE-CNNs [36]	PVA	PVA+LSTM
AP	0.570	0.643	0.724	0.764	0.768	0.781

5.3 Inference Time

Although our method achieves great performance, its speed is not compromised. We employ PVANet, a light-weighted neural network, as the backbone, which greatly reduces the computational burden. We measure the speed using a GTX 1080Ti GPU and Intel Xeon E5-2620 v4@2.1 GHz CPU. Table 3 shows the inference time and AP with respect to different input sizes of our face detector. The max size stands for the long side of the input image while keeping the aspect ratio.

Table 3. The inference time and AP with respect to different input sizes

Max size	800	1120	1440	1760	1920
AP (hard)	0.723	0.829	0.863	0.873	0.872
Time (ms)	60.7	83.9	124.9	172.9	195.0

6 Conclusions

In this paper, we propose a novel FCN-based face detector which is simple and efficient. Unlike other anchor-based methods, our face detector performs dense prediction on a single feature map, which is inherent robust in detecting occluded faces. By using the in-network RNN structure, our face detector is superior to handle the detection of occluded faces. Besides, the size of the final feature map is only 1/4 of the original image, reducing the computational cost while achieving remarkable results in detecting small faces. The experiments demonstrate that the proposed method achieves the state-of-the-art performance on the challenging face detection benchmarks, especially for small faces and occluded faces.

References

1. Viola, P., Jones, M.J.: Robust real-time face detection. *IJCV* **57**(2), 137–154 (2004)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, vol. (1), pp. 886–893. IEEE (2005)
3. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: 1994 Proceedings of the 12th IAPR International Conference on Pattern Recognition, Computer Vision and Image Processing, vol. 1, pp. 582–585. IEEE (1994)
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
5. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
6. Girshick, R.: Fast R-CNN. arXiv preprint [arXiv:1504.08083](https://arxiv.org/abs/1504.08083) (2015)
7. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015)
8. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
9. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
10. Zhu, C., Zheng, Y., Luu, K., Savvides, M.: CMS-RCNN: contextual multi-scale region-based CNN for unconstrained face detection. In: Bhanu, B., Kumar, A. (eds.) Deep Learning for Biometrics. ACVPR, pp. 57–79. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61657-5_3
11. Jiang, H., Learned-Miller, E.: Face detection with the faster R-CNN. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), pp. 650–657. IEEE (2017)
12. Wang, H., Li, Z., Ji, X., et al.: Face R-CNN. arXiv preprint [arXiv:1706.01061](https://arxiv.org/abs/1706.01061) (2017)
13. Sun, X., Wu, P., Hoi, S.C.H.: Face detection using deep learning: an improved faster RCNN approach. arXiv preprint [arXiv:1701.08289](https://arxiv.org/abs/1701.08289) (2017)
14. Najibi, M., Samangouei, P., Chellappa, R., et al.: SSH: single stage headless face detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4875–4884 (2017)
15. Zhang, S., Zhu, X., Lei, Z., et al.: S³FD: single shot scale-invariant face detector. arXiv preprint [arXiv:1708.05237](https://arxiv.org/abs/1708.05237) (2017)
16. Hu, P., Ramanan, D.: Finding tiny faces. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1522–1530 (2017)
17. Li, H., Lin, Z., Shen, X., et al.: A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334 (2015)
18. Zhang, K., Zhang, Z., Li, Z.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Sig. Process. Lett.* **23**(10), 1499–1503 (2016)
19. Qin, H., Yan, J., Li, X., et al.: Joint training of cascaded CNN for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3456–3465 (2016)

20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Yu, J., Jiang, Y., Wang, Z., et al.: UnitBox: an advanced object detection network. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 516–520. ACM (2016)
22. Yang, B., Yan, J., Lei, Z., et al.: Aggregate channel features for multi-view face detection. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8 (2014)
23. Zhu, Q., Yeh, M.C., Cheng, K.T., et al.: Fast human detection using a cascade of histograms of oriented gradients. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1491–1498 (2006)
24. Liao, S., Jain, A.K., Li, S.Z.: A fast and accurate unconstrained face detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(2), 211–223 (2016)
25. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., et al.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2010)
26. Brubaker, S.C., Wu, J., Sun, J., et al.: On the design of cascades of boosted ensembles for face detection. *Int. J. Comput. Vis.* **77**(1–3), 65–86 (2008)
27. Pham, M.T., Cham, T.J.: Fast training and selection of HAAR features using statistics in boosting-based face detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 1–7 (2007)
28. Yang, S., Luo, P., Loy, C.C., et al.: From facial parts responses to face detection: a deep learning approach. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3676–3684 (2015)
29. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
30. Zhu, C., Tao, R., Luu, K., et al.: Seeing small faces from robust anchor’s perspective. arXiv preprint [arXiv:1802.09058](https://arxiv.org/abs/1802.09058) (2018)
31. Zhou, X., Yao, C., Wen, H., et al.: EAST: an efficient and accurate scene text detector. arXiv preprint [arXiv:1704.03155](https://arxiv.org/abs/1704.03155) (2017)
32. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 56–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_4
33. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
34. Lin, T.Y., Dollr, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, no. 2, p. 4 (2017)
35. Kim, K.H., Hong, S., Roh, B., et al.: PVANET: deep but lightweight neural networks for real-time object detection. arXiv preprint [arXiv:1608.08021](https://arxiv.org/abs/1608.08021) (2016)
36. Ge, S., Li, J., Ye, Q., et al.: Detecting masked faces in the wild with LLE-CNNs. In: The IEEE Conference on Computer Vision and Pattern Recognition (2017)



Nuclear Norm Based Superposed Collaborative Representation Classifier for Robust Face Recognition

Yongbo Wu and Haifeng Hu(✉)

School of Electronics and Information Technology, Sun Yat-Sen University,
Guangzhou, China
wuyb6@mail2.sysu.edu.cn, huhaif@mail.sysu.edu.cn

Abstract. In this paper, we propose a novel robust face recognition framework named nuclear norm based superposed collaborative representation classifier (NNSCRC) to handle illumination variations, occlusion and undersampled problems in face recognition. Specifically, we develop a superposed linear collaborative representation classifier for robust face recognition by representing the query image in terms of a superposition of the class centroid, the shared intra-class difference, and the low rank error. By representing a face image as the class centroid and the shared intra-class difference, our model can effectively enhance the face recognition performance on undersampled databases. In addition, since the occlusion and illumination variations generally lead to a low-rank error image, we use nuclear norm matrix regression to obtain these low-rank errors, which makes our model able to reconstruct the test image better. Extensive experiments are performed on Extended Yale-B and AR databases, which show the effectiveness of NNSCRC in robust face recognition.

Keywords: Robust face recognition · Nuclear norm Superposed collaborative representation

1 Introduction

Face recognition (FR) has received extensive research during last thirty years and numerous FR methods have been developed [7, 8, 13, 15, 17, 24]. Classical FR algorithms including principal component analysis (PCA) [19], linear discriminant analysis (LDA) [3] and laplacianface [10] try to employ subspace learning method to represent the intrinsic characteristics of faces. At the same time, many types of image features like scale-invariant feature transform (SIFT) [16], local binary pattern (LBP) [1], speeded-up robust features (SURF) [2] and histogram of oriented gradient (HOG) [21] have been introduced into FR algorithms, while the final recognition result can be easily obtained based on these feature representations. However, these feature descriptors are hand-crafted and always

require many prior knowledge, which limits the improvement of recognition performance.

Regression analysis based methods have also aroused broad interests in face recognition community. For example, Naseem et al. proposed a linear regression classification (LRC) [15] by reconstructing a query image as the linear combination of dictionary faces. Wright et al. proposed a sparse representation based classification algorithm (SRC) [22] for robust face recognition using a sparse constraint. By representing a face image with a sparse linear combination of the dictionary faces, SRC believed that the query image will be reconstructed by the training samples in the same class. However, when the number of training samples is limited, sparsity between classes may lead to misleading solutions. Zhang et al. [25] analyzed the principle of SRC and believed that collaborative representation is more effective than sparsity constraint. Based on ridge regression, they introduced a collaboration representation classifier (CRC) which lead to better FR accuracy and lower complexity than SRC. After that, many improved versions of CRC algorithm have been proposed to further improve the performance of FR. For example, Wang et al. [20] used a relaxed collaborative representation (RCR) by considering locality constraints. Huang et al. [11] introduced group sparse classifier (GSC) which tries to incorporate the class labels to boost FR performance. IRGSC [26] further introduced group sparse classifier with adaptive weights learning, and had achieved good performance in robust face recognition.

Recently, Yang et al. [23] proposed nuclear norm based matrix regression (NMR) classification framework for occlusion face recognition and had achieved good recognition performance. However, NMR relies heavily on the completeness of database. When the number of training samples is limited, NMR suffers from misleading coding coefficients of incorrect classes. More recently, superposed linear representation based classification (SLRC) [9] model was proposed to further improve the robustness of CRC. SLRC decomposed the training sample of CRC into prototype and variation parts, and proposed a superposed linear representation that encodes the test sample as a superposition of the prototype and variation dictionaries. In SLRC, the author simply assumed that the test image can be reconstructed by class-central of corresponding class and the shared intra-class differences. However, when there are unknown illumination variations or occlusion in the test image, the SLRC model will not work effectively since it cannot reconstruct the image properly.

In order to address the limitations of NMR and SLRC, we propose a novel model called nuclear norm based superposed collaborative representation classifier (NNSCRC). In our model, a query image can be decomposed as a class centroid, a shared sample-to-centroid difference and a low rank error image. The main contributions of this paper are outlined as follows:

- We propose a new framework named nuclear norm based superposed collaborative representation classifier for robust face recognition where a test face image can be reconstructed as a superposed of class centroid, intra-class difference and low rank error. The new model can address the misleading coding coefficients of incorrect classes when the dataset is undersampled, since

it has decomposed the image as a class centroid and sample-to-centroid difference. Alternating direction method of multipliers (ADMM) algorithm has been used to obtain the optimal solution of proposed model.

- By introducing a nuclear norm constraint, the low-rank part, generally the occlusion or illumination variations in the image, will be separated out from the dictionary reconstruction. Thus, the NNSCRC model is robust to occlusion or illumination variations.
- NNSCRC model is robust to single sample per person (SSPP) face recognition problem. Specifically, when there is only one train image available in each class, we can borrow the intra-class variations from the subjects outside the gallery since these variations are usually similar across different subjects. The variations between query image and gallery images can be represented by these intra-class variations properly, which will improve the performance of SSPP face recognition.
- Experimental results on Extended Yale-B and AR databases show the proposed NNSCRC model achieves better performance than state-of-the-art regression based methods for illumination variations, occlusion and under-sampled face recognition.

The remainder of this paper is organized as follows: Sect. 2 reviews the related works. Section 3 introduces the proposed nuclear norm based superposed collaborative representation classifier (NNSCRC). In Sect. 4, we conduct experiments on two popular face databases and compare our model with the state-of-the-art regression based methods. Finally, Sect. 5 concludes this paper.

2 Related Works

In this section, we briefly review the regression based methods and introduce SLRC method in detail, which is related to our model.

Regression based methods have long been a research hotspot in face recognition community. Started by SRC, which represents a query image as a sparse reconstruction of dictionary images, many regression based approaches like CRC have been proposed in succession and have achieved good performance in face recognition task. Collaborative representation based methods believe that l_2 -norm constraint is more important than l_1 -norm constraint in classifier. They use training samples to reconstruct the test sample and believe the training samples in the same class will become the major components in the reconstruction process. Although these regression based methods have achieved good performance on general face recognition, their generalization ability to illumination variations, occlusion and undersampled face recognition problems is still weak.

Recently, superposed linear representation based classification (SLRC) [9] is proposed to decompose the collaborative dictionary in a manner similar to the decomposed representation in LDA. Specifically, given a sample x from one of the classes in the training set, SLRC assume it can be naturally reconstructed by two parts:

$$x = c_{(x)} + (x - c_{(x)}) \quad (1)$$

where $c_{(x)}$ is the centroid of corresponding class, and $x - c_{(x)}$ is the intra-class difference from the sample to its class centroid. SLRC has achieved promise performance when the test images have similar attributes to the training images. However, when there are unknown variations in the test image such as illumination changes or occlusion, the SLRC model will not work properly since it cannot reconstruct these variations in test image.

Considering these limitations, we propose a novel framework to incorporate the nuclear norm constraint into superposed linear representation based classification, which not only makes use of the general variation information of training samples, but also improves the robustness to unknown illumination changes and occlusions. The proposed model will be introduced in detail in the next section.

3 Nuclear Norm Based Superposed Collaborative Representation Classifier (NNSCRC)

Although CRC methods have received great success in face recognition, it still suffers from undersampled and occlusion problems. Firstly, when the training images are insufficient or unrepresentative, the test sample has to be reconstructed by the samples of other classes, which usually generates misleading coding coefficients. Secondly, when there are illumination changes or occlusion in the test images, the reconstructed error will be dominated by these noise, which will also lead to erroneous results. In order to overcome these difficulties, we propose a novel robust face recognition framework called nuclear norm based superposed collaborative representation classifier (NNSCRC). We will introduce our NNSCRC model in detail and provide the optimization algorithm of NNSCRC in this section.

3.1 NNSCRC Model

Inspired by NMR [23] and SLRC [9], we represent a test image as a superposition of three parts, i.e., the class centres, the shared intra-class differences, and the

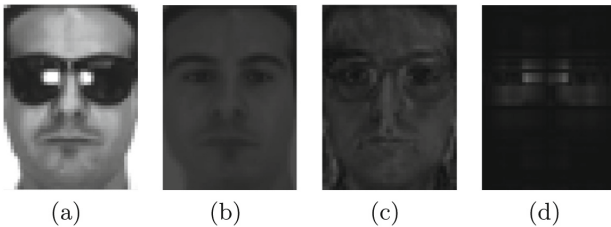


Fig. 1. In the proposed NNSCRC model, we try to reconstruct a test image as a linear superposition of the class centroid, the shared intra-class differences, and the low-rank error. (a) the original test image (b) the class centroid image (c) the shared intra-class differences image (shown in absolute value) (d) the low-rank error image (shown in absolute value)

low-rank error, as shown in Fig. 1. Specifically, given a test image \mathbf{Y} , we assume it can be reconstructed by the mentioned three parts, which can be formulated as:

$$\mathbf{Y} = \mathcal{P}(\boldsymbol{\alpha}) + \mathcal{V}(\boldsymbol{\beta}) + \mathbf{B}. \quad (2)$$

where $\mathcal{P}(\boldsymbol{\alpha}) = \alpha_1 \mathbf{P}_1 + \alpha_2 \mathbf{P}_2 + \dots + \alpha_n \mathbf{P}_n$, $\mathcal{V}(\boldsymbol{\beta}) = \beta_1 \mathbf{V}_1 + \beta_2 \mathbf{V}_2 + \dots + \beta_n \mathbf{V}_n$, and \mathbf{P}_i is the central of class i , \mathbf{V}_i is the variation dictionary of class i . α_i, β_i are the corresponding reconstruction coefficients of class i . \mathbf{B} is the low rank error image. To obtain the optimal reconstruction coefficients $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, we can naturally construct the objective function as:

$$\begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \arg \min \|\mathbf{y} - [\mathbf{P}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{b}\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 + \lambda_2 \|\mathbf{B}\|_*, \quad (3)$$

where $\mathbf{P} \in \mathbb{R}^{d \times k}$ is the prototype dictionary and $\mathbf{V} \in \mathbb{R}^{d \times n}$ is the variation dictionary, d is the dimension of face image, k represents the class number and n is the number of training images. $\|\mathbf{B}\|_*$ represents the nuclear norm of low rank error \mathbf{B} , and \mathbf{b} is the vectorization of matrix \mathbf{B} . $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are the coefficient vectors to be determined. λ_1, λ_2 are the penalty parameters. The prototype dictionary \mathbf{P} consists of centroid from all classes, and the variation dictionary \mathbf{V} consists of intra-class difference from the sample to its class centroid. The construction of dictionaries \mathbf{P} and \mathbf{V} is similar to [9]. For most collaborative representation based methods, undersampled training images usually lead to misleading coding coefficients. The main reason is that when the training images is insufficient, the difference between test image and corresponding prototype class need to be make up by images from other class, which make the major components of reconstruction might be found in the error class. By integrating superposed linear representation classifier with nuclear norm, our model can address the problem of misleading coefficients and enhance the robustness to illumination changes and occlusion. The reasons are listed as follows:

Firstly, we introduce a superposed linear representation into our model, which constructs a prototype dictionary \mathbf{P} and a variation dictionary \mathbf{V} . When the dataset is undersampled, the shared variation dictionary \mathbf{V} will make up the difference between the test image and the corresponding prototype class. The major components of reconstructed test image will be the class centroid of corresponding class, the intra-class variations from all classes, and the low rank error, which makes our model can handle the misleading coefficients problem.

Secondly, since occlusion and illumination changes generally lead to a low-rank error image, we apply a nuclear norm constrained matrix to characterize this structured noise (see Fig. 1(d)). When there are unknow occlusion or illumination changes in the test image, the nuclear norm constrained error term will represents this kind of noise properly, which makes the NNSCRC model can work effectively.

3.2 Algorithm of NNSCRC

We provide the theoretical solution of NNSCRC in this section. Since Eq. (3) is not always a convex function, we cannot solve it with traditional methods like

augmented Lagrange Multipliers (ALM). Notice that it satisfies the condition of Alternating Direction Method of Multipliers (ADMM) [4], which will be proved in Sect. 3.3, we use ADMM algorithm to solve the optimization problem. Specifically, we first introduce a matrix variable \mathbf{C} and rewrite Eq. (3), which form the object function as:

$$J(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}, \mathbf{C}) = \min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}, \mathbf{C}} \|\mathbf{y} - [\mathbf{P}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{b}\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 + \lambda_2 \|\mathbf{C}\|_*, \quad s.t. \mathbf{C} - \mathbf{B} = \mathbf{0}. \quad (4)$$

Denote

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}) = \|\mathbf{y} - [\mathbf{P}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{b}\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2. \quad (5)$$

Then the Lagrange form of $J(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}, \mathbf{C})$ is

$$\begin{aligned} L_\rho(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}, \mathbf{C}) &= f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}) + \lambda_2 \|\mathbf{C}\|_* + tr(\mathbf{Z}^T(\mathbf{C} - \mathbf{B})) + \frac{\rho}{2} \|\mathbf{C} - \mathbf{B}\|_F^2 \\ &= f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}) + \lambda_2 \|\mathbf{C}\|_* + \frac{\rho}{2} \|\mathbf{C} - \mathbf{B}\|_F^2 + \frac{1}{\rho} \|\mathbf{Z}\|_F^2 - \frac{1}{2\rho} \|\mathbf{Z}\|_F^2. \end{aligned} \quad (6)$$

where $\rho > 0$ is the Lagrangian multiplier, and \mathbf{Z} is the dual variable. The obtain of the optimal solution contains the following three iterative processes.

Fix $\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}$ to Solve \mathbf{C} . At k -th iterative, when $\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}$ is fixed, Eq. (6) can be rewritten as

$$J_1(\mathbf{C}) = \arg \min_{\mathbf{C}} \lambda_2 \|\mathbf{C}\|_* + \frac{\rho}{2} \|\mathbf{C} - \mathbf{B}_k + \frac{1}{\rho} \mathbf{Z}_k\|_F^2. \quad (7)$$

Let $\mathbf{Q} = \mathbf{B}_k - \frac{1}{\rho} \mathbf{Z}_k \in \mathbb{R}^{m_1 \times m_2}$, where $rank(\mathbf{Q}) = r$. We apply singular value decomposition to \mathbf{Q} as:

$$\mathbf{Q} = \mathbf{U}_{m_1 \times r} \boldsymbol{\Sigma}_{m_2 \times r}^T, \quad (8)$$

where $\boldsymbol{\Sigma} = diag(\sigma_1, \sigma_2, \dots, \sigma_r)$ and $\sigma_1, \sigma_2, \dots, \sigma_r$ are positive singular values. $\mathbf{U}_{m_1 \times r}$ and $\mathbf{V}_{m_2 \times r}$ are corresponding matrices with orthogonal columns. According to [5], the iterative solution of \mathbf{C}_{k+1} can be expressed as

$$\mathbf{C}_{k+1} = \mathbf{U}_{m_1 \times r} (\{max(0, \sigma_j - \frac{\lambda_2}{\rho})\}_{1 \leq j \leq r}) \mathbf{V}_{m_2 \times r}^T. \quad (9)$$

Fix \mathbf{Z}, \mathbf{C} to Solve $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and \mathbf{B} . At k -th iterative, when \mathbf{Z}, \mathbf{C} is fixed, Eq. (6) can be rewritten as

$$\begin{aligned} J_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}) &= \min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}) + \frac{\rho}{2} \|\mathbf{C}_{k+1} - \mathbf{B} + \frac{1}{\rho} \mathbf{Z}_k\|_F^2 \\ &= \min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}} \|\mathbf{y} - [\mathbf{P}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} - \mathbf{b}\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 \\ &\quad + \frac{\rho}{2} \|\mathbf{C}_{k+1} - \mathbf{B} + \frac{1}{\rho} \mathbf{Z}_k\|_F^2. \end{aligned} \quad (10)$$

Define $\mathbf{H}_k = \mathbf{C}_{k+1} + \frac{1}{\rho}\mathbf{Z}_k \in \mathbb{R}^{m_1 \times m_2}$, $\mathbf{h}_k = \text{Vec}\{\mathbf{H}_k\} \in \mathbb{R}^{m_1 m_2 \times 1}$, the optimal solution can be obtained by setting the derivative of $J_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b})$ with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and \mathbf{b} to zero respectively. Therefore, we have the optimal solution of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and \mathbf{B} at k -th iterative as

$$\boldsymbol{\alpha}_{k+1} = (\mathbf{P}^T \mathbf{P} + 2\lambda_1 \mathbf{I})^{-1} \mathbf{P}^T (\mathbf{y} - \mathbf{b}_{k+1} - \mathbf{V} \boldsymbol{\beta}_k), \tag{11}$$

$$\boldsymbol{\beta}_{k+1} = (\mathbf{V}^T \mathbf{V} + 2\lambda_1 \mathbf{I})^{-1} \mathbf{V}^T (\mathbf{y} - \mathbf{b}_{k+1} - \mathbf{P} \boldsymbol{\alpha}_{k+1}), \tag{12}$$

$$\mathbf{b}_{k+1} = \frac{1}{2 + \rho} (2\mathbf{y} - 2\mathbf{P} \boldsymbol{\alpha} - 2\mathbf{V} \boldsymbol{\beta} + \rho \mathbf{h}_k). \tag{13}$$

Fix $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{C}$ and \mathbf{B} to Solve \mathbf{Z} . According to [4], the optimal solution of \mathbf{Z} at iteration k can be directly obtained by

$$\mathbf{Z}_{k+1} = \mathbf{Z}_k + \rho(\mathbf{C}_{k+1} - \mathbf{B}_{k+1}). \tag{14}$$

With the iteration optimal solution in Sect. 3.2, we can finally obtain the optimal solution of $J(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}, \mathbf{C})$ by alternate iteration. Finally, the optimal reconstruction coefficients are:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_{k+1}, \quad \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_{k+1}. \tag{15}$$

3.3 Classification Strategy of NNSCRC

Given test image Y , we need to decide which class it belongs to for face recognition task. By using NNSCRC algorithm, we can obtain the reconstruction coefficients $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. We use the reconstruction residual in each class as the criterion for classification. Specifically, the residual of test image Y is

$$r_i(\mathbf{Y}) = \|\mathbf{Y} - [\mathbf{P}, \mathbf{V}] \begin{bmatrix} \delta_i(\hat{\boldsymbol{\alpha}}) \\ \hat{\boldsymbol{\beta}} \end{bmatrix} - \mathbf{B}\|_2, \quad i = 1, \dots, k. \tag{16}$$

Where $\delta_i(\hat{\boldsymbol{\alpha}}) \in \mathbb{R}^n$ is a new vector whose only nonzero entries are the entries in $\hat{\boldsymbol{\alpha}}$ that are associated with class i . Note that when we calculate the residual, we use intra-class variation matrix of all classes to reconstruct the test image Y , because these intra-class variation are often shareable across different subjects. This is also one of the reason that our model is suitable for SSPP task. From Eq. (16), we can find that the normal variations and error image are separated out from the original query image, which can remove the influence of illumination changes and occlusions. Based on the reconstruction residual, we can decide the class label by

$$\text{class}(\mathbf{Y}) = \arg \min_i r_i(\mathbf{Y}). \tag{17}$$

4 Experiments

In this section, we perform extensive experiments on two publicly available face datasets to demonstrate the effectiveness of NNSCRC. Section 4.1 first gives the

experimental settings of our experiments. In Sect. 4.2, we evaluate NNSCRC for FR with different training sizes under controlled conditions. Section 4.3 verifies the robustness of NNSCRC to illumination changes and occlusion face recognition. Section 4.4 compares our method with existing methods for face recognition task under real face disguise. Finally, in Sect. 4.5, face recognition experiment with single sample per person has been performed.

4.1 Experimental Settings

We apply Aleix Martinez and Robert Benavente (AR) dataset [14] and the Extended Yale B (ExYaleB) dataset [12] to test the effectiveness and robustness of proposed model. The AR dataset contains over 4000 images of 126 individuals (70 men and 56 women). The faces in AR dataset contain variations such as lighting conditions, expressions and occlusions. Some examples of face images in AR database are shown in Fig. 2. For this dataset, we randomly select 100 subjects (50 men and 50 women) for our experiments. The Extended Yale B face dataset contains 38 human subjects under 9 poses and 64 illumination conditions. The 64 samples of each subject are acquired in a particular pose, which are all frontal view facial images. Figure 3 shows some facial images in ExYaleB database. All face images marked with P00 are used in our experiments.



Fig. 2. Facial image samples in AR database



Fig. 3. Facial image samples in the Extended Yale B face database

The proposed model is compared to state-of-the-art regression based representation methods including NMR [23], WGSC [18], RCRC [6], RSRC [22], and IRGSC [26]. For NNSCRC, the Lagrangian multiplier ρ is set to 1, and the parameter λ_1 , λ_2 are both traversed in $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ to obtain best result. For all the comparative methods, the related parameters are set to the values suggested by the authors.

4.2 Face Recognition with Different Sample Sizes

We first validate the performance of NNSCRC without occlusion on ExYaleB database. In order to explore the effect of sample size on experimental results, we randomly split the dataset into two parts. One part is used as the dictionary, which contains $n(=10, 20, 30, 40, 50)$ images for each person, and the other part is used for testing. The results are shown in Fig. 4, which compares our method with the state-of-the-art method, IRGSC. Two most classical regression based face recognition methods including CRC and SRC have also been used for comparison.

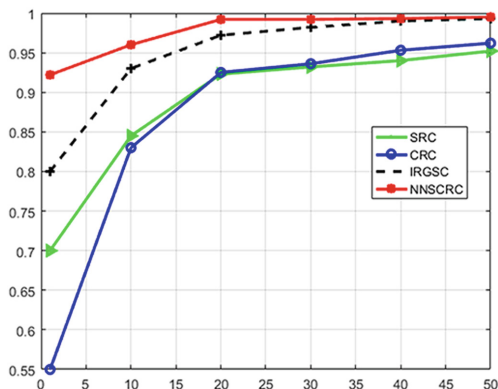


Fig. 4. Face recognition with different sample sizes on ExYaleB database

From Fig. 4, we can find that the performances of all methods improved when the sample size increases. Though the test faces suffers from illumination problems, for all groups of sample size, our NNSCRC model outperforms SRC and CRC for over five percentage, which shows our model is more robust to illumination variations compared to original collaborative representation methods. IRGSC achieves higher accuracy than SRC and CRC because it use the reconstruction residuals to obtain the feature weights, which can reduce the influence of the pixel errors. However, there are still some variations between train images and test images which will influence the reconstruction and classification, and these variations cannot easily removed by the adaptive weights in IRGSC. In comparison, our model still achieves higher accuracy than IRGSC for all groups of sample sizes. The main reason is that our model can reconstruct the variations by using the variation dictionary which is constructed by all classes. The nuclear norm constraint can also handle the illumination variations problem, which make NNSCRC achieve better performance compared to IRGSC.

4.3 Face Recognition with Occlusion

To validate the robustness of proposed NNSCRC model to occlusion, we conduct two types of experiments on ExYaleB dataset, including random block occlusion experiment and random face occlusion experiment.

Random Block Occlusion. We select 20 samples per subject in ExYaleB dataset for training, and 20 for testing. Similar to the work in IRGSC, for each test image, we randomly select a location in the image and replace 10–60% pixels using a black block. Figure 5 shows the examples of different percentage of occlusions. The recognition rates of different methods are shown in Table 1. From Table 1, we can see that for all group of block occlusion, our method achieve the best performance compared with state-of-the-art regression based methods. Note that for 60% occlusion, our method still achieves 80.3% recognition rate, which is 7.7% higher than IRGSC. NMR has worse performance compared to IRGSC because it simply ignores the general variations, which will also influence the reconstruction error. By considering the general variations and the low-rank error, the proposed model can achieve better performance than other methods.

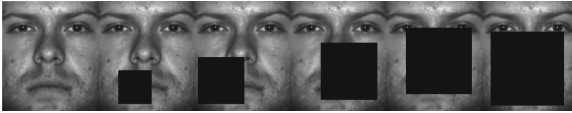


Fig. 5. Samples with different percentage of pixel corruption (0%–60%)

Table 1. Recognition accuracy of different methods versus different percentage of block occlusion

Occlusion (%)	10	20	30	40	50	60
RSRC	98.6	96.2	95.2	93.5	69.7	56.4
RCRC	99.0	97.9	96.7	94.3	81.2	62.0
WGSC	94.1	93.4	85.3	73.9	57.1	41.3
NMR	99.0	98.0	95.9	92.5	81.1	69.3
IRGSC	99.1	98.2	96.7	94.2	83.8	72.6
NNSCRC	99.4	98.4	96.7	94.7	87.5	80.3

Random Face Occlusion. In this experiment, we replace 10–50% pixels of each test images with other face images. As shown in Fig. 6, both the location of occlusion position and the occlusion face images are randomly selected. Table 2 lists the recognition accuracy of different methods. As can be seen, our method still achieve better performance compared to others methods. The recognition

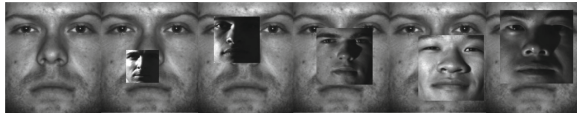


Fig. 6. Samples with different percentage of face occlusion (0%–50%)

Table 2. Recognition accuracy of different methods versus different percentage of face occlusion

Occlusion (%)	10	20	30	40	50
RSRC	96.9	95.6	91.2	88.6	72.9
WGSC	97.6	96.4	90.2	84.0	67.8
NMR	98.9	95.3	93.8	83.1	72.1
IRGSC	99.1	96.4	94.1	89.2	81.7
NNSCRC	99.3	97.2	96.0	91.4	83.2

rate of our model is a little lower than that of random block occlusion, which is due to the reason that face occlusion is not strictly low rank. Still, our model outperforms about 2% than IRGSC under large percentage face occlusion, which indicates the effectiveness of NNSCRC to address occlusions.

4.4 Face Recognition with Real Disguise

To evaluate the robustness of our model to real possible disguise, we further conduct experiments on AR dataset. As shown in Fig. 2, there are some samples with sunglasses or scarves in AR database, which reflects the real FR conditions in practical application. This kind of occlusion is irregular, thus brings a large challenge for FR tasks. In our experiment, the face images of these 100 persons were separated into 2 sessions according to the shooting time of photos. For each person, we select 3 images in session 1 which has no illumination changes or occlusion problem as training samples. 1200 face images are used for test, which are divided into 4 groups as: 300 face images with illumination changes and sunglasses in session 1, and 300 face images with illumination changes and scarves in session 1, and the same divided in session 2.

The experiment results of competing methods are listed in Table 3. Clearly, the NNSCRC method achieves better result in all 4 groups of experiments compared with WGSC, RCRC, RSRC, and NMR. WGSC has the worst performance, while WGSC tried to regress the query images only with the training samples, and failed to consider the influence caused by occlusion. RCRC tries to solve the problem of occlusion, and in fact achieves better performance than WGSC. Note that our model outperform NMR by around 14%, which indicates that by introducing a superposed linear collaborative representation to NMR model, our model can enhance the robustness of face recognition effectively.

Table 3. Recognition rates (%) of different methods on AR database

Classifier	Session 1		Session 2	
	Sunglasses	Scarves	Sunglasses	Scarves
WGSC	66.3	62.7	32.0	36.3
RSRC	89.3	32.3	57.3	12.7
RCRC	80.3	70.3	46.3	42.0
NMR	72.3	72.3	35.3	45.3
NNSCRC	90.0	79.7	59.7	50.7

4.5 Face Recognition with Single Sample per Person

We further conduct experiments on ExYaleB dataset to evaluate the robustness of our model to single sample per person (SSPP) face recognition. 20 persons in ExYaleB are used for SSPP test and the other persons are used to construct intra-class variations. We use the first image of these 20 persons in ExYaleB dataset as gallery, and select 30 images each person as probe set. The results are shown in Table 4. As can be seen, the recognition rate of NNSCRC is 9.9% and 3.9% higher than that of NMR and IRGSC respectively. Though NMR and IRGSC can handle the problem of differences between query and gallery images in some kind, both of them suffers from the misleading coding coefficients of incorrect classes when there is only one sample per subject. Different from these methods, our model can borrow the intra-class variations from other subjects which are not in the gallery set because these variations are usually similar across different subjects. Clearly, the NNSCRC method achieves much better result than NMR and IRGSC since NNSCRC can borrow the intra-class variations from other subjects, which demonstrate our model is capable for SSPP face recognition task.

Table 4. SSPP FR accuracy of different methods on ExYaleB database

	NMR	IRGSC	NNSCRC
Accuracy	79.3	85.3	89.2

5 Conclusion

In this paper, we present a NNSCRC model for robust face recognition task. In the proposed framework, a superposed collaborative representation is adopted to obtain robust representation of reconstruct face images. By representing a face image as a superposed of a class centroid, a shared sample-to-centroid difference and a low rank error, our method can address the misleading coding coefficients of incorrect classes when the dataset is undersampled. Specially, when there

is only a single sample per class available, the proposed model can still have promised performance by acquiring the intra-class variation base from the generic subjects outside the gallery. Furthermore, our model is robust to occlusion and illumination changes by introducing nuclear norm constrained. Experiments on the famous Extended Yale-B and AR databases show the superiority of our model compared with the state-of-the-art regression based face recognition methods.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (61673402, 61273270, 60802069), the Natural Science Foundation of Guangdong Province (2017A030311029, 2016B010123005, 2017B090909005), the Science and Technology Program of Guangzhou of China (201704020180, 201604020024) and the Fundamental Research Funds for the Central Universities of China.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(12), 2037–2041 (2006)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **110**(3), 346–359 (2008)
3. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. In: Buxton, B., Cipolla, R. (eds.) *ECCV 1996*. LNCS, vol. 1064, pp. 43–58. Springer, Heidelberg (1996). <https://doi.org/10.1007/BFb0015522>
4. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**(1), 1–122 (2010)
5. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20**(4), 1956–1982 (2008)
6. Cai, S., Zhang, L., Zuo, W., Feng, X.: A probabilistic collaborative representation based approach for pattern classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2950–2959 (2016)
7. Chien, J.T., Wu, C.C.: Discriminant waveletfaces and nearest feature classifiers for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(12), 1644–1649 (2003)
8. Choi, S.I., Lee, S.S., Sang, T.C., Shin, W.Y.: Face recognition using composite features based on discriminant analysis. *IEEE Access* **6**, 13663–13670 (2018)
9. Deng, W., Hu, J., Guo, J.: Face recognition via collaborative representation: its discriminant nature and superposed representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 1 (2017)
10. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using Laplacian-faces. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 328–340 (2005)
11. Huang, J., Nie, F., Huang, H., Ding, C.: Supervised and projected sparse coding for image classification. In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 438–444 (2013)
12. Lee, K.C., Ho, J., Kriegman, D.J.: Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(5), 684–698 (2005)

13. Lu, J., Liong, V.E., Zhou, X., Zhou, J.: Learning compact binary face descriptor for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(10), 2041–2056 (2015)
14. Martinez, A.M.: The AR face database. CVC Technical report 24 (1998)
15. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(11), 2106–2112 (2010)
16. Ng, P.C., Henikoff, S.: Sift: predicting amino acid changes that affect protein function. *Nucl. Acids Res.* **31**(13), 3812–3814 (2003)
17. Pentland, A., Moghaddam, B., Starner, T.: View-based and modular eigenspaces for face recognition. In: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 84–91 (1994)
18. Tang, X., Feng, G., Cai, J.: Weighted group sparse representation for undersampled face recognition. *Neurocomputing* **145**(18), 402–415 (2014)
19. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 586–591 (1991)
20. Wang, S.: Relaxed collaborative representation for pattern classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2224–2231 (2012)
21. Wang, X.: An HOG-LBP human detector with partial occlusion handling. In: Proceedings of IEEE International Conference on Computer Vision, Kyoto, Japan, September, vol. 30, no. 2, pp. 32–39 (2009)
22. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(2), 210–227 (2008)
23. Yang, J., Luo, L., Qian, J., Tai, Y., Zhang, F., Xu, Y.: Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(1), 156–171 (2016)
24. Yang, M., Zhang, L., Yang, J., Zhang, D.: Regularized robust coding for face recognition. *IEEE Trans. Image Process.* **22**(5), 1753–1766 (2013)
25. Zhang, L., Yang, M.: Sparse representation or collaborative representation: which helps face recognition? In: International Conference on Computer Vision, pp. 471–478 (2011)
26. Zheng, J., Yang, P., Chen, S., Shen, G., Wang, W.: Iterative re-constrained group sparse face recognition with adaptive weights learning. *IEEE Trans. Image Process.* **26**(5), 2408–2423 (2017)



Face Image Set Recognition Based on Bilinear Regression

Wen-Wen Hua¹ and Chuan-Xian Ren^{1,2}(✉)

¹ School of Mathematics, Sun Yat-sen University, Guangzhou 510275, China
huaww@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn

² Shenzhen Research Institute of Sun Yat-sen University, Shenzhen 518000, China

Abstract. Image sets-based face recognition receives growing research interest in pattern recognition and machine learning. The most challenging problem focuses on how to formulate a computable and discriminative model by using given data sets. In this paper, we propose a new method, which is called Bilinear Regression Classifier (BLRC) for short, to address the image sets-based face recognition problem. BLRC classifies a given test set by choosing the category that simultaneously maximizes the unrelated subspace and minimize the related subspace. In particular, the unrelated subspace is used to characterize the distances between the query set and the unrelated image sets, while the related subspace is used to characterize the distances between the query set and the related sets. In our work, the Mahalanobis metric, rather than the Euclidean metric, is exploited to compute the subspace distance. The subspace coefficient vectors are obtained by solving an Elastic-Net regularized regression model. Extensive experiments are conducted on several benchmark datasets to evaluate the real recognition performance of the new method. The results show that our BLRC method obtains competitive accuracies with some state-of-the-art methods.

Keywords: Face recognition · Image sets · Linear regression

1 Introduction

Face recognition has traditionally been posed as the problem of identifying a face from a single image. Good performance is usually rely on smartly designed classifiers. A number of classifiers were proposed, such as the Nearest Neighbor (NN) [4], A Local Support Vector Machine Approach [12], Sparse Representation-based Classifier [16] and Linear Regression Classification (LRC) [11]. These classifiers use a single test sample for classification and assume that images are taken

C.-X. Ren—This work is supported in part by the Science and Technology Program of Shenzhen under Grant JCYJ20170818155415617, the National Natural Science Foundation of China under Grants 61572536, and the Science and Technology Program of GuangZhou under Grant 201804010248.

in controlled environments. Their classification performance is generally dependent on the representation of individual test samples. However, facial appearance changes dramatically under variations in pose, illumination, expression, etc., and images captured under controlled conditions may not suffice for reliable recognition under the more varied conditions, that occur in real surveillance and video retrieval applications. Recently there has been growing interest in face recognition from image sets. Rather than supplying a single query image, the system supplies a set of images of the same unknown individual, and we expect that rich information provided in the image sets can improve the recognition rate.

Image sets classification algorithms include parametric methods [1, 8, 14] and non-parametric methods [2, 3, 5–7, 9, 13, 17]. Parametric method, firstly use the probability density functions to represent the image sets, then they use distance of divergence functions to measure the similarity between the image set (probability distribution), and they finally classify the test image set into the category which the closest image collection belongs. There are various difficulties in parametric methods, and the recognition performance is usually unsatisfactory. In recent years, researchers have focused on nonparametric methods that are independent of models. These methods do not have any assumptions about the distribution of image sets. Typical example of such methods is subspace algorithm.

This paper makes a brief review on dual linear regression classification (DLRC), then proposes the bilinear regression classification (BLRC) for image set retrieval. For BLRC algorithm, we first give the concept of uncorrelated subspace. Then, we introduce two strategies to constitute the unrelated subspace. Next, we calculate related distance metric and unrelated distance metric. Last, we introduce a combination metric for two new classifiers based on two constitution strategies of the unrelated subspace. Experimental results shows that the performance of BLRC is better than DLRC and several state-of-the-art classifiers for some benchmark.

2 Dual Linear Regression Classification

Suppose a and b be height and width of an image. Let two sets of (down-scaled) face images be represented by

$$X = [x_1, x_2, \dots, x_m], \quad (1)$$

$$Y = [y_1, y_2, \dots, y_n], \quad (2)$$

where x_i ($i = 1, 2, \dots, m$) and y_j ($j = 1, 2, \dots, n$) are column vectors of size ab .

Column vectors of the image set X and the image set Y determine a subspace respectively, and an image located at the intersection of the two subspaces. That is, the “virtual” face image can be assumed vector V should be a linear combination of the column vectors of two image sets respectively. To calculate the distance between two image sets, our task is to find the “virtual” face V and Coefficient vectors $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)^T$, $\beta = (\beta_1, \beta_2, \dots, \beta_n)^T$ such that

$$V = X\alpha = Y\beta. \quad (3)$$

Considering that we have all down-scaled images standardized into unit vectors, we further require that

$$\sum_{i=1}^m \alpha_i = \sum_{j=1}^n \beta_j = 1. \tag{4}$$

When $\hat{x}_i = x_i - x_m$ ($i = 1, 2, \dots, m - 1$), $\hat{y}_j = y_j - y_n$ ($j = 1, 2, \dots, n - 1$). We have

$$V = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{m-1}] \hat{\alpha} + x_m = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{n-1}] \hat{\beta} + y_n, \tag{5}$$

where $\hat{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_{m-1})^T$, $\hat{\beta} = (\beta_1, \beta_2, \dots, \beta_{n-1})^T$. Assume that there is an approximate solution $\gamma = (\alpha_1, \alpha_2, \dots, \alpha_{m-1}, \beta_1, \beta_2, \dots, \beta_{n-1})^T \in \mathbb{R}^{(m+n-2) \times 1}$ for the equation

$$y_n - x_m = \hat{X}Y\gamma, \tag{6}$$

where $\hat{X}Y = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{m-1}, -\hat{y}_1, -\hat{y}_2, \dots, -\hat{y}_{n-1}]$.

After obtaining the estimated value of the regression coefficient γ , the “virtual” face image may be represented by the image set X and the image set Y respectively. Specifically, the “virtual” face image V_X reconstructed from the image set X is

$$V_X = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{m-1}] [\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_{m-1}]^T + x_m, \tag{7}$$

while the “virtual” face image V_Y reconstructed from the image set Y is

$$V_Y = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{n-1}] [\hat{\gamma}_m, \hat{\gamma}_{m+1}, \dots, \hat{\gamma}_{m+n-2}]^T + y_n. \tag{8}$$

Obviously, difference between the two reconstructed “virtual” face images is essentially the residual of the linear regression equation. Since the difference between the image set X and the image set Y can be expressed by calculating the difference between the two reconstructed “virtual” face images, we can use the residual of the linear regression equation to estimate the similarity of the two image sets subspace X, Y , namely

$$D(X, Y) = \|V_Y - V_X\| = \|(y_n - x_m) - \hat{X}Y\hat{\gamma}\|. \tag{9}$$

If the $D(X, Y)$ value is smaller, the two image sets are closer to each other.

3 Bilinear Regression Classification

Inspired by DLRC, this section proposes bilinear regression classification. We show a simple flowchart in Fig. 1. The main contents of this section are organized as follows. First, the concept of unrelated subspaces is presented in Subsect. 3.1. Second, two strategies of constituting the unrelated subspace are described in Subsect. 3.2. Then, both related metric and unrelated metrics are computed in Subsect. 3.3. Last, the final distance metric for classification called combination metric, is described in Subsect. 3.4.

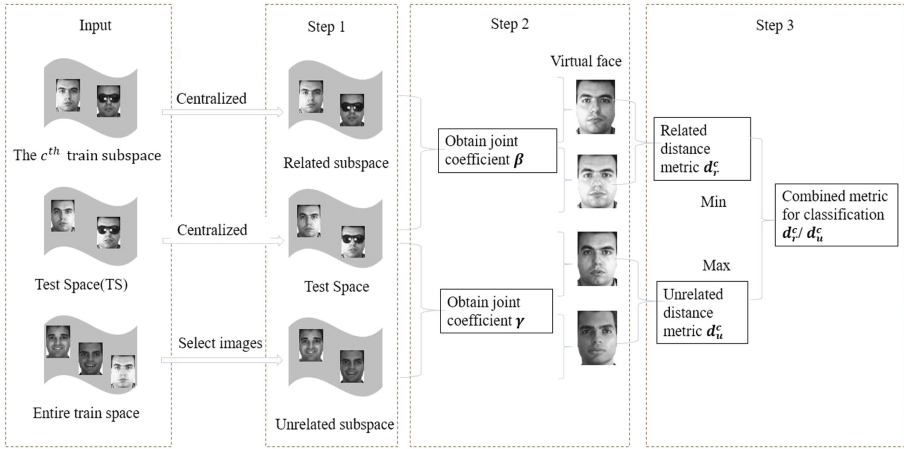


Fig. 1. The flowchart of the proposed BLRC

3.1 Definition of Unrelated Image Set Subspace

Definition 1. Suppose that there are C -classes image set in the training set, there are a total of M test image sets in the test set. For each image set in the test set, it is assumed that we need to calculate the distance between the test image set and the c^{th} image set, where $c = 1, 2 \dots C$, and the c^{th} image set in the training image set has N_c image samples. If there is a set U , U also contains N_c samples, and these N_c samples are from the other $C - 1$ classes except for the c^{th} class, then set U is called the unrelated image set subspace of the above test image set.

According to Definition 1, we need to select N_c image samples from the remaining $C - 1$ class samples that exclude c^{th} category to construct the unrelated image set subspace. In next subsection we will describe how to construct unrelated image set subspace.

3.2 Constructions of the Unrelated Subspace

The c^{th} image set X^c in the training image set is represented as follows:

$$X^c = [x_1^c, x_2^c, \dots, x_{N_c}^c] \in \mathbb{R}^{q \times N_c}. \tag{10}$$

That means that the c^{th} image set in the training set defines a subspace, which can be represented by X^c .

The subspace X determined by all images on the training set is as follows:

$$X = [X^1, X^2, \dots, X^C] \in \mathbb{R}^{q \times l}, \tag{11}$$

in which $l = \sum_{c=1}^C N_c$.

The overall mean of training image set X is

$$X_{mean} = \frac{1}{l} \sum_{c=1}^C \sum_{i=1}^{N_c} x_i^c. \quad (12)$$

The mean of the c^{th} image set on training image sets is $X_{mean}^c = \frac{1}{N_c} \sum_{i=1}^{N_c} x_i^c$. Images in class c are centralized as $\hat{x}_i^c = x_i^c - X_{mean}^c$ ($c = 1, 2, \dots, C; i = 1, 2, \dots, N_c$), then the centralized training image set \hat{X} is formulated as follows:

$$\hat{X} = [x_1^1, x_2^1, \dots, x_{N_1}^1, \dots, x_{N_C}^c] \in \mathbb{R}^{q \times l}. \quad (13)$$

Similarly, the image subspace determined by the test image set Y presented by

$$Y = [y_1, y_2, \dots, y_n] \in \mathbb{R}^{q \times n}. \quad (14)$$

For image set Y , $y_{mean} = \frac{1}{n} \sum_{i=1}^n y_i$, centralized as $\hat{y}_i = y_i - y_{mean}$ ($i = 1, 2, \dots, n$), and then the centralized testing image set \hat{Y} is formulated as follows:

$$\hat{Y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n]. \quad (15)$$

Strategy 1. When calculating the manhatta distance between the test image set and the c^{th} image set, the distance between y_{mean} and a training sample X_i can be computed as:

$$d_i = |X_i - y_{mean}| (i = 1, 2, \dots, l). \quad (16)$$

The distance metric set D of the training image set X and y_{mean} is as follows:

$$D = [d_1, d_2, \dots, d_l] \in \mathbb{R}^{1 \times l}. \quad (17)$$

First, we remove the elements corresponding to the c^{th} class from D as $\hat{D} \in \mathbb{R}^{1 \times (L - N_c)}$. Then we sort the elements in \hat{D} in ascend order and select N_c samples x_i^p ($p \neq c$) from X , which corresponds to the smallest N_c distances from \hat{D} to constitute the unrelated subspace U_c .

$$U_c = [u_1^c, u_2^c \dots u_{N_c}^c] \in \mathbb{R}^{q \times N_c}. \quad (18)$$

The classifier based on strategy 1 will be called bilinear regression classification-I (BLRC-I).

Strategy 2. When calculating the distance between the test image set and the c^{th} training image set, assuming that training image set X and test image set Y determine a “virtual” face image space. Different from strategy 1, Strategy 2 does not directly calculate the distance between each image in the training image set X and the center y_{mean} of the test image set. Instead, it calculates the distance between the projection of each image in the training image set on the “virtual” face space and the center of the test image set y_{mean} .

In order to obtain the joint coefficient vector of the two image sets \hat{X} and \hat{Y} , the joint image set E and the test vector e can be constituted as:

$$E = [\hat{X}, -\hat{Y}] \in \mathbb{R}^{q \times (l+n)}, \tag{19}$$

$$e = y_{mean} - x_{mean}. \tag{20}$$

Suppose that $\theta \in \mathbb{R}^{(L+n) \times 1}$ is the joint coefficient vector of \hat{X} and \hat{Y} , which can be calculated by solving the optimization problem

$$\hat{\theta} = \arg \min_{\theta} \|e - E\theta\|^2 + \lambda_1 \|\theta\|_2^2 + \lambda_2 \|\theta\|_1, \tag{21}$$

where $\lambda_1 > 0, \lambda_2 > 0$ and $\lambda_1 + \lambda_2 = 1$.

After solving the regression coefficient $\hat{\theta}$. Then, the Mahalanobis distance between the projection of each image in the training image set X on the ‘‘virtual’’ face space and the center of the test image set can be expressed by the following equation:

$$d_i = |\hat{X}_i \hat{\theta}_i - y_{mean}| (i = 1, 2, \dots, l). \tag{22}$$

The distance metric set D is formulated by

$$D = [d_1, d_2, \dots, d_l] \in \mathbb{R}^{1 \times l}. \tag{23}$$

First, we remove the elements corresponding to the c^{th} class from D as $\hat{D} \in \mathbb{R}^{1 \times (L-N_c)}$. Then we sort the elements in \hat{D} in ascend order and select N_c samples $x_i^p (p \neq c)$ from X , which corresponds to the smallest N_c distances from \hat{D} to constitute the unrelated subspace U_c ,

$$U_c = [u_1^c, u_2^c, \dots, u_{N_c}^c] \in \mathbb{R}^{q \times N_c}. \tag{24}$$

The classifier based on strategy 2 will be called bilinear regression classification-II (BLRC-II).

3.3 Related and Unrelated Distance Metric

Related Distance Metric. In Subject. 3.2, we have obtained the class mean X_{mean}^c for each class in the training set. After centralized processing, the training image set of class c can be converted to

$$\hat{X}_c = [\hat{x}_1^c, \hat{x}_2^c, \dots, \hat{x}_{N_c}^c] \in \mathbb{R}^{q \times N_c}. \tag{25}$$

Now we need to calculate the distance between the test image set \hat{Y} and the c^{th} image set \hat{X}_c in the training set. To obtain the joint regression coefficients of the two image sets, the joint image set S_r^c and test vector s_r^c can be constituted as:

$$S_r^c = [\hat{X}_c, -\hat{Y}] \in \mathbb{R}^{q \times (N_c+n)}, \tag{26}$$

and

$$s_r^c = y_{mean} - x_{mean}^c. \tag{27}$$

Assume that $\gamma^c \in \mathbb{R}^{(N_c+n) \times 1}$ is the joint regression coefficient of \hat{X}_c and \hat{Y} . According to the regression equation $s_r^c = S_r^c \gamma^c$, we can see that the solution of $\gamma^c \in \mathbb{R}^{(N_c+n) \times 1}$ is

$$\hat{\gamma}^c = (S_r^{cT} S_r^c + \lambda I)^{-1} S_r^{cT} s_r^c. \quad (28)$$

Then, the reconstructed “virtual” face image r_1 obtained from the c^{th} training image set \hat{X}_c is

$$r_1 = \hat{X}_c[\hat{\gamma}_1^c, \hat{\gamma}_2^c, \dots, \hat{\gamma}_{N_c}^c]^T + x_{mean}^c. \quad (29)$$

The reconstructed “virtual” face image r_2 obtained from the test image set Y is

$$r_2 = \hat{Y}[\hat{\gamma}_{N_c+1}^c, \hat{\gamma}_{N_c+2}^c, \dots, \hat{\gamma}_{N_c+n}^c]^T + y_{mean}. \quad (30)$$

Finally, the distance between r_1 and r_2 can be used to represent the distance between the test image set and the c^{th} image set in the training set, which is expressed by

$$d_r^c = \|r_1 - r_2\| = \|s_r^c - S_r^c \gamma^c\|. \quad (31)$$

That is, the residual of the linear regression equation $s_r^c = S_r^c \gamma^c$ can be used to represent the distance between the test image set and the c^{th} image set in the training set.

Unrelated Distance Metric. The unrelated image set subspace U_c of the test image set has been obtained in Sect. 3.2. The mean vector of U_c is

$$u_{mean}^c = \frac{1}{N_c} \sum_{i=1}^{N_c} u_i^c. \quad (32)$$

After centralization, the unrelated image set subspace U_c can be converted to

$$\hat{U}_c = [\hat{u}_1^c, \hat{u}_2^c, \dots, \hat{u}_{N_c}^c] \in \mathbb{R}^{q \times N_c}. \quad (33)$$

Now we need to calculate the distance between the test image set \hat{Y} and the unrelated image set subspace U_c . To obtain the joint regression coefficients of two image sets, the joint image set S_u^c and test vector s_u^c can be constituted as

$$S_u^c = [\hat{U}_c, -\hat{Y}] \in \mathbb{R}^{q \times (N_c+n)}, \quad (34)$$

and

$$s_u^c = y_{mean} - u_{mean}^c. \quad (35)$$

Assume that $\delta^c \in \mathbb{R}^{(N_c+n) \times 1}$ is the joint regression coefficient of \hat{U}_c and \hat{Y} . According to the regression equation $s_u^c = S_u^c \delta^c$, it indicates that the solution of $\delta^c \in \mathbb{R}^{(N_c+n) \times 1}$ is

$$\hat{\delta}^c = (S_u^{cT} S_u^c + \lambda I)^{-1} S_u^{cT} s_u^c. \quad (36)$$

Then, the reconstructed “virtual” face image r_1 obtained from the unrelated image set subspace \hat{U}_c is

$$r_1 = \hat{U}_c[\hat{\delta}_1^c, \hat{\delta}_2^c, \dots, \hat{\delta}_{N_c}^c]^T + u_{mean}^c. \quad (37)$$

The reconstructed ‘‘virtual’’ face image r_2 obtained from the test image set Y is

$$r_2 = \hat{Y}[\delta_{N_c+1}^c, \delta_{N_c+2}^c, \dots, \delta_{N_c+n}^c]^T + y_{mean}. \quad (38)$$

Finally, the distance between r_1 and r_2 can be used to represent the distance between the test image set and the unrelated image set subspace, which is expressed by

$$d_u^c = \|r_1 - r_2\| = \|s_u^c - S_u^c \delta^c\|. \quad (39)$$

That is, the residual of the linear regression equation $s_u^c = S_u^c \delta^c$ can be used to represent the distance between the test image set and the unrelated image set subspace.

3.4 Combined Distance Metric

After obtaining the related distance metric d_r^c and the unrelated distance metric d_u^c , we can construct a discriminative criterion by combine the two metric results in a suitable manner. It is obvious that if the test image set belongs to category c , we hope that the distance between the test image set \hat{Y} and the c^{th} image set \hat{X}_c is closer, that is, the d_r^c is as small as possible. on the other hand, it is desirable to make the feature representations between the test image set \hat{Y} and the unrelated image set \hat{U}_c further, that is, the d_u^c is as large as possible. So we propose a new metric d_p^c as

$$d_p^c = \frac{d_r^c}{d_u^c}. \quad (40)$$

The smaller the value of d_p^c , the greater similarity between the test image set and the c^{th} image set. In other words our face image set recognition criterion selects the image set category c when d_p^c takes the minimum value, i.e.

$$\min_{c^*} \{d_p^c \mid c = 1, 2, \dots, C\}. \quad (41)$$

4 Experimental Results

This section provides extensive experimental results to evaluate the performance of two proposed classifiers: BLRC-I and BLRC-II. These experiments are conducted by using several benchmark datasets, i.e., image-based face recognition on the LFW face database [18] and AR face database [10], video-based face recognition on Honda/UCSD face database [8].

4.1 Experiments on LFW

LFW face database were captured in unconstrained environments such that there will be large variations in face images including pose, age, race, facial expression, lighting, occlusions, and background, etc. We use the aligned version of the LFW database, LFW-a to evaluate the recognition performance.

LFW-a contains more than 5,000 subjects. Each subject including images of the same individual in different poses. Note that all the images in LFW-a are of size 250×250 . We manually crop the images into size of 90×78 (by removing 88 pixel margins from top, 72 from bottom, and 86 pixel margins from both left and right sides). An subset of LFW containing 62 persons, each people has more than 20 face images, is used for evaluating the algorithms. Our experimental setting is identical to that in [3]. The first 10 images of each subject are selected to form the training set, while the last 10 images are used as the probe images.

The proposed classifiers are compared with methods including sparse approximated nearest points (SANP) [5,6], affine hull based image set distance (ASIHD) [2], convex hull based image set distance (CSIHD) [2], manifold discriminant analysis (MDA) [13], Dual Linear Regression Based Classification for Face Cluster Recognition (DLRC) [3] and Pairwise Linear Regression Classification for Image Set Retrieval (PLRC) [19]. All methods use the down-scaled images of size of 10×10 and 15×10 as in [3]. The classification results of all methods are illustrated in Table 1. For the images with size of 10×10 , the proposed BLRC-I achieves identical performances with the MDA and PLRC-I method, and the recognition rate is 93.55%, which exceeds other classifiers. For BLRC-II, the recognition rate is 98.39%, obtains the best recognition rate compared with other methods. For images with size of 15×10 , BLRC-I reaches 96.77% recognition rate, BLRC-II, recognition rate is as high as 98.39%. The effects of BLRC-II are higher than those of other classifiers as shown in Table 1.

Table 1. The recognition rates (RR) on LFW database.

Method	10×10	15×10
SANP	85.48	92.55
ASIHD	87.10	95.16
CSIHD	90.32	93.55
MDA	93.55	95.16
DLRC	91.94	95.16
PLRC-I	93.55	96.77
PLRC-II	95.16	96.77
BLRC-I	93.55	96.77
BLRC-II	98.39	98.39

4.2 Experiments on AR

In this section, we study the performance of the proposed classifiers by using the well-known AR database. There are over 4000 face images of 126 subjects (70 men and 56 women) in the database. The face images of each individual contain different expressions, lighting conditions, wearing sun glasses and wearing

scarf. We use the cropped AR database that includes 2600 face images of 100 individuals. First, we manually crop images into a size of 90×70 (by removing 38 pixel margins from top, 39 from bottom, and 24 pixel margins from left and 25 pixel margins right sides). Then downscale the clipped image to get 40×40 resolutions. In the experiments, the first 13 images of each subject are selected to form a training image set, and the remaining 13 images are composed of test image sets.

For this database, the proposed classifiers are compared with following state-of-the-art approaches: SANP [5,6], ASIHD [2], CSIHD [2], DLRC [3] and PLRC [19]. The recognition rates of different classifiers have been presented in Table 2. Experimental results show that compared with other algorithms, the recognition accuracy of the BLRC-I and BLRC-II for image set recognition is as high as 97.98%, which shows obvious improvement on the classification performance.

Table 2. The recognition rates (RR) on AR database.

Methods	RR
SANP	77.00
ASIHD	87.67
CSIHD	84.67
DLRC	96.00
PLRC-I	95.00
PLRC-II	97.33
BLRC-I	97.98
BLRC-II	97.98

4.3 Honda/UCSD Face Database

The Honda/UCSD dataset contains 59 video clips of 20 subjects [8], all but one have at least 2 videos. 20 videos are called training videos and the remainder 39 test videos. The lengths of videos vary from 291 to 1168 frames. In order to maintain the comparability of the experimental results, we use face images consistent with other proceeding work [6].

This dataset has been used extensively for image-based face recognition, the accuracy has reached 100% or close to 100%. Therefore, researchers have turned to experiment on the settings using a small amount frames. We carry out the experiment using the first 50 frames in each video for this database. The shared database by [5] is used. For the video clips that contain less than 50 frames, all frames are selected in the experiment. The following methods are chosen for comparison: DCC [7], MMD [15], MDA [13], AHISD [2], CHISD [2], MSM [17], SANP [5,6], DLRC [3] and PLRC [19]. Table 3 lists all recognition rates of these classifiers on this database. We find that the recognition rates of BLRC-I,

AHISD, RNP, DLRC and PLRC-I are all equal 87.18%, which is much better than those of DCC and MMD methods. The BLRC-II classifier obtains the highest accuracy 92.31% for this database, which is obviously superior to the results of other types of recognition algorithms.

Table 3. The recognition rates (RR) on Honda/UCSD database.

Methods	RR
DCC	70.92
MMD	69.32
MDA	82.05
ASISD	87.18
CSISD	82.05
MSM	74.36
SANP	84.62
DLRC	87.18
PLRC-I	87.18
PLRC-II	89.74
BLRC-I	87.18
BLRC-II	92.31

5 Conclusion

In this paper, bilinear regression classification method (BLRC) is proposed for face image set recognition. Compared to DLRC, BLRC increases the unrelated subspace for classification. Based on different methods of constituting the unrelated subspace, two classifiers are proposed in this paper. In order to validate the performance of two classifiers, some experiments are evaluated on three database for face image set classification tasks. All experimental results confirm the effectiveness of two proposed classification algorithms.

References

1. Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 581–588 (2005)
2. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: Computer Vision and Pattern Recognition, pp. 2567–2573 (2010)
3. Chen, L.: Dual linear regression based classification for face cluster recognition. In: Computer Vision and Pattern Recognition, pp. 2673–2680 (2014)
4. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Press (1967)

5. Hu, Y., Mian, A.S., Owens, R.: Sparse approximated nearest points for image set classification, vol. 42, no. 7, pp. 121–128 (2011)
6. Yiqun, H., Mian, A.S., Owens, R.: Face recognition using sparse approximated nearest points between image sets. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(10), 1992–2004 (2012)
7. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 1005 (2007)
8. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 313–320 (2003)
9. Mahmood, A., Mian, A., Owens, R.: Semi-supervised spectral clustering for image set classification. In: *Computer Vision and Pattern Recognition*, pp. 121–128 (2014)
10. Martínez, A.M., Kak, A.C.: PCA versus LDA. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(2), 228–233 (2001)
11. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(11), 2106–2112 (2010)
12. Sch, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: *International Conference on Pattern Recognition*, pp. 32–36 (2004)
13. Wang, R., Chen, X.: Manifold discriminant analysis. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 429–436 (2009)
14. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: a natural and efficient approach to image set classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2496–2503 (2012)
15. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2008)
16. Wright, J., Ganesh, A., Zhou, Z., Wagner, A., Ma, Y.: Demo: robust face recognition via sparse representation. In: *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–2 (2009)
17. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: *1998 Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 318–323 (1998)
18. Zhu, P., Zhang, L., Hu, Q., Shiu, S.C.K.: Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. In: *European Conference on Computer Vision*, pp. 822–835 (2012)
19. Feng, Q., Zhou, Y., Lan, R.: Pairwise linear regression classification for image set retrieval. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4865–4872 (2016)



Semi-supervised Learning of Deep Difference Features for Facial Expression Recognition

Can Xu, Ruyi Xu, Jingying Chen^(✉), and Leyuan Liu

National Engineering Research Center for E-Learning,
Central China Normal University, Wuhan, China
chenjy@mail.ccnu.edu.cn

Abstract. Facial expression recognition (FER) is an important means of detecting human emotions and is widely applied in many fields, such as affective computing and human-computer interaction. Currently, several methods for FER heavily rely on large amounts of manually labeled data, which are costly and not available in real-world applications. To address this problem, this paper proposes a semi-supervised method based on the deep difference features. First, a cascaded structure is introduced to the original safe semi-supervised SVM (S4VM) to solve the multi-classification task. Then, multiple deep different features are fed to the cascaded S4VM to train the six basic facial expressions using the information of the unlabeled data safely. Extensive experiments show that the proposed method achieved encouraging results on public databases even when using a small labeled sample set.

Keywords: Facial expression recognition · Deep learning · Cascaded S4VM
Semi-supervised method

1 Introduction

Analyzing facial expressions is one of the most important methods of human emotion recognition and facial expressions are defined as the corresponding facial changes in response to a person's inner emotional state and intentions [1]. Nowadays, automatic facial expression recognition (FER) has miscellaneous applications, such as affective computing, interactive games, social psychology, synthetic animation, and intelligent robots [2].

Automatic FER systems can be divided into two categories: those that based on static images and those that based on dynamic image sequences [3]. The static-based method only contains information of the currently input image, while the sequence-based method can use temporal information from multi frames to identify the expression. FER systems receive static images or dynamic sequences as input and then output the corresponding expression category. This work focuses on methods based on the key frames extracted from dynamic image sequences.

In the past two decades, many attempts have been made to recognize facial expressions, and the effectiveness of these attempts depends largely on the size of the labeled training set. A large-scale training set can better reflect the real distribution of samples and hence acquire a better generalization error. However, manual annotation is

demanding, time consuming and expensive [4]. A semi-supervised method can simultaneously use labeled and unlabeled data to improve the classification performance with small datasets, reduce the workload of manual labeling and enhance the practicability of FER [5].

There have been few attempts to recognize facial expressions using a semi-supervised method. Existing methods can be roughly divided into two categories: semi-supervised learning (SSL) [6–8] and semi-supervised clustering [9–11]. SSL exploits the distribution of the unlabeled data to enhance training. Semi-supervised clustering sets the pairwise constraints with labeled data for cluster analysis. In 2004, Cohen et al. [6] were the first to apply SSL to facial expression recognition. They trained probabilistic classifiers with labeled and unlabeled data based on Bayesian networks and achieved an average recognition accuracy of 74.8% on the Cohn-Kanade dataset. Hady et al. [7] mentioned a learning framework to exploit the unlabeled data by the combination of the Co-Training and the one-against-one output-space decomposition approach, which uses Tri-Class SVMs as binary classifiers. The average recognition accuracy on the four basic expressions of the Cohn-Kanade dataset was 86.95%. Jiang et al. [8] focused on the problem of multi-pose facial expression recognition by bringing transfer learning into SSL. Liu et al. [9] addressed the expression recognition in the wild under a semi-supervised frame that combined reference manifold learning with Semi-Supervised Non-negative Matrix Factorization to select discriminant unlabeled data for enhanced training. Liliana et al. [10] proposed a semi-supervised clustering method based on Fuzzy C-means (FCM) to consider the level of ambiguity of facial expressions. Araujo et al. [11] mentioned a semi-supervised temporal clustering method and applied it to the complex problem of facial emotion categorization.

Although the unlabeled samples are helpful to construct the exact model for facial expression classification, experiments show that the effect of some SSL methods is even worse than simply using the methods employed for labeled samples [12, 13]. To address this problem, Li and Zhou presented the safe semi-supervised vector machine (S4VM) [14] to explore multiple candidate low-density separators, estimate the decision boundary closest to the real situation and ensure the best classification effect. The researchers define S4VM as a safe semi-supervised classifier whose performance never degenerates, even when using unlabeled data.

Inspired by Li and Zhou, this work proposes a semi-supervised learning method based on the DPND feature. The DPND feature proposed in our previous work [15] extract the deep representations of the peak (the fully expressive) frame and the neutral frame, respectively, and use the difference between them to represent the facial expression. In this paper, to further improve the robustness, a set of DPND features is extracted from each facial expression sequence which select the key frames near to the cluster centroids. Then, a cascaded semi-supervised classifier is constructed to classify facial expressions with both labeled and unlabeled samples. The final classification result of each sequence is decided by the voting of all key-frame pairs.

The rest of this paper is organized as follows. The details of the semi-supervised FER method are presented in Sect. 2. The experimental setup is described in detail, and the experiment results are given in Sect. 3. Section 4 concludes the paper.

2 The Proposed Method

In this section, the proposed semi-supervised FER approach will be described in detail. The proposed method consists of two main parts: (1) Multiple DPND feature extraction from expression sequences and (2) construction of a cascaded semi-supervised classifier for FER.

2.1 Multiple DPND Feature Extraction

To address the FER problem, researchers have proposed many elaborate features to represent facial expressions during past decades [16]. However, some recent works show that features learned from millions of training samples by deep learning outperform manually designed features in face-related tasks, such as face detection [17] and face recognition [18]. Encouraged by these advancements, the popular VGG-16 [19] is adopted as the network architecture for deep representation extraction in this study. The VGG-16 is pre-trained on the VGG face dataset, which contains 2.6 M face images from 2,622 subjects. When face images are put into the VGG-16, the output of neuron responses by one of the intermediate layers of the VGG-16 network can be extracted as images' deep representation. In this paper, the DPND feature is employed to describe the change between the neutral frame and the peak frame as our previous work [15]:

$$f_{DPND} = (f^P - f^N)/N \quad (1)$$

where f^P and f^N are deep representation features extracted from the peak frame and neutral frame, respectively, and N is the normalized factor. The DPND feature can effectively retain facial expression information while eliminating individual differences and environmental noises.

For some standard facial expression datasets, such as CK+ [20], in which each sequence begins with the neutral expression and ends with the peak expression, the DPND feature can be easily obtained by the deep representation feature of the beginning frame and the end frame. However, the neutral frame and the peak frame of an expression sequence are not directly available in some datasets, such as the BU-4DFE [21]. To extract the DPND feature from expression sequences, a joint method of K-means clustering and rank-SVM is presented.

However, a single DPND feature [15] from each sequence to represent the facial expression has two limitations: first, the extraction of key frames has a certain randomness due to the random initialization of cluster centroids; second, the extracted key frames can only approximately represent the neutral frames and peak frames. In order to further improve the robustness, in this work, a set of DPND features is extracted from each facial expression sequence which select the key frames near to the cluster centroids obtained using K-means. The final classification result of each sequence is decided by the voting of all key-frame pairs. In this way, the multiple DPND feature can effectively avoid the problem caused by the inaccurate selection of key frames. And the subsequent experiments prove that, compared to the single DPND feature, the multiple DPND feature can indeed improve the accuracy of FER.

2.2 Construct a Cascaded Multi-class Classifier for FER

In this subsection, a cascaded classifier is introduced to the S4VM construct to recognize the six basic facial expressions using the proposed DPND feature. The original S4VM proposed by Li and Zhou [14] is an inductive binary classifier. For applying it to FER tasks, a set of S4VMs is combined with a cascaded structure, and each S4VM divides a kind of facial expression from the given dataset. A brief introduction of S4VM is first given.

Safe Semi-Supervised Support Vector Machine (S4VM). Let \mathcal{X} be the input space and $\mathcal{Y} = \{\pm 1\}$ be the label space. A set of labeled data as $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^l$ and a set of unlabeled data are given as $\{\hat{\mathbf{x}}_j\}_{j=1}^u$. Semi-Supervised learning SVM (S3VM) aims to find a decision function $f : \mathcal{X} \rightarrow \{\pm 1\}$ and a label assignment on unlabeled instances $\mathbf{y} = \{y_{l+1}, \dots, y_{l+u}\} \in \mathcal{B}$ such that the following objective function is minimized,

$$h(f, \hat{\mathbf{y}}) = \frac{\|f\|_H}{2} + C_1 \sum_{i=1}^l l(y_i, f(x_i)) + C_2 \sum_{j=1}^u l(\hat{y}_j, f(\hat{x}_j)) \tag{2}$$

S4VM focuses on the safeness of SSL algorithms. Its main idea is to generate multiple low-density separators to approximate the ground truth decision boundary and maximize the improvement in performance of inductive SVMs for any candidate separator. To generate a pool of diverse separators $\{f_i\}_{i=1}^T$, the following function is minimized:

$$\min_{\{f, \hat{\mathbf{y}}_t \in \mathcal{B}\}_{t=1}^T} \sum_{t=1}^T h(f_t, \hat{\mathbf{y}}_t) + M\Omega(\{\hat{\mathbf{y}}_t\}_{t=1}^T), \tag{3}$$

where T is the number of separators, Ω is a penalty coefficient about the diversity of separators, and M is a large constant to ensure diversity. A variety of methods can be adopted to solve this optimization problem, such as global simulated annealing search and representative sampling.

To learn a label assignment \mathbf{y} such that the performance against the inductive SVM, \mathbf{y}^{svm} , is improved, the worst-case improvement over inductive SVM is maximized and $\bar{\mathbf{y}}$ is denoted as the optimal solution:

$$\bar{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y}} \min_{\hat{\mathbf{y}}} \operatorname{gain}(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{y}^{svm}) - \operatorname{loss}(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{y}^{svm}) \tag{4}$$

where $\operatorname{gain}(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{y}^{svm})$ and $\operatorname{loss}(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{y}^{svm})$ are the gained and lost accuracies compared to the inductive SVM, respectively. It has been shown that the accuracy of $\bar{\mathbf{y}}$ is never worse than that of \mathbf{y}^{svm} and achieves the maximal performance improvement over that of \mathbf{y}^{svm} in the worst cases.

Multi-class Classification with the Cascaded S4VM. The original S4VM is typically designed for binary classification problems; thus, S4VM must be extended into a

multi-class classifier for FER. The most common strategies are called one-against-one and one-against-all, however, S4VM, as an inductive method, cannot use one-against-one to construct a multi-class classification, while adoption of one-against-all is ineffective due to the same large training set for each binary classification.

This paper constructs multi-class classification based on a cascaded structure [22, 23], which can hold inductive and effective to unlabeled data. In detail, the training set that contains labeled and unlabeled data is put into the cascaded classifier, and samples of the specified class are picked out for each S4VM classifier. The identified unlabeled data and the corresponding labeled data are removed from the training set, while the remaining samples are passed to the next S4VM classifier.

It is worth noting that the performance of multi-class classifiers varies widely according to different cascaded order. To design a more effective cascaded classifier, the order of the S4VM classifiers is determined according to a discriminant measure of labeled data. The ratio of the inner-class distance and the inter-class distance is defined as the separable measure:

$$S_p = \frac{D_{pp}}{\sum_{q \neq p} D_{pq}} \quad (5)$$

where $D_{pq} = \frac{1}{|p||q|} \sum_{i \in p, j \in q} d_{ij}$ is the average distance between any two samples in the class p and q . The class p is separated from the training set according to the ascending order of S_p . The corresponding classes are sorted to p_1, p_2, \dots, p_m . Then, a classifier with a cascaded structure is constructed, such as that shown in Fig. 1.

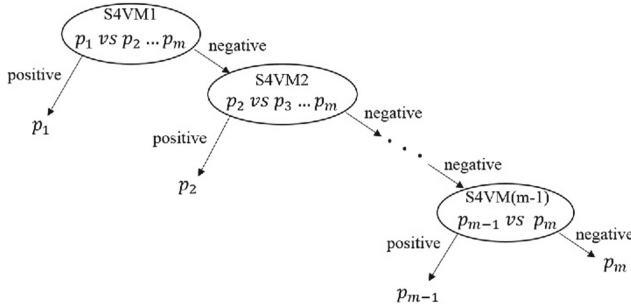


Fig. 1. Multi-class classification based on a cascaded structure.

Samples of class p_1 are assigned to the positive category, and samples of the rest classes are assigned to the negative category; then, the first sub-binary classifier S4VM1 is trained. After that, samples of class p_1 are removed from the training set. Similarly, samples of class p_2 are assigned to the positive category, and the rest of the samples are assigned to the negative category; then, the second sub-binary classifier SVM2 is trained until all the sub-classifiers are trained. Finally, a cascaded S4VM is obtained.

3 Experiments

3.1 Experimental Protocol

To evaluate the effectiveness of the proposed algorithm, two public sequence-based datasets, CK+ [20] and BU-4DFE [21], were chosen for the experiment; the CK+ dataset has been used in [10]. The details of these two datasets are listed in Table 1. In our experiment, only six basic expressions (angry, disgust, fear, happy, sad and surprise) were considered, and we extracted a subset of 53 subjects from the CK+ and a subset of 64 subjects from the BU-4DFE. Some samples of the two databases are shown in Fig. 2. For the CK+ dataset, the DPND feature is the difference between the deep representation feature of the first frame and the last frame; for BU-4DFE, the DPND feature is extracted from the facial sequences directly by our proposed method.

Table 1. Details of the CK+, BU-4DFE dataset.

Dataset	Subjects	Sequences	Gender(F/M)	Age	Ethnicity
CK+	97	486	65%/35%	18–30	Multiethnic
BU-4DFE	101	606	56%/44%	18–70	Multiethnic



Fig. 2. Exemplar expression images in the CK+, BU-4DFE dataset.

3.2 Comparison Among the Multiple DPND, the Single DPND and the DPR Feature

In order to show the effectiveness of the DPND feature, we compared it with the static feature that the deep representations of peak frames (DPR feature) extracted from the VGG-16 network. Then, the proposed cascaded S4VM was employed to evaluate the effects of the different features. For BU-4DFE, the multiple DPND feature was extracted from a set of key-frame pairs near to the cluster centroids. It is noteworthy that the labeled samples only accounted for 10% of the training set in the experiment. The average accuracies of the different features are listed in Table 2. The results indicate that the accuracy of the single DPND feature on the CK+ and BU-4DFE are 8.5% and 21% higher than that of the DPR feature, and the performance of the multiple DPND feature is 3.4% higher than that of the single DPND feature on the BU-4DFE, which strongly proves the excellence of the DPND feature, especially the multiple DPND feature.

Table 2. Average accuracy of the DPND and DPR features.

Feature	CK+	BU-4DFE
Multiple DPND	—	71.8%
Single DPND	89.4%	68.4%
DPR	80.9%	47.4%

3.3 Comparisons with the State-of-the-Art Method

In this subsection, we compare the proposed method (the cascaded S4VM with the DPND feature) with the current state-of-the-art method [10] on the CK+ dataset. The method [10] is based on an SSL algorithm. It first employed an Active Appearance Model to detect human facial points for feature extraction and then utilized semi-supervised Fuzzy C-Means to work as the classifier system; we refer to the method as SSFCM. It selected 329 images of eight emotions from the CK+ dataset, of which 63% were used as a training set and the remaining samples were used for testing. The average accuracies of the proposed method and SSFCM method are shown in Table 3. The proposed method outperforms the SSFCM method [10] even though the SSFCM method selected the peak frames out from the sequences manually and used more labeled data than our method.

Table 3. Average accuracies of the proposed method and the current state-of-the-art method on the CK+.

Method	CK+
Proposed method	89.4%
SSFCM	80.7%

3.4 Comparison with the Supervised Classification

In this subsection, we aimed to use the CK+ and BU-4DFE dataset to evaluate the capability of the SSL method for FER. To this end, the proposed cascaded S4VM and SVM were used as expression classifiers and SVM was considered the baseline because it has been demonstrated as a successful approach for FER tasks. The performance of the cascaded S4VM was calculated based on its outputs, including the list of generated labels for unlabeled data. Using the same data, SVM was applied as a fully supervised version of the cascaded S4VM (see Table 4) for comparison of the semi-supervised learning and supervised learning. The results demonstrate that although a small proportion of each dataset was labelled (10%), the accuracy of the cascaded S4VM for FER on the CK+ and BU-4DFE are 5% and 12% higher than that of SVM.

For more evaluation, the accuracy of the cascaded S4VM was considered with different amounts of labeled data (10%, 12.5%, 17%, 20%, 25% and 50%), as shown in Fig. 3. In all these experiments, the cascaded S4VM achieved better accuracy than SVM, especially in the case of few labelled data, which confirms the cascaded S4VM's efficiency. The results illustrate that combined with information from labeled and

unlabeled samples, the cascaded S4VM can predict the distribution of data more reasonably and then adjust the decision boundary to improve the classification accuracy. Figure 3 also shows that as the number of labeled data increases, the accuracy of the cascaded S4VM and SVM also increase and match.

Table 4. Accuracy of the cascaded S4VM compared to SVM.

Dataset (10%)	SVM	Cascaded S4VM
CK+	84.9%	89.4%
BU-4DFE	59.9%	71.8%

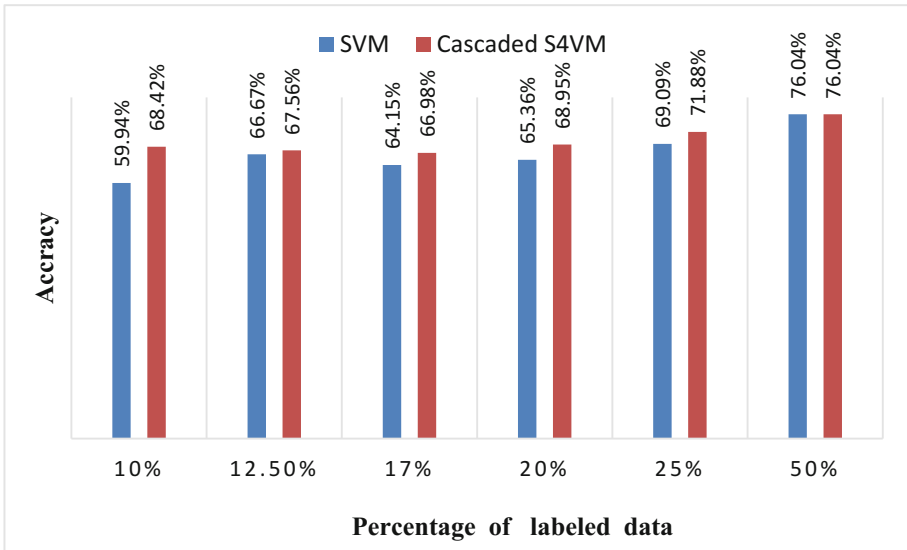


Fig. 3. Accuracy with different percentages of labelled data.

4 Conclusion

In this paper, we propose a semi-supervised method based on the multiple DPND feature for FER. The DPND feature tends to emphasize the facial parts that are changed in the transition from the neutral to the expressive face and to eliminate differences in individual face identities and environmental noises. In this work, the multiple DPND feature are extracted from each sequence to improve the robustness of feature representation. Then, a cascaded semi-supervised classifier is constructed to recognize six basic facial expressions using both labeled and unlabeled data. The proposed method achieves an accuracy of 89.4% on the CK+ dataset and an accuracy of 71.8% on the BU-4DFE dataset when only 10% of the training samples are labeled. The encouraging results on public databases suggests that our method has strong potential to recognize facial expressions in real-world applications.

Acknowledgment. This work was supported by Research funds from the National Key Research and Development Program of China (No. 2018YFB1004500, No. 2018YFB1004504), Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU17ZDJC04) Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJAZH005), National Natural Science Foundation of China (No. 61702208), and Natural Science Foundation of Hubei Province (No. 2017CFB504).

References

1. Li, S.Z., Jain, A.K.: Handbook of Face Recognition, vol. 132, no. 3, pp. 470–487. Springer, Heidelberg (2011)
2. Fang, T., Zhao, X., Ocegueda, O., Shah, S.K.: 3D facial expression recognition: a perspective on promises and challenges. In: IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, vol. 28, pp. 603–610. IEEE (2011)
3. Lopes, A.T., Aguiar, E.D., Souza, A.F.D., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recogn.* **61**, 610–628 (2016)
4. Jiang, B., Jia, K., Sun, Z.: Research on the algorithm of semi-supervised robust facial expression recognition. In: Yoshida, T., Kou, G., Skowron, A., Cao, J., Hacid, H., Zhong, N. (eds.) *AMT 2013*. LNCS, vol. 8210, pp. 136–145. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-02750-0_14
5. Jadidi, Z., Muthukkumarasamy, V., Sithirasanen, E., Singh, K.: Flow-based anomaly detection using semi supervised learning. In: International Conference on Signal Processing and Communication Systems, pp. 1–5. IEEE (2016)
6. Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., Huang, T.S.: Semi supervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(12), 1553–1566 (2004)
7. Hady, M.F.A., Schels, M., Schwenker, F., Palm, G.: Semi-supervised facial expressions annotation using co-training with fast probabilistic tri-class SVMs. In: Diamantaras, K., Duch, W., Iliadis, Lazaros S. (eds.) *ICANN 2010*. LNCS, vol. 6353, pp. 70–75. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15822-3_8
8. Jiang, B., Jia, K.: Semi-supervised facial expression recognition algorithm on the condition of multi-pose. *J. Inf. Hiding Multimed. Sig. Process.* **4**(3), 138–146 (2013)
9. Liu, M., Li, S., Shan, S., Chen, X.: Enhancing expression recognition in the wild with unlabeled reference data. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) *ACCV 2012*. LNCS, vol. 7725, pp. 577–588. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-15822-3_8
10. Liliiana, D.Y., Widyanto, M.R., Basaruddin, T.: Human emotion recognition based on active appearance model and semi-supervised fuzzy C-means. In: International Conference on Advanced Computer Science and Information Systems, pp. 439–445. IEEE (2017)
11. Araujo, R., Kamel, M.S.: A semi-supervised temporal clustering method for facial emotion analysis. In: IEEE International Conference on Multimedia and Expo Workshops, pp. 1–6. IEEE (2014)

12. Wang, L., Chan, K.L., Zhang, Z.: Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval. In: Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 1-629-1-634. IEEE (2003)
13. Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi-supervised support vector machines. *J. Mach. Learn. Res.* **9**(1), 203-233 (2008)
14. Li, Y.F., Zhou, Z.H.: Towards making unlabeled data never hurt. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(1), 175-188 (2015)
15. Chen, J., Xu, R., Liu, L.: Deep peak-neutral difference feature for facial expression recognition. *Multimed. Tools Appl.* **77**(2), 1-17 (2018)
16. Corneanu, C.A., Oliu, M., Cohn, J.F., Escalera, S.: Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(8), 1548-1568 (2016)
17. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: *Computer Vision and Pattern Recognition*, pp. 5325-5334. IEEE (2015)
18. Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S.Z., et al.: When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition, pp. 384-392 (2015)
19. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1-12 (2015)
20. Tian, Y.L., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(2), 97 (2001)
21. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: *International Conference on Automatic Face and Gesture Recognition*, vol. 2006, pp. 211-216. IEEE (2006)
22. Saatci, Y., Town, C.: Cascaded classification of gender and facial expression using active appearance models. In: *International Conference on Automatic Face and Gesture Recognition*, vol. 47, pp. 393-398. IEEE (2006)
23. Li, L., Gao, Z.P., Ding, W.Y.: Fuzzy multi-class support vector machine based on binary tree in network intrusion detection. In: *International Conference on Electrical and Control Engineering*, vol. 28, pp. 1043-1046. IEEE (2010)

Feature Extraction and Selection



Noise Level Estimation for Overcomplete Dictionary Learning Based on Tight Asymptotic Bounds

Rui Chen^{1(✉)} and Changshui Yang²

¹ Tianjin University, Tianjin, China
ruichen@tju.edu.cn

² Peking University, Beijing, China
csyang@pku.edu.cn

Abstract. In this paper, we address the problem of estimating Gaussian noise level from the trained dictionaries in update stage. We first provide rigorous statistical analysis on the eigenvalue distributions of a sample covariance matrix. Then we propose an interval-bounded estimator for noise variance in high dimensional setting. To this end, an effective estimation method for noise level is devised based on the boundness and asymptotic behavior of noise eigenvalue spectrum. The estimation performance of our method has been guaranteed both theoretically and empirically. The analysis and experiment results have demonstrated that the proposed algorithm can reliably infer true noise levels, and outperforms the relevant existing methods.

Keywords: Dictionary learning · Sample covariance matrix
Random matrix theory · Noise level estimation

1 Introduction

The dictionary learning is a matrix factorization problem that amounts to finding the linear combination of a given signal $\mathbf{Y} \in \mathbb{R}^{N \times M}$ with only a few atoms selected from columns of the dictionary $\mathbf{D} \in \mathbb{R}^{N \times K}$. In an overcomplete setting, the dictionary matrix \mathbf{D} has more columns than rows $K > N$, and the corresponding coefficient matrix $\mathbf{X} \in \mathbb{R}^{K \times M}$ is assumed to be sparse. For most practical tasks in the presence of noise, we consider a contamination form of the measurement signal $\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{w}$, where the elements of noise \mathbf{w} are independent realizations from the Gaussian distribution $\mathcal{N}(0, \sigma_n^2)$. The basic dictionary learning problem is formulated as:

$$\min_{\mathbf{D}, \mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \quad s.t. \quad \|\mathbf{x}_i\|_0 \leq L \quad \forall i \quad (1)$$

Therein, L is the maximal number of non-zero elements in the coefficient vector \mathbf{x}_i . Starting with an initial dictionary, this minimization task can be solved by the popular alternating approaches such as the method of optimal directions (MOD) [1] and K-SVD [2]. The dictionary training on noisy samples can incorporate the denoising together

into one iterative process [3]. For a single image, the K-SVD algorithm is adopted to train a sparsifying dictionary and the developed method in [3] denoises the corrupted image by alternating between the update stages of the sparse representations and the dictionary. In general, the residual errors of learning process are determined by noise levels. Noise incursion in a trained dictionary can affect the stability and accuracy of sparse representation [4]. So the performance of dictionary learning highly depends on the estimation accuracy of unknown noise level σ_n^2 when the noise characteristics of trained dictionaries are unavailable.

The main challenge of estimating the noise level lies in effectively distinguishing the signal from noise by exploiting sufficient prior information. The most existing methods have been developed to estimate the noise level from image signals based on specific image characteristics [5–8]. Generally, these works assume that a sufficient amount of homogeneous areas or self-similarity patches are contained in natural images. Thus empirical observations, singular value decomposition (SVD) or statistical properties can be applied on carefully selected patches. However, it is not suitable for estimating the noise level in dictionary update stage because only few atoms for sparse representation cannot guarantee the usual assumptions. To enable wider applications and less assumptions, more recent methods estimate the noise level based on principal component analysis (PCA) [9, 10]. These methods underestimate the noise level since they only take the smallest eigenvalue of block covariance matrix. Although later work [11] has made efforts to tackle these problems by spanning low dimensional subspace, the optimal estimation for true noise variance is still not achieved due to the inaccuracy of subspace segmentation. As for estimating the noise variance techniques, the scaled median absolute deviation of wavelet coefficients has been widely adopted [12]. Leveraging the results from random matrix theory (RMT), the median of sample eigenvalues is also used as an estimator of noise variance [13]. However, these estimators are no longer consistent and unbiased when the dictionary matrix has high dimensional structure.

To solve the aforementioned problems, we propose to accurately estimate the noise variance in a trained dictionary by using exact eigenvalues of a sample covariance matrix. The proposed method can also be applied to estimate the noise level for the noisy image. As a novel contribution, we construct the tight asymptotic bounds of extreme eigenvalues to separate the subspaces between the signal and the noise based on random matrix theory (RTM). Moreover, in order to eliminate the possible bias caused by the high-dimensional settings, a corrected estimator is derived to provide the consistent inference on the noise variance for a trained dictionary. Based on these asymptotic results, we develop an optimal variance estimator which can well deal with the settings with different sample sizes and dimensions. The practical usefulness of our method is numerically illustrated.

2 Tight Bounds for Noise Eigenvalue Distributions

In this section, we analyze the asymptotical distribution of the ratio of extreme eigenvalues of a sample covariance matrix based on the limiting RTM law. Then a tight bound is derived.

2.1 Eigenvalue Subspaces of Sample Covariance Matrix

We consider the sparse approximation of each observed sample $\mathbf{y}_i \in \mathbb{R}^N$ with s prototype atoms selected from learned dictionary \mathbf{D} . With respect to the sparse model (1), we aim at estimating the noise level σ_n^2 for an elementary trained dictionary \mathbf{D}_s containing a subset of the atoms $\{\mathbf{d}_i\}_{i=1}^s$. Note that $\mathbf{D}_s = \mathbf{D}_s^0 + \mathbf{w}_s$, where \mathbf{D}_s^0 denotes original dictionary and \mathbf{w}_s is the additive Gaussian noise. At each iterative step, the noise level σ_n^2 goes gradually to zero when updating towards true dictionary \mathbf{D}_s^0 [14]. The known noise variance is helpful to avoid noise incursion and determine the sample size, the sparsity degree and even the performance of the true underlying dictionary [15]. To derive the relationship between the eigenvalues and noise level, we first construct the sample covariance matrix of dictionary \mathbf{D}_s as follows:

$$\Sigma_S = \frac{1}{s-1} \sum_{i=1}^s (\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}})^T, \quad \bar{\mathbf{d}} = \frac{1}{s} \sum_{i=1}^s \mathbf{d}_i \tag{2}$$

According to (2), the square matrix Σ_S has N dimensions with the sparse condition $N \gg s$. Based on the symmetric property, this matrix is decomposed into the product of three matrices: an orthogonal matrix \mathbf{U} , a diagonal matrix and a transpose matrix \mathbf{U}^T , which can be selected by satisfying $\mathbf{U}^T \mathbf{U} = \mathbf{I}$. Here, this transform process is written as:

$$\mathbf{U}^T \Sigma_S \mathbf{U} = \mathbf{diag}(\lambda_1, \dots, \lambda_m, \lambda_{m+1}, \dots, \lambda_N) \tag{3}$$

Given $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$, we exploit the eigenvalue subspaces to enable the separation of atoms from noise. To be more specific, we divide the eigenvalues into two sets $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2$ by finding the appropriate bound in a spiked population model [16]. Most structures of an atom lie in low-dimension subspace and thus the leading eigenvalues in set $\mathbf{S}_1 = \{\lambda_i\}_{i=1}^m$ are mainly contributed by atom itself. The redundant-dimension subspace $\mathbf{S}_2 = \{\lambda_i\}_{i=m+1}^N$ is dominated by the noise. Because the atoms contribute very little to this later portion, we take all the eigenvalues of \mathbf{S}_2 into consideration to estimate the noise variance while eliminating the influence of trained atoms. Moreover, the random variables $\{\lambda_i\}_{i=m+1}^N$ can be considered as the eigenvalues of pure noise covariance matrix Σ_w , whose dimensions are N .

2.2 Asymptotic Bounds for Noise Eigenvalues

Suppose the sample matrix Σ_w has the form $(s-1)\Sigma_w = \mathbf{H}\mathbf{H}^T$, where the sample entries of \mathbf{H} are independently generated from the distribution $\mathcal{N}(0, \sigma_n^2)$. Then the real matrix $\mathbf{M} = \mathbf{H}\mathbf{H}^T$ follows a standard Wishart distribution [17]. The ordered eigenvalues of \mathbf{M} are denoted by $\bar{\lambda}_{\max}(\mathbf{M}) \geq \dots \geq \bar{\lambda}_{\min}(\mathbf{M})$. In the high dimensional situation: $N/s \rightarrow \gamma \in [0, \infty)$ as $s, N \rightarrow \infty$, the Tracy-Widom law gives the limiting distribution of the largest eigenvalue of the large random matrix \mathbf{M} [18]. Then we have the following asymptotic expression:

$$\Pr \left\{ \frac{\bar{\lambda}_{\max}/\sigma_n^2 - \mu}{\xi} \leq z \right\} \rightarrow F_{\text{TW1}}(z) \tag{4}$$

where $F_{\text{TW1}}(z)$ indicates the cumulative distribution function with respect to the Tracy-Widom random variable. In order to improve both the approximation accuracy and convergence rate, even only with few atom samples, we need choose the suitable centering and scaling parameters μ, ξ [19]. By the comparison between different values, such parameters are defined as

$$\begin{cases} \mu = 1/s \cdot (\sqrt{s-1/2} + \sqrt{N-1/2})^2 \\ \xi = 1/s \cdot (\sqrt{s-1/2} + \sqrt{N-1/2}) \left(\frac{1}{\sqrt{s-1/2}} + \frac{1}{\sqrt{N-1/2}} \right)^{1/3} \end{cases} \tag{5}$$

The empirical distribution of the eigenvalues of the large sample matrix converges almost surely to the Marcenko-Pastur distribution on a finite support [20]. Based on the generalized result in [21], when $N \rightarrow \infty$ and $\gamma \in [0, \infty)$, with probability one, we derive limiting value of the smallest eigenvalue as

$$\bar{\lambda}_{\min}/\sigma_n^2 \rightarrow (1 - \sqrt{\gamma})^2 \tag{6}$$

According to the asymptotic distributions described in the theorems (4) and (6), we further quantify the distribution of the ratio of the maximum eigenvalue to minimum eigenvalue in order to detect the noise eigenvalues. Let T_1 be a detection threshold. Then we find T_1 by the following expression:

$$\begin{aligned} \Pr \left\{ \frac{\bar{\lambda}_{\max}}{\bar{\lambda}_{\min}} \leq T_1 \right\} &= \Pr \left\{ \frac{\bar{\lambda}_{\max}}{\sigma_n^2} \leq T_1 \cdot \frac{\bar{\lambda}_{\min}}{\sigma_n^2} \right\} \approx \Pr \left\{ \frac{\bar{\lambda}_{\max}}{\sigma_n^2} \leq T_1 \cdot (1 - \sqrt{N/s})^2 \right\} \\ &= \Pr \left\{ \frac{\bar{\lambda}_{\max}/\sigma_n^2 - \mu}{\xi} \leq \frac{T_1 \cdot (1 - \sqrt{N/s})^2 - \mu}{\xi} \right\} \approx F_{\text{TW1}} \left\{ \frac{T_1 \cdot (1 - \sqrt{N/s})^2 - \mu}{\xi} \right\} \end{aligned} \tag{7}$$

Note that there is no closed-form expression for the function F_{TW1} . Fortunately, the values of F_{TW1} and the inverse F_{TW1}^{-1} can be numerically computed at certain percentile points [16]. For a required detection probability α_1 , this leads to

$$\frac{T_1 \cdot (1 - \sqrt{N/s})^2 - \mu}{\xi} = F_{\text{TW1}}^{-1}(\alpha_1) \tag{8}$$

Plugging the definitions of μ and ξ into the Eq. (8), we finally obtain the threshold

$$T_1 = \frac{s(\sqrt{s-1/2} + \sqrt{N-1/2})^2}{(\sqrt{s} - \sqrt{N})^2} \cdot \left(\frac{(\sqrt{s-1/2} + \sqrt{N-1/2})^{-2/3}}{(s-1/2)^{1/6}(N-1/2)^{1/6}} \cdot F_{\text{TW1}}^{-1}(\alpha_1) + 1 \right) \tag{9}$$

When the detection threshold T_1 is known in the given probability, it means that an asymptotic upper bound can also be obtained for determining the noise eigenvalues of the matrix $\Sigma_{\mathbf{w}}$ because the equality $\lambda_{m+1}/\lambda_N = \bar{\lambda}_{\max}/\bar{\lambda}_{\min}$ holds. In general, the noise eigenvalues in the set \mathbf{S}_2 surround the true noise variance as it follows the Gaussian distribution. The estimated largest eigenvalue λ_{m+1} should be no less than σ_n^2 . The known smallest eigenvalue λ_N is no more than σ_n^2 by the theoretical analysis [11]. The location and value of λ_{m+1} in \mathbf{S} are obtained by checking the bound $\lambda_{m+1} \leq T_1 \cdot \lambda_N$ with high probability α_1 . In addition, λ_1 cannot be selected as noise eigenvalue λ_{m+1} .

3 Noise Variance Estimation Algorithm

3.1 Bounded Estimator for Noise Variance

Without requiring the knowledge of signal, the threshold T_1 can provide good detection performance for finite s , N even when the ratio N/s is not too large. Based on this result, more accurate estimation can be obtained by averaging all elements in \mathbf{S}_2 . Hence, the maximum likelihood estimator of σ_n^2 is

$$\hat{\sigma}_n^2 = \frac{1}{N - m} \sum_{j=m+1}^N \lambda_j \tag{10}$$

In the low dimensional setting where N is relatively small compared with s , the estimator $\hat{\sigma}_n^2$ is consistent and unbiased as $s \rightarrow \infty$. It follows asymptotically normal distribution as

$$\sqrt{s}(\hat{\sigma}_n^2 - \sigma_n^2) \rightarrow \mathcal{N}(0, t^2), \quad t^2 = \frac{2\sigma_n^4}{N - m} \tag{11}$$

When N is large with respect to the sample size s , the sample covariance matrix shows significant deviations from the underlying population covariance matrix. In this context, the estimator $\hat{\sigma}_n^2$ might have a negative bias, which leads to overestimation of true noise variance [22, 23]. We investigate the distribution of another eigenvalue ratio. Namely, the ratio of the maximum eigenvalue to the trace of the eigenvalues is

$$U = \frac{\lambda_{m+1}}{1/(N - m) \cdot \text{tr}(\Sigma_{\mathbf{w}})} = \frac{\lambda_{m+1}}{1/(N - m) \cdot \sum_{j=m+1}^N \lambda_j} \tag{12}$$

According to the result in (4), the ratio U also follows a Tracy-Widom distribution as both N , $s \rightarrow \infty$. The denominator in the definition of U is distributed as an independent $\sigma_n^2 \chi_N^2/N$ random variable, and thus has $E(\hat{\sigma}_n^2) = \sigma_n^2$ and $\text{Var}(\hat{\sigma}_n^2) = 2\sigma_n^4/(N \cdot s)$. It is easy to show that replacing σ_n^2 by $\hat{\sigma}_n^2$ results in the same limiting distribution in (4). Then we have

$$\Pr\left\{\frac{\lambda_{m+1}/\hat{\sigma}_n^2 - \mu}{\xi} \leq z\right\} \rightarrow F_{\text{TW1}}(z) \tag{13}$$

Unfortunately, the asymptotic approximation present in (13) is inaccurate for small and even moderate values of N [24]. This approximation is not a proper distribution function. The simulation observations imply that the major factor contributing to the poor approximation is the asymptotic error caused by the constant ξ [24]. Therefore, a more accurate estimate for the standard deviation of $\lambda_{m+1}/\hat{\sigma}_n^2$ will provide a significant improvement. For finite samples, we have

$$\mathbb{E}\left(\frac{\lambda_{m+1}}{\sigma_n^2}\right) = \mu, \quad \mathbb{E}\left(\frac{\lambda_{m+1}^4}{\sigma_n^4}\right) = \mu^2 + \xi^2 \tag{14}$$

Using these asymptotic results, we get the corrected deviation

$$\xi' = \sqrt{\frac{N \cdot s}{2 + N \cdot s} \left(\xi^2 - \frac{2}{N \cdot s} \mu^2\right)} \tag{15}$$

Note that this formula in (15) has corrected the overestimation in the high dimensional setting. thus the better approximation for the probabilities of the ratio is

$$\Pr\left\{\frac{\lambda_{m+1}/\hat{\sigma}_n^2 - \mu}{\xi'} \geq z\right\} \approx 1 - F_{\text{TW1}}(z) \tag{16}$$

The determination of the distribution for the ratio U is devoted to the correction of the variance estimator. In order to complete the detection of the large deviations of the initial estimator $\hat{\sigma}_n^2$, we provide a procedure to set the threshold T_2 . Based on the result in (16), an approximate expression for the overestimation probability is given by

$$\Pr\left\{\frac{\hat{\sigma}_n^2}{\lambda_{m+1}} \leq T_2\right\} = \Pr\left\{\frac{\lambda_{m+1}/\hat{\sigma}_n^2 - \mu}{\xi'} \geq \frac{1/T_2 - \mu}{\xi'}\right\} \approx 1 - F_{\text{TW1}}\left(\frac{1/T_2 - \mu}{\xi'}\right) \tag{17}$$

Hence, for a desired probability level α_2 , the above equation can be numerically inverted to find the decision threshold. After some simplified manipulations, we obtain

$$T_2 = \frac{1}{\xi' \cdot F_{\text{TW1}}^{-1}(1 - \alpha_2) + \mu} \tag{18}$$

Asymptotically, the spike eigenvalue λ_{m+1} converges to the right edge of the support $\sigma_n^2(1 + \sqrt{N/s})$ as N, s go to infinity. According to the expression in (18), this function turns out to have a simple approximation $T_2 = 1/\mu$ in the high probability case. Then the upper bound $T_2 \cdot \lambda_{m+1}$ for the known $\hat{\sigma}_n^2$ yields a bias estimation. Finally, the following expectation holds true:

$$\mathbb{E} \left(\frac{\mu \cdot T_2 \cdot \lambda_{m+1}}{1 + \sqrt{N/s}} \right) \approx \sigma_n^2 \ll \hat{\sigma}_n^2 \quad (19)$$

By analyzing the statistical result in (19), the correction for $T_2 \cdot \lambda_{m+1}$ can be approximated as the better estimator than $\hat{\sigma}_n^2$ because this bias-corrected estimator is closer to the true variance under the high dimensional conditions. If $\hat{\sigma}_n^2$ can satisfy the requirement of no excess of the bound $T_2 \cdot \lambda_{m+1}$, the sample eigenvalues are consistent estimates of their population counterparts. Hence, the optimal estimator is given by

$$\hat{\sigma}_*^2 = \min \left\{ \hat{\sigma}_n^2, \frac{\mu \cdot T_2 \cdot \lambda_{m+1}}{1 + \sqrt{N/s}} \right\} \quad (20)$$

3.2 Implementation

Based on the construction of two thresholds, we propose a noise estimation algorithm for dictionary learning as follows:

Algorithm 1 Noise Estimation for Dictionary Learning

- 1: Input:** Noisy dictionary \mathbf{D}_s , the dimension N , the sample number s , the probability levels α_1 and α_2 .
 - 2: Compute** the eigenvalues $\{\lambda_i\}_{i=1}^N$ of the sample covariance matrix Σ_s , and order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$.
 - 3: Compute** two thresholds T_1 and T_2 .
 - 4: for** $i=1:N-1$ **do**
 - if** $\lambda_{i+1} \leq T_1 \cdot \lambda_N$ **then**
 - Obtain the location $m+1=i+1$, $\lambda_{m+1}=\lambda_{i+1}$ and **break**
 - end if**
 - end for**
 - 5: Estimate** an initial noise variance $\hat{\sigma}_n^2$ using (10).
 - 6: Compare** the values of two estimators of (20) and select the minimum as an optimal estimator $\hat{\sigma}_*^2$.
 - 7: Output:** noise level estimation $\sigma_n^2 = \hat{\sigma}_*^2$.
-

4 Numerical Experiments

The proposed estimation method is evaluated on two benchmark datasets: Kodak [7] and TID2008 [9]. The subjective experiment is to compare our method with three state-of-the-art estimation methods by Liu et al. in [8], Pyatykh *et al.* in [9] and Chen *et al.* in [11], which are relevant in SVD domain. The testing images are added to the

independent white Gaussian noise with deviation level 10 and 30, respectively. We set the probabilities $\alpha_1, \alpha_2 = 0.97$ and choose $N = 256, s = 3$. In general, a higher noise estimation accuracy leads to a higher denoising quality. We use the K-SVD method to denoise the images [3]. Figures 1 and 2 show the results using our method outperform other competitors. Moreover, our peak signal-to-noise ratios (PSNRs) are nearest to true values, 32.03 dB and 27.01 dB, respectively.

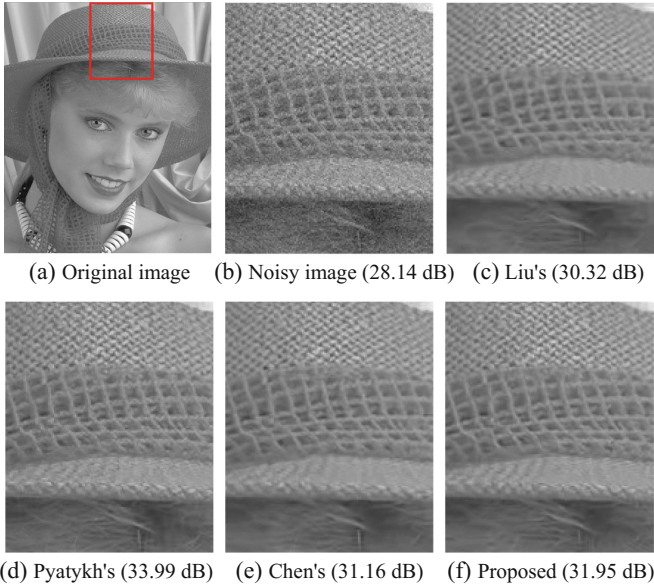


Fig. 1. Denoising results on the *Woman* image using K-SVD.

To quantitatively evaluate the accuracy of noise estimation, the average of standard deviations, mean square error (MSE), mean absolute difference (MAD) are computed by randomly selecting 1500 image patches from 20 testing images. The results shown in Table 1 indicate that the proposed method is more accurate and stable than other methods. Next, we compare our optimal estimator $\hat{\sigma}_*^2$ with $\hat{\sigma}_n^2$ and other two existing estimators in the literatures. The simulated realization of a sample covariance matrix is followed a Gaussian distribution with different variances. As presented in Table 2, the performance of $\hat{\sigma}_*^2$ is invariably better than other estimators. To test robustness of our estimation method, we further obtain the empirical probabilities of the estimated eigenvalues at typical confidence levels. Figure 3 illustrates that two asymptotic bounds can achieve very high success probabilities.

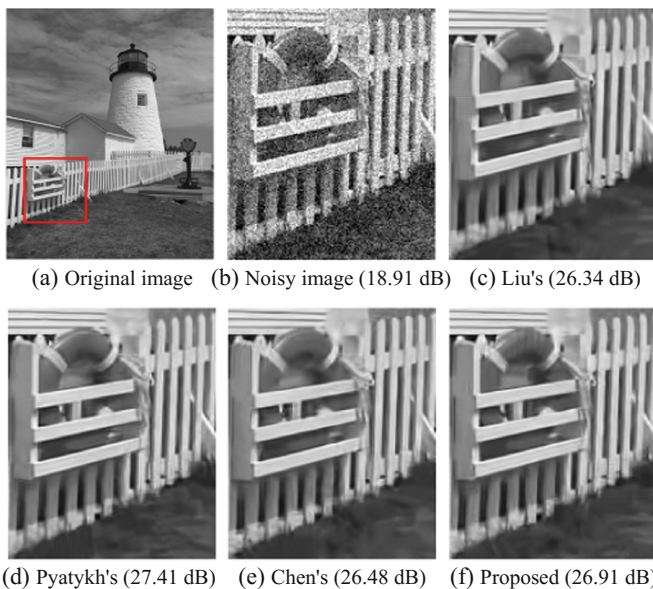


Fig. 2. Denoising results on the *House* image using K-SVD.

Table 1. Estimation results of different methods (Best results are highlighted).

σ_n	Liu's [8]	Pyatykh's [9]	Chen's [11]	Proposed
1	2.18	1.34	0.59	1.16
5	7.30	3.83	5.41	5.27
10	13.86	7.19	11.83	10.19
15	16.72	13.91	15.92	15.17
20	20.99	18.75	20.62	19.90
25	26.64	23.29	24.34	25.06
30	32.38	27.27	31.98	30.12
MAD	3.30	1.59	0.98	0.15
MSE	4.84	3.22	1.39	0.03

Table 2. Estimation results of four estimators (Best results are highlighted).

σ_n	$\hat{\sigma}_{\text{median}}$ [23]	$\hat{\sigma}_{\text{US}}$ [13]	$\hat{\sigma}_n$	$\hat{\sigma}_*$
1	1.27	1.99	1.14	1.06
5	4.59	5.27	6.24	5.18
10	8.76	11.28	9.97	9.94
15	15.22	14.29	16.17	14.93
20	20.85	19.14	20.96	20.10
25	25.87	25.98	26.31	25.28
30	30.59	30.37	31.16	30.11
MAD	0.64	0.78	0.86	0.12
MSE	0.52	0.72	0.99	0.02

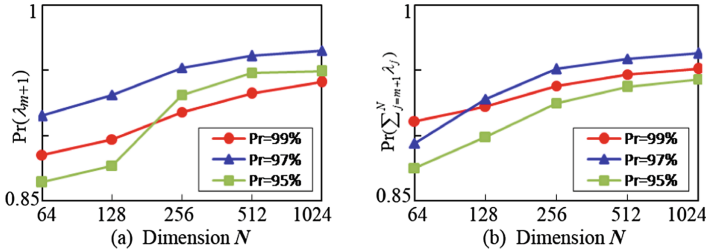


Fig. 3. Empirical probabilities of exact noise eigenvalue estimation.

5 Conclusions

In this paper, we have shown how to infer the noise level from a trained dictionary. The eigen-spaces of the signal and noise are transformed and separated well by determining the eigen-spectrum interval. In addition, the developed estimator can effectively eliminate the estimation bias of a noise variance in high dimensional context. Our noise estimation technique has low computational complexity. The experimental results have demonstrated that our method outperforms the relevant existing methods over a wide range of noise level conditions.

References

- Engan, K., Aase, S., Husoy, J.: Method of optimal directions for frame design. In: Proceedings of International Conference on Acoustics, Speech, and Signal Pattern Process (ICASSP), pp. 2443–2446 (1999)
- Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **54**(11), 4311–4322 (2006)
- Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.* **15**(12), 3736–3745 (2006)
- Sahoo, S., Makur, A.: Enhancing image denoising by controlling noise incursion in learned dictionaries. *IEEE Signal Process. Lett.* **22**(8), 1123–1126 (2015)
- Li, D., Zhou, J., Tang, Y.: Noise level estimation for natural images based on scale-invariant kurtosis and piecewise stationarity. *IEEE Trans. Image Process.* **26**(2), 1017–1030 (2017)
- Hashemi, M., Beheshti, S.: Adaptive noise variance estimation in BayesShrink. *IEEE Signal Process. Lett.* **17**(1), 12–15 (2010)
- Tang, C., Yang, X., Zhai, G.: Noise estimation of natural images via statistical analysis and noise injection. *IEEE Trans. Circuit Syst. Video Technol.* **25**(8), 1283–1294 (2015)
- Liu, W., Lin, W.: Additive white gaussian noise level estimation in SVD domain for images. *IEEE Trans. Image Process.* **22**(3), 872–883 (2013)
- Pyatykh, S., Hesser, J., Zhang, L.: Image noise level estimation by principal component analysis. *IEEE Trans. Image Process.* **22**(2), 687–699 (2013)
- Liu, X., Tanaka, M., Okutomi, M.: Single-image noise level estimation for blind denoising. *IEEE Trans. Image Process.* **22**(12), 5226–5237 (2013)

11. Chen, G., Zhu, F., Heng, P.: An efficient statistical method for image noise level estimation. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 477–485 (2015)
12. Donoho, L., Johnstone, I.: Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455 (1994)
13. Ulfarsson, M., Solo, V.: Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. Signal Process.* **56**(12), 5804–5816 (2008)
14. Gribonval, R., Jenatton, R., Bach, F.: Sparse and spurious: dictionary learning with noise and outliers. *IEEE Trans. Inf. Theory* **61**(11), 6298–6319 (2015)
15. Jung, A., Eldar, Y., Gortz, N.: On the minimax risk of dictionary learning. *IEEE Trans. Inf. Theory* **62**(3), 1501–1515 (2016)
16. Johnstone, I.M.: On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**(2), 295–327 (2001)
17. Chiani, M.: On the probability that all eigenvalues of Gaussian, Wishart, and double Wishart random matrices lie within an interval. *IEEE Trans. Inf. Theory* **63**(7), 4521–4531 (2017)
18. Karoui, N.E.: A rate of convergence result for the largest eigenvalue of complex white Wishart matrices. *The Annals of Probability* **34**(6), 2077–2117 (2006)
19. Ma, Z.M.: Accuracy of the Tracy-Widom limits for the extreme eigenvalues in white Wishart matrices. *Bernoulli* **18**(1), 322–359 (2012)
20. Marcenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb.* **1**(4), 457–483 (1967)
21. Bai, Z., Silverstein, J.: *Spectral Analysis of Large Dimensional Random Matrices*, 2nd edn. Springer, New York (2010). <https://doi.org/10.1007/978-1-4419-0661-8>
22. Kritchman, S., Nadler, B.: Determining the number of components in a factor model from limited noisy data. *Chem. Int. Lab. Syst.* **94**(1), 19–32 (2008)
23. Passelier, D., Li, Z., Yao, J.: On estimation of the noise variance in high dimensional probabilistic principal component analysis. *J. R. Stat. Soc. B* **79**(1), 51–67 (2017)
24. Nadler, B.: On the distribution of the ratio of the largest eigenvalue to the trace of a Wishart matrix. *J. Multivar. Anal.* **102**, 363–371 (2011)



Perceptual Compressive Sensing

Jiang Du, Xuemei Xie^(✉), Chenye Wang, and Guangming Shi

School of Artificial Intelligence, Xidian University, Xi'an 710071, China
jiangdu@ieee.org, xmxie@mail.xidian.edu.cn, cywang_dd@163.com,
gmshi@xidian.edu.cn

Abstract. Compressive sensing (CS) works to acquire measurements at sub-Nyquist rate and recover the scene images. Existing CS methods always recover the scene images in pixel level. This causes the smoothness of recovered images and lack of structure information, especially at a low measurement rate. To overcome this drawback, in this paper, we propose perceptual CS to obtain high-level structured recovery. Our task no longer focuses on pixel level. Instead, we work to make a better visual effect. In detail, we employ perceptual loss, defined on feature level, to enhance the structure information of the recovered images. Experiments show that our method achieves better visual results with stronger structure information than existing CS methods at the same measurement rate.

Keywords: Compressive sensing · Perceptual loss
Fully convolutional network · Low-level computer vision
Semantic reconstruction

1 Introduction

Nowadays, information is one of the most important component in human world. Visual information takes up most of the percentage. There are billions of images and videos around our daily life. Computer vision has underwent huge resurgence in recent years, since deep learning has made a significant difference in this field. Researchers have shown that deep learning has made breakthrough achievements in the following two broad categories. The first category is the high-level computer vision tasks. For example, image and video classification or recognition [26, 27], object detection [8, 29], image caption [19], and visual tracking [21]. The second category is low-level reconstruction tasks. For example, denoising [17, 35], super-resolution [16], style transfer [13], and optical flow estimation [10].

Researches on inverse problems in imaging [20, 22] have been carried on for decades, which cover various low-level computer vision tasks. Compressive sensing (CS) [2, 3, 5] is a typical inverse problem in imaging. Conventional CS works to recover the signal by optimization algorithms [4, 7]. However, this model is hard to be implemented and costs much computational complexity. The application of deep neural networks in inverse problems in imaging makes it possible

that the CS measurements can be recovered real-time. Data-driven CS [14, 24, 25] learns the recovery network from the training data. Adp-Rec [36] jointly train the coder-decoder and brings significant improvement on reconstruction quality. Fully convolutional measurement network (FCMN) [6] firstly measures and recovers full images. However, all the above methods focus on pixel level, and ignore the high-level structure information. This makes the reconstructed results look smooth and have unsatisfactory visual effect. To overcome the drawback, we consider to add high-level perceptual information to CS. So the question is, how to add high-level perceptual information on the low-level CS task.

Design of loss function is a promising solution for perceptual recovery. Study on loss functions for low-level computer vision tasks has provided a variety of approaches. For example, mean square error (MSE) loss, L1 loss [17], NRMSE loss [33] and constraint loss [39]. Recently, perceptual loss [13] has been proposed and employed in many reconstruction tasks, such as style transfer [13] and super-resolution [16]. They are a combination of low-level detailed information and high level semantic information. Perceptual loss is widely used to achieve these goals. It is because perceptual loss is defined in feature space, which can convert the ability of extracting high-level semantic information to recovery network. Thus, the recovered images will contain rich structure information. Inspired by the above applications, we propose perceptual CS, which focuses more on sensing and recovering structure information. We use FCMN [6] as base network to measure and recover scene images, and adopt perceptual loss to train it. We surprisingly find that this framework is capable of capturing and recovering the structure information, especially at extremely low measurement rate, where the measurements can merely contain very limited amount of information.

The contribution of this paper is that, we propose perceptual CS, which can measure and recover the structure information of scene images. It should be noted that, only one deconvolution layer and one Res-block are used in our framework as an illustration. One can employ a deeper network if necessary.

Moreover, perceptual CS indicates an universal architecture. One can change the loss network using pre-trained or dynamic feature extractors for more specific tasks. In this paper, we use VGG [32] as an example. Our code is available on github¹ for further reproduction.

The organization of the rest part of this paper is as follows. Section 2 introduces some related works of this paper. Section 3 describes the technical design and theoretical analysis of the proposed framework. Section 4 presents experimental results of perceptual CS and gives detailed analysis. Section 5 draws the conclusion.

2 Related Work

2.1 Compressive Sensing

CS [5, 15, 34] proves signal can be reconstructed after being sampled at sub-Nyquist rates as long as the signal is sparse in a certain domain. Reconstructing

¹ <https://github.com/jiang-du/Perceptual-CS>.

signal from measurements is an ill-posed problem. Traditional CS usually solves an optimization problem, which leads to high computational complexity.

Recently, deep neural networks (DNNs) has been applied to CS tasks [6, 14, 23–25, 36]. These DNN-based methods can be divided into two categories depending on whether measurement and reconstruction process are trained jointly. The first category trains the recovery network while the measurement part is fixed, like SDA [25], ReconNet [14], and DeepInverse [24]. SDA [25] first applies deep learning approach to solve the CS recovery problem, which uses fully-connected layers in the recovery part. ReconNet [14] uses a fully-connected layer along with convolutional layers to recover signals block by block. While, DeepInverse [24] uses pure convolutional layers. The random Gaussian fashion of the measurement part would mismatch the learned recovery part.

The second category jointly trains the measurement part and the recovery part, such as Deepcodec [23], Adaptive [36], and FCMN [6]. These methods totally overcome the problem that the measurement part is independent from the recovery part. Deepcodec [23] is a framework where both measurement and approximate inverse process are learned end-to-end by a deep fully-connected encoder-decoder network. In [36], a fully-connected layer as the measurement matrix along with a super-resolution network as the recovery part is trained. FCMN [6] firstly uses a fully convolutional network where the measurement part is implemented with an overlapped convolution operation. All these methods recover the scene image on pixel level. They ignore the structure information of images.

2.2 Perceptual Loss

Recently, perceptual loss [13] is widely used in many image reconstruction tasks [9, 11, 13, 16, 30, 38]. It can recover the image with better visual effect since it is defined on feature space. Typically, perceptual loss calculates the Euclidean distance between the features maps of the reconstructed images and the labels from the same layer of the same pre-trained classification network. Perceptual loss reflects the similarity in the feature level between the label and output images, which makes the reconstructed images retain high-level structure information. In contrast, per-pixel loss focuses on similarity in pixel level, which only preserves low-level pixel information.

Perceptual loss achieves more excellent performance than per-pixel loss in most of image restoration tasks. For example, Johnson et al. [13] use perceptual loss for style transfer and super resolution. The output images have sharper edges compared to per-pixel loss. SRGAN [16] trained by perceptual loss generates more photo-realistic super-resolved images than by MSE loss. When used in image inpainting [30], perceptual loss produces satisfactory results due to the addition of high-level context. Additionally, perceptual loss helps to remain finer details for image editing [38]. Inspired by the advantages of perceptual loss in preserving structure and detail, we attempt to apply it to CS field and it accordingly performs well.

3 Perceptual CS Framework

In this section, we mainly introduce the technical design of the perceptual CS framework. The architecture is shown in Fig. 1. It consists of two parts: *compressive sensing network* and *perceptual loss network*. The compressive sensing network originally performs reconstruction in pixel-wise manner. With the perceptual loss network added, the perceptual CS network preserves the structure information of the recovered images. With the help of perceptual recovery, the proposed network is able to acquire high-level perceptual information.

The compressive sensing network measures and recovers the full scene images. The full image processing fashion provides an enough receptive field that makes it possible to perform perceptual reconstruction. While, in the perceptual loss network, we employ a classification network, VGG19, as an auxiliary network. It plays the role of extracting the perceptual information of the images.

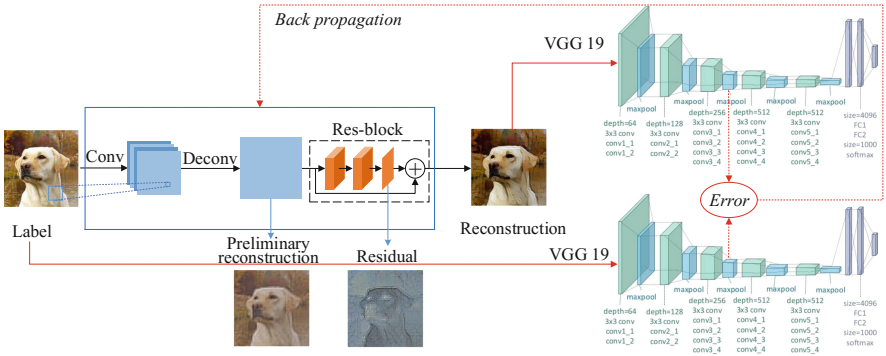


Fig. 1. The architecture of perceptual CS network.

3.1 Full Image Compressive Sensing Network

In most existing CS methods, the scene image is measured and recovered block by block, and each block is reshaped into a column vector. This breaks the structure of the full image. Besides, the computational complexity of the existing methods will extremely increase when the size of the image becomes larger. For example, when an image with the size of $n \times n$ is measured, the memory consumption of the sensing matrix can be up to $S(n) = O(n^4)$. Thus, it is nearly impossible to design a large sensing matrix, let alone measuring the full image. This is because the mapping from the scene image to the measurements is fully-connected, leading to an extremely large-scale parameter nightmare.

Inspired by fully convolutional measurement network (FCMN) [6], we employ a fully convolutional architecture to measure and recover the scene images in the proposed framework, which can get rid of the disaster of the exploding number

of parameters. The first convolution layer plays the role of measurement matrix, with kernel size 32 and stride 16. This indicates that the size of measurement matrix is 32×32 and the sliding step is 16 pixels. The deconvolution layer right after the measurement part transforms the dimension of feature map back to the same as input image. Moreover, the fully convolutional architecture can preserve the correspondence among pixels (instead of reshaping into column vector). In this way, block-effect has been largely removed in the recovered images due to the overlapped convolutional measurement. This preserves the structure information of the whole image. Furthermore, the full image method makes it possible to use perceptual loss for semantic reconstruction.

Although the convolution and deconvolution layers can recover the image, for better visual effect, we enhance the proposed framework with residual learning. In detail, we add one residual block and it works quite well, as is shown in Fig. 2(b). One can add more residual blocks for further improvements if necessary.

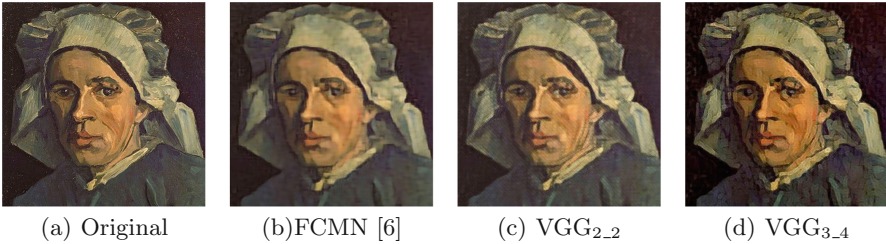


Fig. 2. The original image ‘head of a peasant woman with white cap’ by Van Gogh and the reconstructed images with different methods at 4% measurement rate. Here the proposed method uses the conv2.2 and conv3.4 of VGG19 [32] as different scale of loss respectively.

3.2 Perceptual Reconstruction for Compressive Sensing

In the proposed network, we focus on the perceptual recovery. In the classic CS task, the recovery network approximates the error in the pixel-wise space. To extract the structure information, we recover the scene image in feature-level space. Instead of MSE loss, we consider the perceptual loss, which focuses on perceptual recovery.

MSE loss: In classic CNN-based CS, the loss function is usually defined with pixel-wise loss:

$$l_{pixel}(w) = \|f\{x, w\} - x\|_2^2. \quad (1)$$

This pixel-wise loss will force the image to have the minimized average Euclidean distance between the reconstruction images $f\{x, w\}$ and the labels x . Here, w represents the parameters of the whole network, including the measurement and the recovery parts. Although MSE loss in (1) can help to achieve the reconstructed images with high peak signal-to-noise ratio (PSNR), the reconstructed

images usually look smooth and the structure information is not clear. We can see in Fig. 2(b) that the face and the hat of the person is very smooth compared with the original image in Fig. 2(a). Especially the wrinkle on the face cannot be clearly seen.

Perceptual Loss: Considering the current popular classification network works by extracting the features in an image, we can take this advantage into our proposed method. Thus, we apply the perceptual loss. It is formulated as

$$l_{feat}^{\phi,j}(w) = \|\phi_j(f\{x, w\}) - \phi_j(x)\|_2^2, \quad (2)$$

where $\phi_j(x)$ denotes the feature map of the j -th layer of VGG19 with the input image x . Different from (1), a typical kind of perceptual loss is defined with the (squared, normalized) Euclidean distance between the feature maps generated from the reconstructed image and the label. Actually, when applying CS at a very low measurement rate, we do not care much about the detailed texture of it. Correspondingly, we emphasize the importance of the structural information. As is shown in Fig. 2(c) and (d), the structure information recovered better, especially the hat of the person has richer structure information compared with Fig. 2(b).

In practical, we define the loss function on VGG_{2,2} or VGG_{3,4} of VGG19 (actually pooling 2 or pooling 3) as examples. The results can be addressed in Fig. 2(c) and (d). The feature map of bottom layers contains detailed low-level information and the top layers have more high-level semantic features. We can also choose other layers by different requirements. In this paper, We do not apply perceptual loss by too high level layers because in terms of compressive sensing, higher level drops too much information that it is nearly impossible to inverse, even if pre-trained.

4 Experiments with Analysis

In this section, we conduct the experiments to illustrate the performance of the proposed perceptual CS framework. We test our framework with a standard dataset [14] containing 11 grayscale images. We also compare the reconstruction results with some typical CS methods. Furthermore, we take some reconstruction results as examples to make a detailed analysis of the performance of the proposed method.

Experiment Setup. The learning rate is set to 10^{-8} when perceptual loss is defined on VGG_{2,2}, and 10^{-9} when perceptual loss is defined on VGG_{3,4}. The bench size is set to 5 while training. For each measurement rate, the iteration time is 10^6 . We use the caffe [12] framework for network training and MATLAB for testing. Our computer is equipped with Intel Core i7-6700K CPU with frequency of 4.0 GHz, 4 NVidia GeForce GTX Titan XP GPUs, 128 GB RAM, and the framework runs on the Ubuntu 16.04 operating system. The training dataset consists of 800 pieces of images with size 256×256 down sampled and cropped from 800 images in DIV2K dataset [1].

Results with Analysis. The following is the analysis of the experimental results at different measurement rates.

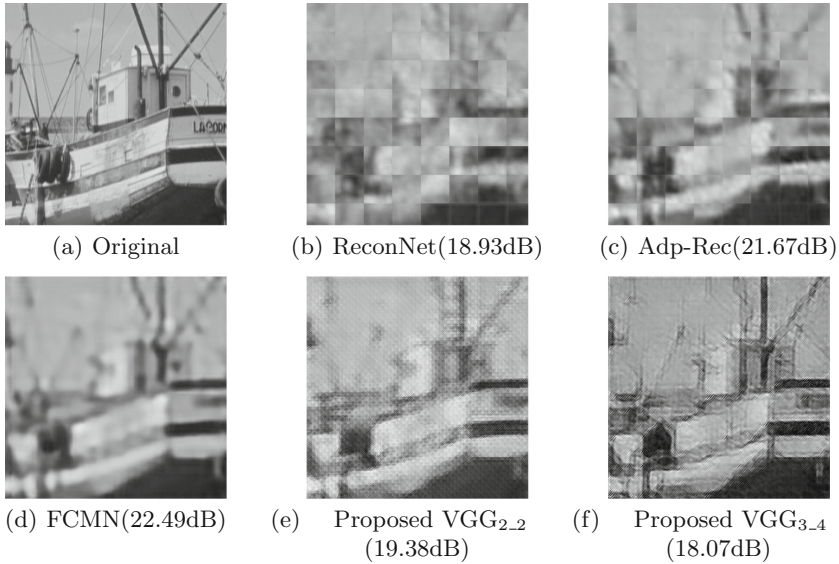


Fig. 3. Boats at measurement rate 1%. (b) and (c) are of block-wise. (d) is of full-image. They all use MSE loss. (e) and (f) are improved by using perceptual loss [2]. Perceptual CS brings stronger structure information compared with FCMN.

The explanation from Fig. 3 at measurement rate 1% is as follows.

- (1) Block effect occurs in Fig. 3(b) and (c) by block-wise methods such as ReconNet [14] and Adp-Rec [36].
Based on the standard ReconNet [14], the improved ReconNet [18] adds several tricks such as adaptive measurement and adversarial loss. Its performance is even lower than Adp-Rec [36].
- (2) Figure 3(d) has no block artifacts in FCMN [6] where fully-convolutional measurement is employed. This work achieves the state-of-the-art results in terms of PSNR and SSIM.
In this experiment, all existing CS-based image reconstruction works rely on MSE loss. While, FCMN [6] makes perceptual loss promising.
- (3) Perceptual loss in Fig. 3(e) and (f) enhances structure information, even if PSNR is lower compared with Fig. 3(d).

The explanation of measurement rate at 4% in Fig. 4 is as follows:

- (1) Block effect also occurs in Fig. 4(c) in DR²-Net [37].
DR²-Net achieves highest PSNR among random Gaussian methods, since it adds several Res-blocks that fully convergence in the reconstruction stage.

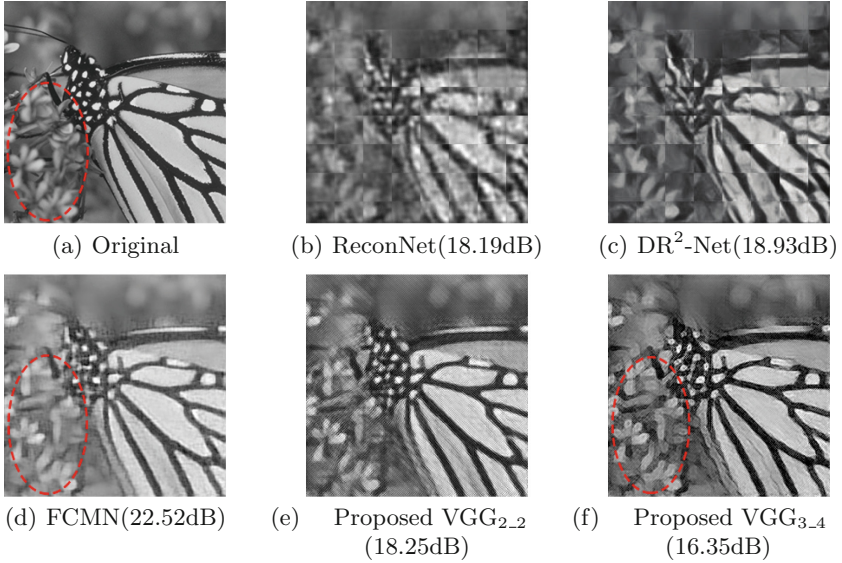


Fig. 4. Monarch at measurement rate 4%. (b) and (c) are of block-wise. (d) is of full-image. They all use MSE loss. (e) and (f) are improved with perceptual loss. They have stronger structure information than the state-of-the-art result in FCMN. Specially, we can see in the **red circle** of (f), compared with (a) and (d), that even blurry image can be enhanced. (Color figure online)

- (2) The method with adaptive measurement for Fig. 4(d) adopts one Res-block, achieving the highest PSNR. The comparison among several typical methods including DR²-Net is in Fig. 4, where FCMN [6] with full image gets the best result in terms of PSNR. It should be pointed out that only one Res-block is used in both FCMN [6] and the proposed framework in this paper. One can add more Res-blocks for further improvement.
- (3) With just one Res-block, perceptual loss in Fig. 4(e) and (f) works well, which improves FCMN [6]. Structure information is kept. In some case, even weak structure can become strong (see Fig. 4(f) compared to Fig. 4(a) and (d)).

It should be noted that, even if PSNR is worse with perceptual loss, the structure information is clearly reconstructed.

Evaluation of Perceptual CS. To evaluate the performance of the proposed method, we evaluate quality of the reconstructed images with PSNR and SSIM. Furthermore, we also use Mean Opinion Score (MOS) [28] to test the visual effect of these methods. In this metric, an image is scored by 26 volunteers and the final score is the average value. The quality ranking is represented by scores from 1 to 5, where 1 denotes lowest quality and 5 denotes the highest. All the test images are ranked randomly before being scored and they are displayed group

Table 1. Mean PSNR, SSIM and MOS of different methods

MR = 1%	ReconNet	DR ² -Net	Adp-Rec	FCMN	VGG _{2,2}	VGG _{3,4}
PSNR	17.27	17.44	20.32	21.27	18.30	16.80
SSIM	0.4083	0.4291	0.5031	0.5447	0.2478	0.2565
MOS	1.0734	1.1188	1.8496	2.6328	2.6818	2.9510
MR = 4%						
PSNR	19.99	20.80	24.01	23.87	19.38	16.72
SSIM	0.5287	0.5804	0.7021	0.7042	0.3522	0.4729
MOS	1.5979	1.7237	3.0489	3.4230	3.4755	3.3566

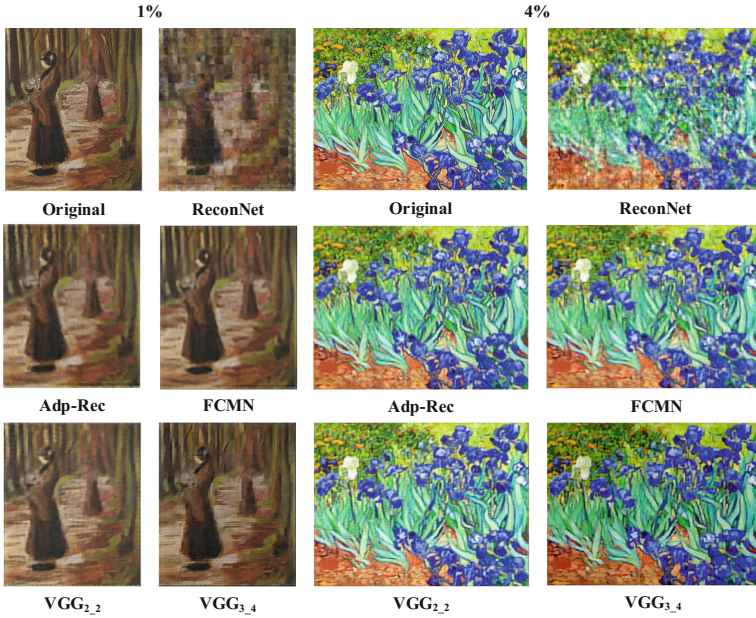


Fig. 5. The reconstructed results of ReconNet [14], Adp-Rec [36], FCMN [6], and the proposed method using the conv2.2 and conv3.4 of VGG19 [32] with measurement rate 1% and 4% and their corresponding original scene image.

by group. Each group has six reconstruction images, in different methods. All participants take this test on the same computer screen, from the same angle and distance. Here the distance from the screen to the tested persons is 50 cm and the eyes of those persons are of the same height of the center of the screen.

The detailed comparison results of mean PSNR, SSIM and MOS is shown in Table 1. we can draw the following conclusion. Our method achieves the highest MOS rating. The PSNR and SSIM value of typical methods is higher, since their loss function is defined as the Euclidean distance between the output and label.

While, perceptual CS concentrates more on the visual effect. Thus, it is helpful for MOS, instead of PSNR and SSIM.

Moreover, we give some examples of color images. In terms of color channels, we measure and recover the RGB channels respectively, and then combine them to a whole color image. The results of perceptual CS with color images are shown in Fig. 5. Of course, we give the comparison with existing methods. We can see obviously from the figure that the visual effect of perceptual CS is quite well.

In terms of hardware implementation, we follow the approach of the existing work proposed in [31] in which sliding window is used to measure the scene. Similarly, we can replace the random Gaussian measurement matrix with the learned pre-defined parameters in the convolution layer of the measurement network. The reconstruction part is not on optical device, so only the measurement part needs to be implemented with the approach above.

5 Conclusion

In this paper, we propose perceptual CS for sensing and recovering structured scene images. The proposed framework managed to recover structure information from CS measurements. Our work is of profound significance, which may open a door towards alternative to semantic sensing and recovery.

Acknowledgements. This work is supported by Natural Science Foundation (NSF) of China (61836008, 61472301, 61632019), and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 61621005).

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: dataset and study. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, vol. 3, p. 2 (2017)
2. Baraniuk, R.G.: Compressive sensing [lecture notes]. IEEE Sig. Process. Mag. **24**(4), 118–121 (2007). <https://doi.org/10.1109/MSP.2007.4286571>
3. Baraniuk, R.G.: More is less: signal processing and the data deluge. Science **331**(6018), 717–719 (2011)
4. Candes, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
5. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006). <https://doi.org/10.1109/TIT.2006.871582>
6. Du, J., Xie, X., Wang, C., Shi, G., Xu, X., Wang, Y.: Fully convolutional measurement network for compressive sensing image reconstruction. Neurocomputing (2018). <https://doi.org/10.1016/j.neucom.2018.04.084>
7. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Sel. Top. Sig. Process. **1**(4), 586–597 (2008)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision (ICCV), October 2017

9. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2458–2467, July 2017
10. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
12. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, MM 2014, pp. 675–678. ACM, New York (2014). <https://doi.org/10.1145/2647868.2654889>
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
14. Kulkarni, K., Lohit, S., Turaga, P., Kerviche, R., Ashok, A.: ReconNet: non-iterative reconstruction of images from compressively sensed measurements. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 449–458, June 2016
15. Kunis, S., Rauhut, H.: Random sampling of sparse trigonometric polynomials, II. Orthogonal matching pursuit versus basis pursuit. *Found. Comput. Math.* **8**(6), 737–763 (2008)
16. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
17. Lefkimmiatis, S.: Non-local color image denoising with convolutional neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5882–5891. IEEE (2017)
18. Lohit, S., Kulkarni, K., Kerviche, R., Turaga, P., Ashok, A.: Convolutional neural networks for non-iterative reconstruction of compressively sensed images. *IEEE Trans. Comput. Imaging* **4**(3), 326–340 (2018)
19. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7219–7228 (2018)
20. Lucas, A., Iliadis, M., Molina, R., Katsaggelos, A.K.: Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Sig. Process. Mag.* **35**(1), 20–36 (2018). <https://doi.org/10.1109/MSP.2017.2760358>
21. Luo, W., Sun, P., Zhong, F., Liu, W., Zhang, T., Wang, Y.: End-to-end active object tracking via reinforcement learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, stockholmssäsan, Stockholm, Sweden, 10–15 July 2018, vol. 80, pp. 3286–3295 (2018)
22. McCann, M.T., Jin, K.H., Unser, M.: Convolutional neural networks for inverse problems in imaging: a review. *IEEE Sig. Process. Mag.* **34**(6), 85–95 (2017). <https://doi.org/10.1109/MSP.2017.2739299>
23. Mousavi, A., Dasarathay, G., Baraniuk, R.G.: DeepCodec: adaptive sensing and recovery via deep convolutional neural networks. In: 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), p. 744, October 2017. <https://doi.org/10.1109/ALLERTON.2017.8262812>

24. Mousavi, A., Baraniuk, R.G.: Learning to invert: signal recovery via deep convolutional networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2272–2276 (2017)
25. Mousavi, A., Patel, A.B., Baraniuk, R.G.: A deep learning approach to structured signal recovery. In: Communication, Control, and Computing, pp. 1336–1343 (2016)
26. Rahmani, H., Mian, A., Shah, M.: Learning a deep model for human action recognition from novel viewpoints. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(3), 667–681 (2018). <https://doi.org/10.1109/TPAMI.2017.2691768>
27. Ranjan, R., et al.: Deep learning for understanding faces: machines may be just as good, or better, than humans. *IEEE Sig. Process. Mag.* **35**(1), 66–83 (2018). <https://doi.org/10.1109/MSP.2017.2764116>
28. ITU-R Recommendation: Recommendation 500-10; methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT. 500-10 (2000)
29. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767) (2018)
30. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1225–1233, July 2017
31. Shi, G., Gao, D., Song, X., Xie, X., Chen, X., Liu, D.: High-resolution imaging via moving random exposure and its simulation. *IEEE Trans. Image Process.* **20**(1), 276–282 (2011). <https://doi.org/10.1109/TIP.2010.2052271>
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
33. Sun, J., Li, H., Xu, Z., et al.: Deep ADMM-Net for compressive sensing MRI. In: Advances in Neural Information Processing Systems, pp. 10–18 (2016)
34. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **53**(12), 4655–4666 (2007)
35. Xie, X., Du, J., Shi, G., Hu, H., Li, W.: An improved approach for visualizing dynamic vision sensor and its video denoising. In: Proceedings of the International Conference on Video and Image Processing, ICVIP 2017, pp. 176–180. ACM, New York (2017). <https://doi.org/10.1145/3177404.3177411>
36. Xie, X., Wang, Y., Shi, G., Wang, C., Du, J., Han, X.: Adaptive measurement network for CS image reconstruction. In: Yang, J., et al. (eds.) CCCV 2017. CCIS, vol. 772, pp. 407–417. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-7302-1_34
37. Yao, H., Dai, F., Zhang, D., Ma, Y., Zhang, S., Zhang, Y.: DR²-net: deep residual reconstruction network for image compressive sensing. arXiv preprint [arXiv:1702.05743](https://arxiv.org/abs/1702.05743) (2017)
38. Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6882–6890, July 2017
39. Zhang, J., Ghanem, B.: ISTA-Net: interpretable optimization-inspired deep network for image compressive sensing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1828–1837 (2018)



Differential and Integral Invariants Under Möbius Transformation

He Zhang^{1,2}(✉), Hanlin Mo^{1,2}, You Hao^{1,2}, Qi Li^{1,2}, and Hua Li^{1,2}

¹ Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
zhanghe@ict.ac.cn

² University of Chinese Academy of Sciences, Beijing, China

Abstract. One of the most challenging problems in the domain of 2-D image or 3-D shape is to handle the non-rigid deformation. From the perspective of transformation groups, the conformal transformation is a key part of the diffeomorphism. According to the Liouville Theorem, an important part of the conformal transformation is the Möbius transformation, so we focus on Möbius transformation and propose two differential expressions that are invariable under 2-D and 3-D Möbius transformation respectively. Next, we analyze the absoluteness and relativity of invariance on them and their components. After that, we propose integral invariants under Möbius transformation based on the two differential expressions. Finally, we propose a conjecture about the structure of differential invariants under conformal transformation according to our observation on the composition of above two differential invariants.

Keywords: Conformal transformation · Möbius transformation
Differential invariant · Integral invariant

1 Introduction

One of the most challenging problems in the domain of 2-D image or 3-D shape is to handle the non-rigid deformation, especially in the situation of anisotropy, which is universal in the real world. In the viewpoint of transformation groups, the isometric transformation is a prop subgroup of the conformal transformation, which is a prop subgroup of the diffeomorphism. Obviously, the anisotropic non-rigid transformation exceeds the boundary of isometric transformation and contains conformal transformation. Based on the Erlangen program of Klein, geometry is a discipline that studies the properties of space that remain unchanged under a particular group of transformation. In order to solve the anisotropic transformation problem, it is necessary to find the invariants under the conformal transformation.

The original motivation of conformal mapping is how to flatten the map of globe, and the Mercator projection produce an angle-preserving map that is very useful for navigation. More generally, the conformal geometry focuses on

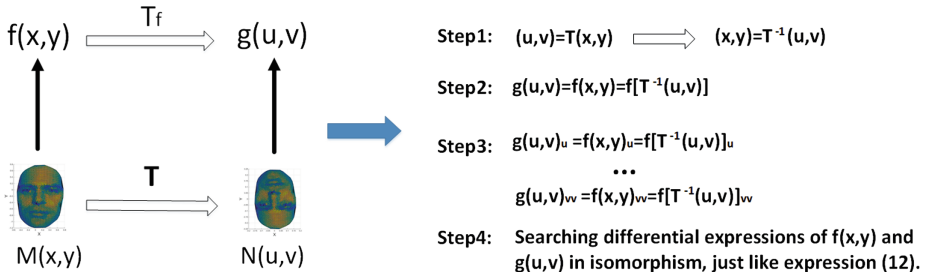


Fig. 1. A brief flowchart of the method.

the shape in which the only measure is angle instead of usually length. The descriptions of conformal mapping contain angle preservation [5, 12, 26], metric rescaling [21, 27], preservation of circles [14, 28], etc. Some key ideas reside in the conformal surface geometry are Dirac equation [6], Cauchy-Riemann equation [22], Möbius transformations [27, 28], Riemann mapping [9, 10, 33, 35], Ricci flow [34], etc. The conformal geometry lies between the topology geometry and the Riemannian geometry, it studies the invariants of the conformal transformation group. The *conformal structures* [9, 10] based on the theories of Riemann surfaces are invariants under conformal transformation. According to conformal geometry [7], the *shape factor* [10] and *conformal module* [35] are conformal invariants. Moreover, the *conformal inner product* [26] defined by an inner product of function is also changeless under conformal transformation. According to the Liouville Theorem [8, 20], the Möbius transformation plays an important role in conformal mapping.

The definition of Möbius transformation [25] shows that it is compounded by a series of simple transformations: Translation, Stretching, Rotation, Reflection and Inversion. In the domain of invariants under translation, stretching and rotation transformations, the Geometric moment invariants (GMIs) [32] and the ShapeDNA [17] show a general method to generate the moment invariants; Hu et al. [13] proposed a general construction method of surface isometric moment invariants based on the intrinsic metric. In the domain of invariants under reflection transformation, the chiral invariants [36] show the moment invariants based on the generating functions of ShapeDNA [17]. In the domain of invariants under conformal transformation, Hu [12] proposed limited conformal invariants based on geodesic tangent vectors. In the domain of invariants under Möbius transformation, the expression $(H^2 - K)dA$ proposed by Blaschke [1] is proved to be a conformal invariant by Chen [4]; based on the *Gauss-Bonnet* Theorem, White [30] proposed that $\int_M H^2 dA$ is a global conformal invariant if M is an oriented and closed surface. The *Gauss-Bonnet* Theorem associates the differential expression (Gaussian curvature) of the surface S with its topological invariant $\chi(S)$ (the Euler's characteristic). This great theorem motivates us to explore the differential invariants under the Möbius transformation since the differential expressions play essential roles in some procedures of physics, mathematics,

computer science and other fields. In the domain of differential invariants, rotation and affine differential invariants were proposed by Olver [23] based on the moving frame method; a special type of affine differential invariants was presented by Wang et al. [29]; Li et al. [19] prove the existence of projective moment invariants of images with relative projective differential invariants; the research [18] on the relationship between differential invariants and moment invariants show that they are isomorphic under affine transformation.

In this article, we study invariants by combining functional map [24] and the derivatives of function (see Fig. 1). In Sect. 2, we show the background of this paper. In Sect. 3, we propose the invariants under Möbius transformation. In Sect. 4, we show another Möbius invariant from the functional view. Finally, we propose a conjecture about the structure of differential invariants under conformal transformation. The main contributions of this paper are as follows.

- We propose two differential expressions that are invariant under 2-D and 3-D Möbius transformation respectively. According to the Liouville Theorem, the 3-D differential invariant is a conformal invariant.
- Based on the analysis on absoluteness and relativity of invariance about the two differential expressions and their components, we propose integral invariants under Möbius transformation.
- We propose a conjecture about the composition of differential invariants under conformal transformation.

2 Notion and Background

2.1 Notion

The formulation in this paper is same with the functional maps framework [24]. Assuming M and N are two manifolds, a bijective mapping $T : M \rightarrow N$ induces the transformation $T_F : \mathcal{F}(M, \mathbb{R}) \rightarrow \mathcal{F}(N, \mathbb{R})$ of derived quantities, where $\mathcal{F}(\cdot, \mathbb{R})$ is scalar function defined on manifold. It means that any function $f : M \rightarrow \mathbb{R}$ have a counterpart function $g : N \rightarrow \mathbb{R}$ and $g = f \circ T^{-1}$.

To make the invariants under Möbius transformation clear, we partially modify original definition and theorem in this paper with this formulation.

2.2 Theoretic Background

According to the Liouville Theorem [20], the only conformal mapping in $R^n (n > 2)$ are Möbius transformation [11, 15, 25]. Furthermore, the Generalized Liouville Theorem shows that any conformal mapping defined on $D (D \in \mathbb{R}^n, n > 2)$ must be a restriction of Möbius transformation.

Theorem 1 (Generalized Liouville Theorem [8]). *Suppose that D, D' are domains in \mathbb{R}^n and that $T : D \rightarrow D'$ is a homeomorphism. If $n = 2$, then T is 1-quasiconformal if and only if T or its complex conjugate is a meromorphic function of a complex variable in D . If $n \geq 3$, then T is 1-quasiconformal if and only if T is the restriction to D of a Möbius transformation, i.e., the composition of a finite number of reflections in $(n - 1)$ -spheres and planes.*

Next, we will show the common expressions of Möbius transformation in different dimensions ($n \geq 2$).

In the field of complex analysis, a Möbius transformation could be expressed as

$$T(z) = \frac{az + b}{cz + d}, \tag{1}$$

where $a, b, c, d, z \in \mathbb{C}$, $ad - bc \neq 0$. Based on the Liouville Theorem [20], every Möbius transformation in higher dimensions could be given with the form

$$T(x) = b + \frac{\gamma A(x - a)}{\|x - a\|_2^\epsilon}, \tag{2}$$

where $x, a, b \in \mathbb{R}^n$, ϵ is 0 or 2, $\gamma \in \mathbb{R}$ and $A \in \mathbb{R}_{n \times n}$ is an orthogonal matrix. The choice of ϵ decides if $T(x)$ contains inversion transformation, and the sign of $\det(A)$ decides if $T(x)$ contains reflection transformation.

More generally, a Möbius transformation could be composed of a series of simple transformations (Fig. 2), the definition of Möbius transformation is as below.

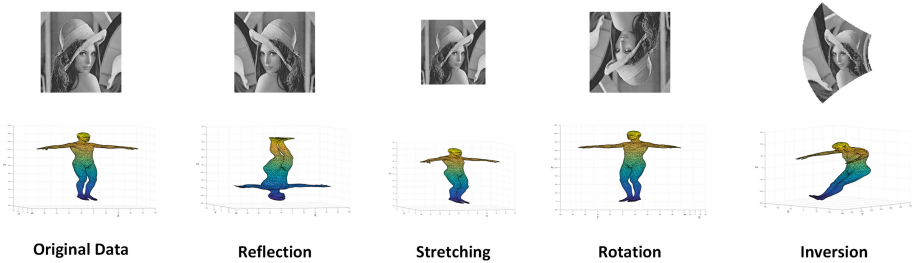


Fig. 2. Some elementary transformations of Möbius transformation.

Definition 1 (Möbius transformation [25]). A n -dimension Möbius transformation is a homomorphism of $\overline{\mathbb{R}^n}$ (the one-point compactification of \mathbb{R}^n), it is a mapping $T : \overline{\mathbb{R}^n} \rightarrow \overline{\mathbb{R}^n}$ that is a finite composition of the following elementary transformations ($x \in \mathbb{R}^n$):

- (1) Translation: $T_a(x) = x + a$, $a \in \mathbb{R}^n$.
- (2) Stretching: $S_s(x) = sx$, $s \in \mathbb{R}$ and $s > 0$.
- (3) Rotation: $Rot_R(x) = Rx$, $R \in \mathbb{R}_{n \times n}$ and R is an orthogonal matrix.
- (4) Reflection about plane $P(a, t)$: $Ref_{a,t}(x) = x - 2(a^T x - t)a$, $a \in \mathbb{R}^n$ is the normal vector of $P(a, t)$, $t \in \mathbb{R}$ is the distance from the origin to $P(a, t)$.
- (5) Inversion about sphere $S^{n-1}(a, r)$: $I_{a,r}(x) = a + \frac{r^2(x - a)}{\|x - a\|_2^2}$, $a \in \mathbb{R}^n$ is the inversion center, r is the inversion radius.

3 Möbius Invariants

3.1 Inversion Invariants

In order to derive the differential invariant under inversion transformation $I_{a,r}$, in the 2-D situation we assume that the $T_{I_{a,r}}$ map the function $f(x, y)$ on domain $D \subset \mathbb{R}^n$ to $g(u, v)$ on domain $D' \subset \mathbb{R}^n$, where $(u, v) = I_{a,r}(x, y)$ and $g(u, v) = f(x, y)$, this means that the coordinates transformations under $I_{a,r}$ are as follows.

$$u = a_x + \frac{r^2(x - a_x)}{(x - a_x)^2 + (y - a_y)^2} \tag{3}$$

$$v = a_y + \frac{r^2(y - a_y)}{(x - a_x)^2 + (y - a_y)^2} \tag{4}$$

At the same time, it means that the coordinates transformations under $I_{a,r}^{-1}$ are as follows.

$$x = a_x + \frac{r^2(u - a_x)}{(u - a_x)^2 + (v - a_y)^2} \tag{5}$$

$$y = a_y + \frac{r^2(v - a_y)}{(u - a_x)^2 + (v - a_y)^2} \tag{6}$$

Based on $g(u, v) = f(x, y)$ and the Eqs. (5) and (6), we obtain the relationships between the partial derivatives of $g(u, v)$ and $f(x, y)$ as follows.

$$g_u = f_x x_u + f_y y_u \tag{7}$$

$$g_v = f_x x_v + f_y y_v \tag{8}$$

$$g_{uu} = (f_{xx}x_u + f_{xy}y_u)x_u + f_x x_{uu} + (f_{yx}x_u + f_{yy}y_u)y_u + f_y y_{uu} \tag{9}$$

$$g_{uv} = (f_{xx}x_v + f_{xy}y_v)x_u + f_x x_{uv} + (f_{yx}x_v + f_{yy}y_v)y_u + f_y y_{uv} \tag{10}$$

$$g_{vv} = (f_{xx}x_v + f_{xy}y_v)x_v + f_x x_{vv} + (f_{yx}x_v + f_{yy}y_v)y_v + f_y y_{vv} \tag{11}$$

Then we obtain a 2-D equation under the inversion transformation, it is

$$\frac{g_{uu} + g_{vv}}{g_u^2 + g_v^2} = \frac{f_{xx} + f_{yy}}{f_x^2 + f_y^2} \tag{12}$$

This means that

$$\frac{f_{xx} + f_{yy}}{f_x^2 + f_y^2} \tag{13}$$

is a differential invariant under inversion transformation. We use the same method in 3-D situation and obtain a differential invariant under the inversion transformation, it is

$$\frac{f_A + f_B}{(f_x^2 + f_y^2 + f_z^2)^2} \tag{14}$$

where

$$\begin{aligned} f_A &= (f_{xx} + f_{yy} + f_{zz})(f_x^2 + f_y^2 + f_z^2) \\ f_B &= f_x^2 f_{xx} + f_y^2 f_{yy} + f_z^2 f_{zz} + 2f_x f_{xy} f_y + 2f_x f_{xz} f_z + 2f_y f_{yz} f_z \end{aligned} \tag{15}$$

3.2 The Boundary of Invariance

We have shown that (13) and (14) are differential invariants under inversion transformation. It is obvious that they are invariants under translation transformation. We prove that (13) and (14) are also differential invariants under rotation, stretching and reflection transformations (see Appendix A¹ for a proof). According to the definition of Möbius transformation, we conclude that the differential expression (13) is a differential invariant under 2-D Möbius transformation. Furthermore, with the Generalized Liouville Theorem we obtain that (14) is a conformal invariant.

3.3 Absoluteness and Relativity of Invariance

If expression Inv_T is an invariant under transformation T , the transformed expression Inv'_T satisfies

$$Inv'_T = W_T \cdot Inv_T \tag{16}$$

where W_T is an expression related to T . In this context, Inv_T is an absolute invariant if $W_T \equiv 1$, otherwise, Inv_T is a relative invariant. Base on the analysis in 3.2, (13) is an absolute invariant under Möbius transformation and (14) is an absolute invariant under conformal transformation. Next, we will show the numerator and denominator of (13) or (14) are relative invariants.

In the derivation of 2-D inversion invariants, we obtain that $W_{I_{a,r}} = ||J||^{-1}$ for the numerator and denominator of (13), this means

$$g_{uu} + g_{vv} = ||J||^{-1}(f_{xx} + f_{yy}) \tag{17}$$

$$g_u^2 + g_v^2 = ||J||^{-1}(f_x^2 + f_y^2) \tag{18}$$

where $|J|$ is the determinant of Jacobian matrix of transformation $I_{a,r}$, $||J||$ is the absolute value of $|J|$. In 3-D situation, we obtain $W_{I_{a,r}} = ||J||^{-\frac{4}{3}}$ for the numerator and denominator of (14). In the stretching transformation, we obtain $W_S = ||J||^{-1}$ in 2-D situation, and $W_S = ||J||^{-\frac{4}{3}}$ in 3-D situation. We also obtain that $W_T = 1$ for the numerator and denominator of (13) or (14) under translation, rotation and reflection transformations.

The result of absoluteness and relativity of invariance on (13) and (14) is shown in Table 1.

3.4 Multiscale and Quantity

Assuming $f(x, y)$ is a regular parameter surface S defined on D , if T_F transform $f(x, y)$ defined on D to $g(u, v)$ defined on D' and $g(u, v) = f(x, y)$, based on the change of variable theorem [16] for multiple integrals and Table 1 we obtain that

$$\iint_{D'} (g_{uu} + g_{vv})dudv = \iint_D W_T(f_{xx} + f_{yy})||J_T||dxdy = \iint_D (f_{xx} + f_{yy})dxdy \tag{19}$$

¹ <https://github.com/duduhe/Differential-and-integral-invariants-under-Mobius-transformation/blob/master/Appendix.pdf>.

Table 1. The form of W_T under transformations

Expression	Translation	Stretching	Rotation	Reflection	Inversion
(13) and (14)	1	1	1	1	1
Num ^a /den of (13)	1	$\ J\ ^{-1}$	1	1	$\ J\ ^{-1}$
Num/den ^b of (14)	1	$\ J\ ^{-\frac{4}{3}}$	1	1	$\ J\ ^{-\frac{4}{3}}$

^aNum means the numerator of fraction.

^bDen means the denominator of fraction.

$$\iint_{D'} (g_u^2 + g_v^2) dudv = \iint_D W_T(f_x^2 + f_y^2) \|J_T\| dx dy = \iint_D (f_x^2 + f_y^2) dx dy \quad (20)$$

where $\|J_T\|$ is the area extension factor, so we obtain that

$$\iint_D (f_{xx} + f_{yy}) dx dy \quad (21)$$

$$\iint_D (f_x^2 + f_y^2) dx dy \quad (22)$$

are integral invariants under 2-D Möbius transformation. In the same way, we obtain that

$$\iiint_D (f_x^2 + f_y^2 + f_z^2)^{\frac{3}{2}} dx dy dz \quad (23)$$

$$\iiint_D (f_A + f_B)^{\frac{3}{4}} dx dy dz \quad (24)$$

are integral invariants under 3-D conformal transformation.

Actually a differential expression Inv_T of function f defined on domain D_f accurately characterize f at point of D_f , it provides extremely wide space to describe the function f .

Multiscale of Invariants. Assuming $F_i(Inv_T)$ is a function of Inv_T , a general method to construct descriptors in different scale is the integral of $\int_{D_j} F_i(Inv_f) dA$ on region $D_j (D_j \subset D_f)$ with different size, and when $D_j = D_f$ the result is a global invariant, for example, the Willmore energy $\int (H^2 - K) dA$ [1] applied in the theory of surfaces [31], digital geometry processing [2] and other fields.

In this view, the only difference between invariant with specify-scale and global invariant is the definition domain, the construction method of specify-scale invariant is same with global invariant. The former could be elaborately modified by selecting domain of integration in different applications.

Quantity of Invariants. A general method to construct a large number of invariants is using various functions $F_i(Inv_T)$ with these functions are independent of each other [3]. We just show a simple method to construct integral invariants based on differential invariants and integral, in addition, more invariant forms can be constructed with differential invariants. Next, we give a possible

form of invariants under Möbius transformation:

$$\iint_D \frac{(f_{xx} + f_{yy})^{n+1}}{(f_x^2 + f_y^2)^n} dx dy \tag{25}$$

$$\iint_D \frac{(f_x^2 + f_y^2)^{n+1}}{(f_{xx} + f_{yy})^n} dx dy \tag{26}$$

$$\iiint_D \frac{(f_A + f_B)^{\frac{3}{4}(n+1)}}{(f_x^2 + f_y^2 + f_z^2)^{\frac{3}{2}n}} dx dy dz \tag{27}$$

$$\iiint_D \frac{(f_x^2 + f_y^2 + f_z^2)^{\frac{3}{2}(n+1)}}{(f_A + f_B)^{\frac{3}{4}n}} dx dy dz \tag{28}$$

if the denominators of (25), (26), (27), (28) are not zero.

3.5 Another Conformal Invariant

The expression $(H^2 - K)dA$ proposed by Biacchke [1] has been proved to be an invariant under Möbius transformation [4,30]. It differs from our method in two important respects: the domain of transformation and the number of functions participated in invariants (see detailed expression at Appendix B).

4 Conjecture of Conformal Invariants

We have shown that (13) is a Möbius invariant and (14) is a conformal invariant. However, the fascinating part of (13) or (14) is that the differential expressions

$$f_x^2 + f_y^2 \quad \text{or} \quad f_x^2 + f_y^2 + f_z^2 \tag{29}$$

$$f_{xx} + f_{yy} \quad \text{or} \quad f_{xx} + f_{yy} + f_{zz} \tag{30}$$

$$f_x^2 f_{xx} + f_y^2 f_{yy} + f_z^2 f_{zz} + 2f_x f_{xy} f_y + 2f_x f_{xz} f_z + 2f_y f_{yz} f_z \tag{31}$$

are differential invariants under rigid transformation. Based on this observation and the fact that the differential expressions play important roles in transformation, we have a bold conjecture about the structure of differential invariants under conformal transformation.

Conjecture: *The differential invariants under conformal transformation are composed of differential invariants under rigid transformation in a self-consistent manner.*

One of the possible self-consistent forms in n-dimensional Euclidean space may be

$$\sum_{i=1}^{n-1} \frac{\prod_{j=1}^{a_i} DRI_j}{(f_{x_1}^2 + f_{x_2}^2 + \dots + f_{x_n}^2)^{n-1}} \tag{32}$$

where DRI is differential invariant under rigid transformation.

5 Experimental Results

We choose a human face model from TOSCA database and treat the z-coordinate value of vertexes of the triangle mesh as a function f defined on x-coordinate and y-coordinate, i.e. $z = f(x, y)$. With least square method, the coordinates of a vertex and its 1-ring neighbors were used to estimate parameters in Taylor expansion of f at the vertex; in order to guarantee the accuracy of descriptor calculation, we only consider vertexes that are located inside the mesh and have enough 1-ring neighbors. After that, we calculate a descriptor at the vertex and the descriptor is composed by (13), (25) and (26) with different $n(\geq 0)$. Moreover, in integral invariants, the area A_{vert} around a vertex is determined by Mixed Voronoi cell.

We deform the definition domain of f with reflection, stretching, rotation and inversion transformation(Fig. 3). In reflection transformation, $a = (1, 0)$ and $t = 0$; the s in stretching transformation is 2; in rotation transformation the original data is rotated 90° counterclockwise; in inversion transformation the inversion center is $(0, 1000)$ and inversion radius is 500 (see more explanation about experiments at Appendix C).

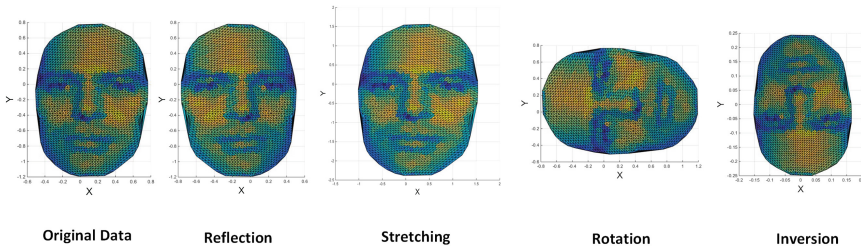


Fig. 3. Elementary transformations of Möbius transformation on human face model.

5.1 Stability of Invariants

In this experiment we choose $n = 0, 1$ and the integral invariants is calculated at the local area of each vertex. After we obtain a 5-dimension descriptor at vertexes of the five mesh in Fig. 3, we calculate the average error of each dimension of the descriptor. In addition, we choose an isometric invariant at the vertex, the Laplacian operator, to compare with above invariants. The average error of each dimension is calculated by the following formula

$$Err = \frac{1}{N} \sum_i \frac{|Inv_{T;i} - Inv_{O;i}|}{|Inv_{T;i}| + |Inv_{O;i}|} \times 100\% \tag{33}$$

where $Inv_{O;i}$ is the value of invariant at vertex i on original data, $Inv_{T;i}$ is the value of invariant at vertex i on deformed data, and N is the total number of vertexes participated in the calculation. The result of this experiment is in Table 2, it shows that (13), (25) and (26) are invariants under Möbius transformations.

Table 2. The average error of Laplacian operator and Möbius invariants.

Expression	Reflection	Stretching	Rotation	Inversion
$f_{xx} + f_{yy}$	0	6.00×10^1	4.82×10^{-13}	8.82×10^1
$\frac{f_{xx}+f_{yy}}{f_x^2+f_y^2}$	0	1.20×10^{-12}	1.33×10^{-12}	1.98×10^{-3}
$\iint_D (f_{xx} + f_{yy}) dx dy$	0	4.38×10^{-13}	4.82×10^{-13}	1.69×10^{-1}
$\iint_D (f_x^2 + f_y^2) dx dy$	0	1.21×10^{-12}	1.27×10^{-12}	1.69×10^{-1}
$\iint_D \frac{(f_{xx}+f_{yy})^2}{f_x^2+f_y^2} dx dy$	0	1.24×10^{-12}	1.47×10^{-12}	1.70×10^{-1}
$\iint_D \frac{(f_x^2+f_y^2)^2}{f_{xx}+f_{yy}} dx dy$	0	2.39×10^{-12}	2.58×10^{-12}	1.70×10^{-1}

5.2 Discrimination of Invariants

In this experiment we use the 5-dimension descriptor of vertex at original to match its corresponding vertex in the deformed mesh with nearest neighbor rule, the metric between vertexes is standardized Euclidean distance. The error rate (percentage) of this experiment is in Table 3.

Table 3. The error rate (percentage) of Möbius invariants in vertex matching.

Reflection	Stretching	Rotation	Inversion
0	0	0	0.87

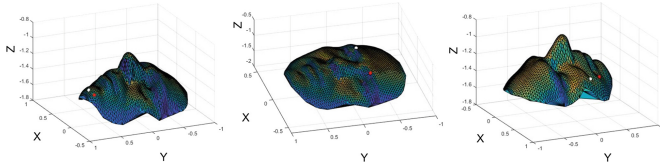


Fig. 4. Some situations where vertex matching fails.

In conformal deformation scenario, this experiment shows the potential of Möbius invariants in matching task. Figure 4 shows some matching-fail situations, where the white point is the real position and the red point is the matching vertex. The reason for most matching failures is that the original white vertex and deformed red vertex have similar functional distribution environments.

6 Conclusions

In this article, we propose two differential invariants under 2-D and 3-D Möbius transformation respectively, in particular, the 3-D expression is a conformal

invariant according to the Liouville Theorem. After that, we analyze the absoluteness and relativity of invariance on the two expressions and their components, and we show an integral construction method that targets to the multiscale and quantity of invariant, the experimental results show that the invariants proposed in this paper perform well. Furthermore, we show another Möbius invariant from the functional view. Finally, we propose a conjecture about the structure of differential invariants under conformal transformation.

This article shows a method of combining functional map and derivatives of function to study conformal invariant, more research about the differential invariants under conformal transformation is necessary in the future. In addition to practical application solutions based on Möbius invariants, questing the generative structure of conformal differential invariant is also an interesting topic.

Acknowledgment. The authors would like to thank Dr. Antti Rrasila of Aalto University for providing help on how to distinguish Möbius invariants and conformal invariants.

This work was partly funded by National Key R&D Program of China (No. 2017YFB1002703) and National Natural Science Foundation of China (Grant No. 60873164, 61227802 and 61379082).

References

1. Biaschke, W.: Vorlesungen über differentialgeometrie iii (1929)
2. Bobenko, A.I., Schröder, P.: Discrete Willmore flow (2005)
3. Brown, A.B.: Functional dependence. *Trans. Am. Math. Soc.* **38**(2), 379–394 (1935)
4. Chen, B.Y.: An invariant of conformal mappings. *Proc. Am. Math. Soc.* **40**(2), 563–564 (1973)
5. Corman, E., Solomon, J., Ben-Chen, M., Guibas, L., Ovsjanikov, M.: Functional characterization of intrinsic and extrinsic geometry. *ACM Trans. Graph. (TOG)* **36**(2), 14 (2017)
6. Crane, K., Pinkall, U., Schröder, P.: Spin transformations of discrete surfaces. *ACM Trans. Graph. (TOG)* **30**(4), 104 (2011)
7. Farkas, H.M., Kra, I.: Riemann surfaces. In: Farkas, H.M., Kra, I. (eds.) *Riemann Surfaces*. GTM, vol. 71, pp. 9–31. Springer, New York (1992). https://doi.org/10.1007/978-1-4612-2034-3_2
8. Gehring, F.W.: Topics in quasiconformal mappings. In: Vuorinen, M. (ed.) *Quasiconformal Space Mappings*. LNM, vol. 1508, pp. 20–38. Springer, Heidelberg (1992). <https://doi.org/10.1007/BFb0094236>
9. Gu, X., Wang, Y., Yau, S.T.: Computing conformal invariants: period matrices. *Commun. Inf. Syst.* **3**(3), 153–170 (2003)
10. Gu, X., Yau, S.T.: Surface classification using conformal structures, p. 701. *IEEE* (2003)
11. Haantjes, J.: Conformal representation of an N-dimensional euclidean space with a non-definite fundamental form on itself (1937)
12. Hu, P.: A class of isometric invariants and their applications (in Chinese). Ph.D. thesis, Institute of Computing Technology, Chinese Academy of Sciences, May 2011
13. Hu, P., Li, H., Lin, Z.: A construction method for surface isometric invariants. *J. Syst. Sci. Math. Sci.* **9**, 006 (2009)

14. Kharevych, L., Springborn, B., Schröder, P.: Discrete conformal mappings via circle patterns. *ACM Trans. Graph. (TOG)* **25**(2), 412–438 (2006)
15. Kühnel, W., Rademacher, H.B.: Liouville's theorem in conformal geometry. *J. de mathématiques pures et appliquées* **88**(3), 251–260 (2007)
16. Lax, P.D.: Change of variables in multiple integrals. *Am. Math. Mon.* **106**(6), 497–501 (1999)
17. Li, E., Huang, Y., Xu, D., Li, H.: Shape DNA: basic generating functions for geometric moment invariants. arXiv preprint [arXiv:1703.02242](https://arxiv.org/abs/1703.02242) (2017)
18. Li, E., Li, H.: Isomorphism between differential and moment invariants under affine transform. arXiv preprint [arXiv:1705.08264](https://arxiv.org/abs/1705.08264) (2017)
19. Li, E., Mo, H., Xu, D., Li, H.: Image projective invariants. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1 (2018). <https://ieeexplore.ieee.org/document/8353142>
20. Liouville, J.: Extension au cas des trois dimensions de la question du tracé géographique. *Applications de l'analyse à la géométrie*, pp. 609–617 (1850)
21. Luo, F.: Combinatorial Yamabe flow on surfaces. *Commun. Contemp. Math.* **6**(05), 765–780 (2004)
22. Mullen, P., Tong, Y., Alliez, P., Desbrun, M.: Spectral conformal parameterization. *Comput. Graph. Forum* **27**(5), 1487–1494 (2008)
23. Olver, P.J.: *Equivalence, Invariants and Symmetry*. Cambridge University Press, Cambridge (1995)
24. Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., Guibas, L.: Functional maps: a flexible representation of maps between shapes. *ACM Trans. Graph. (TOG)* **31**(4), 30 (2012)
25. Rasila, A.: Introduction to quasiconformal mappings in n-space. In: *Proceedings of the International Workshop on Quasiconformal* (2006)
26. Rustamov, R.M., Ovsjanikov, M., Azencot, O., Ben-Chen, M., Chazal, F., Guibas, L.: Map-based exploration of intrinsic shape differences and variability. *ACM Trans. Graph. (TOG)* **32**(4), 72 (2013)
27. Springborn, B., Schröder, P., Pinkall, U.: Conformal equivalence of triangle meshes. *ACM Trans. Graph. (TOG)* **27**(3), 77 (2008)
28. Vaxman, A., Müller, C., Weber, O.: Conformal mesh deformations with möbius transformations. *ACM Trans. Graph. (TOG)* **34**(4), 55 (2015)
29. Wang, Y., Wang, X., Zhang, B.: Affine differential invariants of functions on the plane. *J. Appl. Math.* **2013** (2013). <https://www.hindawi.com/journals/jam/2013/868725/cta/>
30. White, J.H.: A global invariant of conformal mappings in space. *Proc. Am. Math. Soc.* **38**(1), 162–164 (1973)
31. Willmore, T.J.: Surfaces in conformal geometry. *Ann. Glob. Anal. Geom.* **18**(3–4), 255–264 (2000)
32. Xu, D., Li, H.: Geometric moment invariants. *Pattern Recogn.* **41**(1), 240–249 (2008)
33. Xu, J., Kang, H., Chen, F.: Content-aware image resizing using quasi-conformal mapping. *Vis. Comput.* **34**(3), 431–442 (2018)
34. Yu, X., Lei, N., Wang, Y., Gu, X.: Intrinsic 3D dynamic surface tracking based on dynamic Ricci flow and teichmüller map. In: *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017, pp. 5400–5408 (2017)
35. Zeng, W., Gu, X.D.: Registration for 3D surfaces with large deformations using quasi-conformal curvature flow. In: *Computer Vision and Pattern Recognition (CVPR)* (2011)
36. Zhang, H., Mo, H., Hao, Y., Li, S., Li, H.: Fast and efficient calculations of structural invariants of chirality. arXiv preprint [arXiv:1711.05866](https://arxiv.org/abs/1711.05866) (2017)



Automatic Classifier Selection Based on Classification Complexity

Liping Deng¹, Wen-Sheng Chen^{1,2}, and Binbin Pan^{1,2}(✉)

¹ College of Mathematics and Statistics, Shenzhen University, Shenzhen 518060, People's Republic of China
{chenws, pbb}@szu.edu.cn

² Guangdong Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, People's Republic of China

Abstract. Choosing a proper classifier for one specific data set is important in practical application. Automatic classifier selection (CS) aims to recommend the most suitable classifiers to a new data set based on the similarity with the historical data sets. The key step of CS is the extraction of data set feature. This paper proposes a novel data set feature that characterizes the classification complexity of problems, which has a close connection with the performance of classifiers. We highlight two contributions of our work: firstly, our feature can be computed in a low time complexity; secondly, we theoretically show that our feature has connection with generalization errors of some classifiers. Empirical results indicate that our feature is more effective and efficient than the existing data set features.

Keywords: Automatic classifier selection · Data set feature
Data set similarity

1 Introduction

Classification is one of the most important tasks in machine learning. A great number of classifiers were putted forward in recent decades to tackle various kinds of classification problems arose in real world, such as support vector machine, decision tree, AdaBoost, artificial neural networks, and so on. Does there exist a classifier that significantly performs better than any other classifiers on most of data sets? Some literatures have done in-depth investigations on this problem. The No Free Lunch Theorem [1] tells us that there does not exist such classifier. If classifier \mathcal{A}_1 outperforms \mathcal{A}_2 on some data sets, then there must exist as many other data sets on which \mathcal{A}_2 outperforms \mathcal{A}_1 . In [2], authors analyzed the performances of three classifiers on some data sets and they did not observe which classifier is significantly better than the others. Furthermore, [3] conducted classification experiments using 179 classifiers and 121 data sets and showed that there is no optimal classifier. These results indicate that classifiers have preference on different types of data sets. Therefore, which classifier(s) would be selected for a given classification problem?

One idea is to use cross validation for all possible classifiers to find the best classifier. However, this procedure is time-consuming. An efficient alternative approach is automatic classifier selection based on data set similarity [4–7, 10], or classifier selection (CS) for short. We believe that the performances of classifiers on similar data sets should be close. Since different data sets may vary in sample size, dimensions, classes and attributes, how to measure the similarity between data sets is a critical step of CS. The common method is to extract data set feature by designing a feature extraction function (or called meta-learning) and then compute the similarity between these features. There is an intrinsic relationship between classifier performance and data set feature [9]. Therefore, the recommendation heavily depends on the effectiveness of data set feature. Furthermore, the feature should be calculated in a low time complexity, which is a bottleneck of CS.

A number of data set features have proposed. These features are extracted from different aspects of a data set: (i) statistics and information theory (SI) [7, 10]; (ii) model structure (MS) [5]; (iii) problem complexity (PC) [4]; (iv) landmarking (LM) [6]. Especially, PC and LM characterize the classification complexity of problems (we call it *complexity*) using a set of geometrical metrics or basic classifiers. The complexity is expected to highly correlate to the performances of classifiers [11]. In other words, the performances of classifiers on data sets that have similar complexity should be close. Therefore, complexity plays a vital role in CS. However, the data set features extracted by PC and LM have two shortages: (i) time-consuming; (ii) no theoretical connection with performances of classifiers. It is observed that PC and LM did not perform well in some literatures [5, 7], which means that they cannot characterize the complexity accurately.

To remedy the aforementioned shortcomings of PC and LM, this paper uses a set of geometrical and statistical metrics to describe the complexity of two-class data set, then these metrics are united as data set feature. We use KNN classifier as recommendation algorithm for CS. For multi-class classification problem, we split the problem into two-class problems using one-vs-one strategy. Compared with PC and LM, our work has improvements in two aspects: computation efficiency and theoretical guarantee. Empirical results demonstrate the effectiveness and efficiency of our method.

The rest of the paper is structured as follows. We briefly introduce the related works in Sect. 2. Section 3 presents our data set feature. The classifier selection algorithm is given in Sect. 4. Empirical investigations are discussed in Sect. 5 and conclusions are drew in Sect. 6.

2 Related Work

The key problem of CS is feature extraction. To the best of our knowledge, there are four kinds of features.

Statistical Feature: This feature can be categorized into two kinds. The first kind describes the data set using a group of statistical and information theory

characteristics [10]. The second kind is based on summary statistics. Song [7] characterizes the data set structure by computing the frequencies of itemsets generated from binary data sets. Non-binary data set needs to be transformed to binary data set, which would be time-consuming when the attributes of data set are continuous.

Problem Complexity Feature: Twelve measures are designed to describe the geometrical complexity of decision boundary of two-class problems [11]. Cano [12] claimed that some of the measures have little connection with the performances of classifiers. Bernado [4] selected six measures to characterize data set.

Landmarking Feature: This feature [6] utilizes the performances of a set of basic classifiers (called *landmarkers*) to describe the data set. Therefore, the similar features indicate that data sets may belong to the subspace of the same performance. The chosen landmarkers must be significantly different.

Model Structure Feature: The statistical information of a model generated from data set is collected as feature. In this category, decision tree is usually considered [5], from which we gather a set of statistics like maximum/minimum number of nodes, length of longest/shortest branches, and so on.

The aforementioned features belong to experimental origin. However, a theoretical investigation would be more persuasive. Furthermore, these features are computationally expensive.

3 Proposed Feature

In this section, we firstly propose several metrics of complexity for CS. Then the theoretical connections between two metrics and generalization errors of some classifiers are investigated. Finally, we present our data set feature and similarity measurement criterion.

3.1 Metrics of Complexity

Given a two-class data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ in input space \mathcal{X} , where \mathbf{x}_i , $i = 1, 2, \dots, n$ are data points, and y_i is the binary class label, i.e., $y_i \in \{1, -1\}$. Let $\mathbf{y} = [y_1, y_2, \dots, y_n]^\top$ represents the vector formed with n labels. We use n_- and n_+ to represent the amount of samples labeled -1 or 1 , respectively. Note that $n_- + n_+ = n$.

For a given kernel function $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$, where ϕ is a nonlinear mapping that maps $\mathbf{x} \in \mathcal{X}$ to a reproduce kernel hilbert space (RKHS) \mathcal{H} , an $n \times n$ kernel matrix \mathbf{K} is generated from \mathcal{D} as

$$K_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle, i, j = 1, 2, \dots, n.$$

\mathbf{K} is a symmetric positive and semi-definite matrix that totally preserves the geometrical structure of \mathcal{D} . Our five metrics of complexity are based on \mathbf{K} .

Kernel Alignment. This metric, which is known as centered kernel target alignment (KA) [13], is defined as

$$KA(\mathbf{K}_c, \mathbf{y}\mathbf{y}^\top) = \frac{\langle \mathbf{K}_c, \mathbf{y}\mathbf{y}^\top \rangle_F}{\sqrt{\langle \mathbf{K}_c, \mathbf{K}_c \rangle_F \langle \mathbf{y}\mathbf{y}^\top, \mathbf{y}\mathbf{y}^\top \rangle_F}}, \tag{1}$$

where \mathbf{K}_c is a centralized kernel matrix of \mathbf{K} , $\langle \cdot, \cdot \rangle_F$ denotes Frobenius inner-product and $\mathbf{y}\mathbf{y}^\top$ is called the target matrix. $KA \in [0, 1]$ since $\langle \mathbf{K}_c, \mathbf{y}\mathbf{y}^\top \rangle_F \geq 0$.

The numerator of (1) can be expanded as

$$\begin{aligned} \langle \mathbf{K}_c, \mathbf{y}\mathbf{y}^\top \rangle_F &= \mathbf{y}^\top \mathbf{K}_c \mathbf{y} = \sum_{i=1}^n \sum_{j=1}^n y_i y_j (K_c)_{ij} \\ &= \sum_{y_i=y_j} (K_c)_{ij} - \sum_{y_i \neq y_j} (K_c)_{ij}. \end{aligned}$$

Therefore, KA measures the difference between the within-class and between-class distances of data set. A bigger KA indicates that the corresponding data set is more separable. The most time-consuming calculations of KA are the centralization of \mathbf{K} and $\langle \mathbf{K}_c, \mathbf{K}_c \rangle_F$, which take $O(n^2)$ time complexity.

Kernel Space-Based Separability. The centers of two classes in \mathcal{H} are calculated as

$$\begin{aligned} \phi_- &= \frac{1}{n_-} \sum_{y_i=-1} \phi(\mathbf{x}_i), \\ \phi_+ &= \frac{1}{n_+} \sum_{y_i=1} \phi(\mathbf{x}_i), \end{aligned} \tag{2}$$

respectively. KS [14] is defined as

$$KS(\mathbf{K}, \mathbf{y}) = \frac{std_- + std_+}{\|\phi_- - \phi_+\|_2}, \tag{3}$$

where

$$\begin{aligned} std_- &= \sqrt{\frac{\sum_{y_i=-1} \langle \phi(\mathbf{x}_i) - \phi_-, \mathbf{e} \rangle^2}{n_- - 1}}, \\ std_+ &= \sqrt{\frac{\sum_{y_i=1} \langle \phi(\mathbf{x}_i) - \phi_+, \mathbf{e} \rangle^2}{n_+ - 1}}, \end{aligned} \tag{4}$$

are the standard deviations of two classes projected along the direction $\mathbf{e} = \frac{\phi_- - \phi_+}{\|\phi_- - \phi_+\|_2}$ respectively, and $\|\cdot\|_2$ denotes 2-norm of vector.

$KS \in (0, +\infty]$ actually describes the samples' distribution along direction $\phi_- - \phi_+$. A smaller KS means that the data set is more separable. KS needs $O(n^2)$ time complexity.

Overlap Region. We propose a metric that compute the ratio of the overlapped region of two classes to the total region of two classes along direction \mathbf{e} , denoted as ROR. Suppose that the projected data of one class fall into $[a_1, b_1]$, where a_1, b_1 are the minimum and maximum values of the projected data, and the other class falls into $[a_2, b_2]$. Let $U = [a_1, b_1] \cap [a_2, b_2]$ and $V = [a_1, b_1] \cup [a_2, b_2]$ be intersection and union of these two intervals, respectively. ROR is defined as

$$\text{ROR} = \begin{cases} 0, & U = \emptyset, \\ \frac{\max(U) - \min(U)}{\max(V) - \min(V)}, & U \neq \emptyset, \end{cases} \tag{5}$$

where $\min(\cdot)$ and $\max(\cdot)$ are the maximum and minimum values of interval respectively and \emptyset represents empty set. $\text{ROR} \in [0, 1]$ since U is a subset of V . When data set is linear separable, ROR is expected to zero. However, ROR will increase if data set is nonlinear separable. ROR also needs $O(n^2)$ time complexity.

Test of Equality of Means. Now we treat kernel matrix \mathbf{K} as a similarity matrix. The following measure depends on the assumption that the similarity among within-class data is higher than between-class data. We first introduce two vectors extracted from \mathbf{K} :

$$\begin{aligned} \mathbf{k}_W &= \{K_{ij} | i < j \wedge y_i = y_j\}, \\ \mathbf{k}_B &= \{K_{ij} | i < j \wedge y_i \neq y_j\}. \end{aligned} \tag{6}$$

We denote $n_W = \frac{n-(n-1)}{2} + \frac{n+(n-1)}{2}$ and $n_B = n - n_+$ represent the size of vectors \mathbf{k}_W and \mathbf{k}_B respectively. We see that \mathbf{k}_W is the collection of within-class similarity and \mathbf{k}_B is the collection of between-class similarity.

TEM [15] is defined as a variant of t-test to evaluate the equality of means of \mathbf{k}_W and \mathbf{k}_B :

$$\text{TEM}(\mathbf{K}, \mathbf{y}) = \frac{1}{n} \left| \frac{\bar{k}_W - \bar{k}_B}{\sqrt{\frac{\sigma_W^2}{n_W} + \frac{\sigma_B^2}{n_B}}} \right|, \tag{7}$$

where \bar{k}_W and σ_W^2 denote the mean and variance of \mathbf{k}_W respectively, and \bar{k}_B and σ_B^2 denote the mean and variance of \mathbf{k}_B respectively. TEM is very sensitive to the nonlinearity of decision boundary. A larger TEM reflects that the data set is more likely to be linearly separable. Here we normalized TEM by multiplying the reciprocal of n to eliminate the influence of sample size. TEM only utilizes the upper triangle elements of \mathbf{K} , which needs $O(n^2)$ time complexity.

Test of Equality of Variances. Let $\mathbf{k}_{WB} = \mathbf{k}_W \cup \mathbf{k}_B$ be the union of \mathbf{k}_W and \mathbf{k}_B . We define three new vectors as follows:

$$\begin{aligned} \mathbf{z}_W &= |\mathbf{k}_W - \tilde{\mathbf{k}}_W|, \\ \mathbf{z}_B &= |\mathbf{k}_B - \tilde{\mathbf{k}}_B|, \\ \mathbf{z}_{WB} &= |\mathbf{k}_{WB} - \tilde{\mathbf{k}}_{WB}|, \end{aligned} \tag{8}$$

where $|\cdot|$ represents element-wise absolute value, $\tilde{\mathbf{k}}_W$, $\tilde{\mathbf{k}}_B$ and $\tilde{\mathbf{k}}_{WB}$ are the medians of \mathbf{k}_W , \mathbf{k}_B and \mathbf{k}_{WB} respectively. TEV [15] is defined using Brown-Forsythe test to measure the equality of variances of \mathbf{k}_W and \mathbf{k}_B ,

$$\text{TEV}(\mathbf{K}, \mathbf{y}) = \left(1 - \frac{2}{n}\right) \frac{n_W(\bar{z}_W - \bar{z}_{WB})^2 + n_B(\bar{z}_B - \bar{z}_{WB})^2}{\sum_{i=1}^{n_W} [(z_W)_i - \bar{z}_W]^2 + \sum_{i=1}^{n_B} [(z_B)_i - \bar{z}_B]^2}, \quad (9)$$

where \bar{z}_B , \bar{z}_W and \bar{z}_{WB} are the mean values of vectors \mathbf{z}_B , \mathbf{z}_W and \mathbf{z}_{WB} respectively, $(z_W)_i$ and $(z_B)_i$ represent the i^{th} element of \mathbf{z}_W and \mathbf{z}_B . The idea behind TEV is that if \mathbf{k}_W and \mathbf{k}_B have the same variance, then the data set should be difficult to separate. The high value of TEV rejects the hypothesis of equal variance and indicates compact within-class and mutually distant between-class distribution [15]. Here we also normalize TEV by multiplying $1/n$.

Like TEM, TEV also needs $O(n^2)$ time complexity, but TEV needs extra $O(n^2)$ to search the medians.

3.2 Theoretical Analysis

We theoretically investigate the relationship between metrics KA, KS and generalization errors.

Theorem 1. KA is defined as (1). Let $R(h) = \text{Pr}[yh < 0]$ be the error rate of Parzen window predictor

$$h(\mathbf{x}') = \frac{E_{\mathbf{x}}[y k_c(\mathbf{x}, \mathbf{x}')]}{\sqrt{E[k_c^2]}} \quad (10)$$

in binary classification. k_c is the centered kernel function and $E[\cdot]$ is an expectation operator. Suppose that $k(\mathbf{x}, \mathbf{x}) \leq S^2$ for all \mathbf{x} . Then for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:

$$R(h) \leq 1 - \left(\text{KA}(\mathbf{K}_c, \mathbf{y}\mathbf{y}^\top) - 18\beta \left[\frac{3}{n} + 4\sqrt{\frac{\log \frac{6}{\delta}}{2n}} \right] \right) \cdot \frac{1}{\Gamma}, \quad (11)$$

where $\Gamma = \max_{\mathbf{x}'} \sqrt{\frac{E_{\mathbf{x}}[k_c^2(\mathbf{x}', \mathbf{x})]}{E_{\mathbf{x}, \mathbf{x}'}[k_c^2(\mathbf{x}', \mathbf{x})]}}$, $\beta = \max\left(\frac{S^2}{E[k_c^2]}, \frac{S^2}{E[k_c'^2]}\right)$ and $k'(\mathbf{x}_i, \mathbf{x}_j) = y_i y_j$.

Proof. According to Theorem 12 in [13], we have

$$\text{KA}(k_c, k_c') \geq \text{KA}(\mathbf{K}_c, \mathbf{y}\mathbf{y}^\top) - 18\beta \left[\frac{3}{n} + 4\sqrt{\frac{\log \frac{6}{\delta}}{2n}} \right],$$

where $\text{KA}(k_c, k_c') = \frac{E[k_c k_c']}{\sqrt{E[k_c^2]E[k_c'^2]}}$. Unifying Theorem 13 in [13]

$$R(h) \leq 1 - \text{KA}(k_c, k_c') \cdot \frac{1}{\Gamma},$$

We obtain the inequation (11) directly.

Theorem 2. [14] *KS is defined as (3). There is a separating hyperplane*

$$h(\mathbf{x}) = \mathbf{e} \cdot \phi(\mathbf{x}) - \mathbf{e} \cdot \frac{std_- \phi_+ + std_+ \phi_-}{std_- + std_+}, \quad (12)$$

such that the upper bound of training error of data set \mathcal{D} is

$$KSerr = \frac{KS(\mathbf{K}, \mathbf{y})^2}{1 + KS(\mathbf{K}, \mathbf{y})^2}. \quad (13)$$

Theorem 1 tells us that if there is a high KA and Γ is not too large, then the upper bound of generalization error of (10) on \mathcal{D} is small. Theorem 2 indicates if KS is small, then the upper bound of training error of (12) on \mathcal{D} is small, thus we can expect a low generalization error [14].

3.3 Data Set Feature

Based on the above analysis, we define data set feature as follows:

$$\mathbf{v} = [KA, 1 - KSerr, 1 - ROR, TEM, TEV]. \quad (14)$$

The computation of \mathbf{v} has a time complexity of $O(n^2)$. KA, KS and ROR mainly focus on the distributions and the degree of overlap of two classes from a geometrical point of view, while statistical tests (TEM, TEV) are used to characterize the nonlinearity of decision boundary. Employing different kernel functions would produce different features. We adopt Euclidean distance as similarity criterion:

$$\rho(\mathcal{D}, \mathcal{D}') := \|\mathbf{v} - \mathbf{v}'\|_2 = \sqrt{\sum_{i=1}^5 (v_i - v'_i)^2}. \quad (15)$$

The smaller $\rho(\mathcal{D}, \mathcal{D}')$ means that the similarity between data sets \mathcal{D} and \mathcal{D}' is higher.

4 Classifier Selection

Suppose that historical data sets $\mathcal{D}_1, \dots, \mathcal{D}_m$ and testing data set \mathcal{D} are two-class problems. Our CS algorithm is shown in Algorithm 1.

4.1 Recommendation Algorithm

In step 2 of Algorithm 1, we use KNN classifier as \mathcal{A}_R , where the data set similarity is the distance between data set features. Assuming $\mathcal{D}_j, j = 1, 2, \dots, K$ are the K most similar data sets for \mathcal{D} , the recommended classifier is selected as: (i) for each \mathcal{D}_j , we assign a rank to candidate classifiers according to its performances on this problem. The classifier with the best performance has rank 1, while the classifier with the worst performance has rank m . Classifiers with

Algorithm 1. CS for Two-class Problems

Input: historical data sets $\mathcal{D}_1, \dots, \mathcal{D}_m$, candidate classifiers $\mathcal{A}_1, \dots, \mathcal{A}_\ell$, testing data set \mathcal{D}

Output: classifier \mathcal{A}^*

- 1: Evaluate the performances of candidate classifiers on historical data sets using 10-fold cross validation.
 - 2: Design a recommendation algorithm \mathcal{A}_R based on similarity and the performances.
 - 3: Extract the data set features $\mathbf{v}_1, \dots, \mathbf{v}_m$ and \mathbf{v} as (14).
 - 4: Compute the data set similarities using (15).
 - 5: Output a best classifier \mathcal{A}^* for \mathcal{D} using \mathcal{A}_R .
-

the same performance have the same average rank; (ii) let $R_{i,j}, i = 1, 2, \dots, \ell$ denote the rank of classifier \mathcal{A}_i on \mathcal{D}_j , then the rank of classification algorithm \mathcal{A}_i on \mathcal{D} is computed as

$$R_{i,\mathcal{D}} = \frac{1}{K} \sum_{\mathcal{D}_j \in N_c(\mathcal{D})} R_{i,j}, j = 1, 2, \dots, K, \quad (16)$$

where $N_c(\mathcal{D})$ is a set contains the K most similar data sets of \mathcal{D} . In the end, the classifier with the lowest rank is the recommended classifier.

4.2 Multi-class Classification Problem

Our feature only suitable for two-class data sets. We handle multi-class problems as follow.

Step 1: Suppose that data set \mathcal{D} has c classes. We split \mathcal{D} into $m = \frac{c(c-1)}{2}$ two-class problems using one-vs-one strategy.

Step 2: For each sub-problem, we recommend one classifier based on Algorithm 1.

Step 3: The final decision is determined by using voting strategy.

The merit of this method is that we can select the most suitable classifier for each sub-problem, which would make the classification accuracy higher than that of the single classifier.

5 Experiments

We evaluate the proposed feature with three state-of-the-art features with respect to computational efficiency and recommendation performance.

5.1 Experimental Setup

Data Sets. We selected 67 classification problems from the UCI repository which include 49 historical data sets and 18 testing data sets (Table 1). Among

Table 1. Summary of testing data sets in terms of attributes, sample size and classes.

ID	Name	Att.	Ins.	Classes	ID	Name	Att.	Ins.	Classes
1	abalone	8	4117	3	10	page-blocks	10	5473	5
2	car	6	1728	4	11	seeds	7	210	3
3	contrac	9	1473	3	12	segment	18	2310	7
4	dermatology	34	366	6	13	st-landsat	36	6534	6
5	hayes-roth	5	132	3	14	st-vehicle	18	846	4
6	hill-valley	100	1212	2	15	synthetic-control	60	600	6
7	hill-valley-noise	100	1212	2	16	teaching	5	151	3
8	iris	4	154	3	17	waveform	21	2000	3
9	nursery	8	12598	4	18	wine	13	178	3

the historical data sets, the multi-class data sets are split into two-class data sets using one-vs-one technique, then those data sets that are easy to classify or have severely unbalanced/small samples in each class are deleted. We totally have 84 two-class historical data sets. The attributes of data sets are normalized into $[-1, 1]$.

Candidate Classifiers. We employ 20 candidate classifiers. Some candidate classifiers are KNN, LDA, logistics regression, SVM (linear, polynomial kernel, RBF kernel), naive bayes, decision tree C4.5, random forest, Bagging (tree) and AdaBoost (tree). These classifiers are run with the MATLAB statistic toolbox except SVM uses LIBSVM software.

The remaining classifiers are nearest mean classifier, Fisher’s least square linear discriminant, BP neural network, linear perceptron, Bayesian classifier, Gaussians mixture model, Parzen classifier, Parzen density classifier and radial basis neural network classifier, which are adopted from PrTools toolbox 5.0. We run all codes on MATLAB 2017a on Windows operating system with Inter(R) Core(TM) i5-6500 CPU @3.20GHz processor.

Comparative Classifiers. We evaluate 24 classifiers on testing data sets which include 20 candidate classifiers and 4 data set features.

- statistical feature (F_s) [7];
- problem complexity feature (F_p) [4];
- landmarking feature (F_l) [6] with landmarks KNN, C4.5, LR and NB;
- our data set feature using polynomial kernel (F_{poly}). We set $d = 3$.

The attributes of 4 data set features are normalized into $[0, 1]$. F_s , F_p and F_l adopt the CS framework in Algorithm 1. For each testing data set, 10% samples of each class are dropped as testing samples and the rests are used for training (the testing data set in Algorithm 1). The classification model of recommended classifier on training samples are trained using 10-fold cross validation. For the

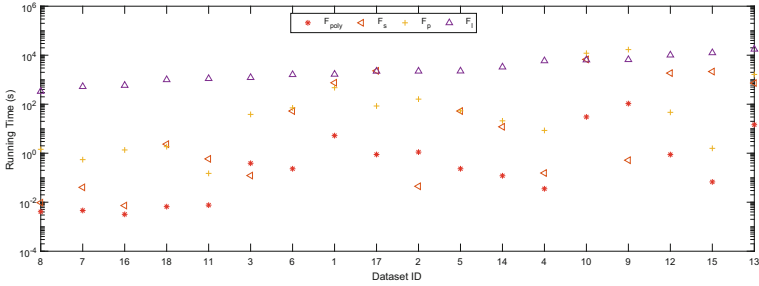


Fig. 1. Running times (s) of F_{poly} , F_s , F_p and F_l on testing data sets. The total times are 160.11s, 14662.85s, 31602.10s and 77109.74s, respectively.

sake of fairness, we also evaluate the performance of candidate classifiers on multi-class testing data sets using splitting and voting strategy.

Performance Metrics. We employ classification accuracy (CA), average recommendation performance ratio (ARPR) [8] and non-parameter statistical tests [16] to evaluate the performance of data set features.

5.2 Computational Efficiency

We collected the computation times of 4 data set features on 18 testing data sets (Fig. 1). The recorded time of each data set is the sum of times of its sub-problems. From Fig. 1, we see that our feature has the fastest computational speed, which spent 160 seconds on overall data sets. However, F_s , F_p and F_l have unacceptable low speeds. Although F_s outperformed our features on data sets 2, 3, and 9, we found that these data sets have discrete variables. For continuous variables, the efficiency of F_s would be degraded rapidly. Therefore, our feature outperforms F_s , F_p and F_l in terms of efficiency.

5.3 Performance Comparisons

In this section, we compare our F_{poly} with three state-of-the-art data set features: F_s , F_p and F_l , as well as 20 candidate classifiers. The comparisons of CA, ARPR and statistical test are listed in Table 2. We observe that F_{poly} has the highest CA and ARPR.

To check the statistical difference between different methods, we calculated the average rank of each feature and shown it in the last row of Table 2. F_{poly} has the lowest average rank 1.36, followed by F_s . F_p has the worst average rank. The Friedman statistic is distributed according to the F-distribution with $(4-1) = 3$ and $(4-1) \times (18-1) = 51$ degrees of freedom. The value of Friedman statistic is 11.64 and the critical value of $F(3, 51)$ is 2.79 at 0.05 significance level. Thus, the null hypothesis is rejected. Then we applied the Nemenyi test for pairwise

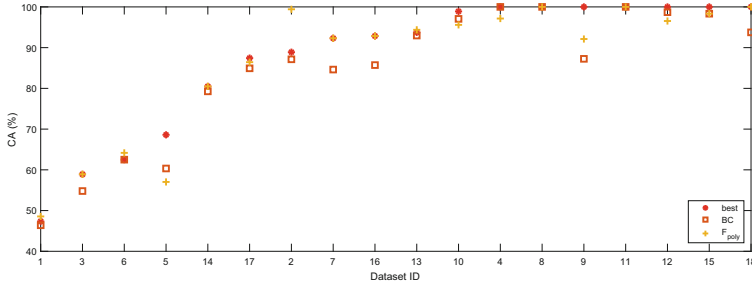


Fig. 2. CA (%) of *best*, BC and F_{poly} . *best* represents the CA of the best candidate classifier.

comparisons. The critical different is 1.11 which means that F_{poly} is significantly better than F_p and F_l .

Finally, we compare the CA of F_{poly} with that of the best candidate classifier and Bayesian classifier (BC) which has the highest ACA among 20 classifiers, shown in Fig. 2. We see that the CA of F_{poly} are very close to the CA of the best candidate classifier except on data sets 7 and 18. F_{poly} is equal to or higher than the best candidate classifier on 11 data sets. F_{poly} has the same CA as or outperforms BC in 14 out of 18 cases. On the 4 data sets that BC outperforms F_{poly} , we see that the CA of BC and F_{poly} are very close.

6 Conclusion

The difficulties of CS mainly stem from the similarity measurement among data sets. So far, people resolve this problem by characterizing data set feature and turn to comparing the similarity of features. In this paper, we proposed a new data set feature to describe the classification complexity of data set. Different

Table 2. CA (%) of F_{poly} , F_s , F_p and F_l on testing data sets. The first column shows the ID of data sets. The last row reports the average rank of each CS algorithm. \mathcal{A}_{best} and \mathcal{A}_{worst} indicate the best and worst CA of candidate classifiers.

ID	F_{poly}	F_s	F_p	F_l	\mathcal{A}_{best}	\mathcal{A}_{worst}	ID	F_{poly}	F_s	F_p	F_l	\mathcal{A}_{best}	\mathcal{A}_{worst}
1	48.56	46.63	41.59	43.99	47.36	44.47	12	96.54	95.67	96.10	95.67	100	98.27
2	99.42	74.85	76.02	91.81	88.89	77.19	13	94.38	93.59	91.72	90.63	93.75	90.47
3	58.90	58.22	50.00	54.79	58.90	53.42	14	80.49	74.39	75.61	74.39	80.49	74.39
4	97.14	94.29	88.57	94.29	100	100	15	98.33	98.33	98.33	95.00	100	98.33
5	57.02	57.02	52.07	52.07	68.60	52.89	16	92.86	78.57	85.71	71.43	92.86	78.57
6	64.17	64.17	50.83	55.83	66.94	48.33	17	86.43	81.41	83.92	84.42	87.44	78.89
7	92.31	92.31	76.92	84.62	92.31	69.23	18	100	100	93.75	100	100	93.75
8	100	100	100	100	100	100							
9	92.12	80.22	78.36	79.29	100	86.01	ACA	86.35	82.44	79.54	81.23		
10	95.59	94.30	92.28	93.93	98.90	95.96	ARPR	0.99	0.95	0.90	0.93		
11	100	100	100	100	100	85.71	Rank	1.36	2.44	3.19	3.00		

from previous works, our feature has merits like low computational complexity and theoretical support. We built a CS framework using the proposed feature. Experimental results show that our feature is effective and efficient. Our method outperforms three data set features, which means that the proposed feature can help to choose suitable classifiers for new classification problems.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant 61602308.

References

1. Wolpert, D.H.: The lack of a priori distinction between learning algorithms. *Neural Comput.* **8**(7), 1341–1390 (1996)
2. Maciá, N., Bernadó-Mansilla, E., Orriols-Puig, A., Kam, H.T.: Learner excellence biased by data set selection: a case for data characterisation and artificial data sets. *Pattern Recogn.* **46**(3), 1054–1066 (2013)
3. Cernadas, E., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**(1), 3133–3181 (2014)
4. Bernadó-Mansilla, E., Ho, T.K.: Domain of competence of XCS classifier system in complexity measurement space. *IEEE Trans. Evol. Comput.* **9**(1), 82–104 (2008)
5. Peng, Y., Flach, P.A., Soares, C., Brazdil, P.: Improved dataset characterisation for meta-learning. In: Lange, S., Satoh, K., Smith, C.H. (eds.) *DS 2002. LNCS*, vol. 2534, pp. 141–152. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-36182-0_14
6. Pfahringer, B., Bensusan, H., Giraud-Carrier, C.G.: Meta-learning by landmarking various learning algorithms. In: *Seventeenth International Conference on Machine Learning*, vol. 11, no. 9, pp. 743–750. Morgan Kaufmann Publishers Inc. (2000)
7. Song, Q., Wang, G., Wang, C.: Automatic recommendation of classification algorithms based on data set characteristics. *Pattern Recogn.* **45**(7), 2672–2689 (2012)
8. Wang, G., Song, Q., Zhu, X.: An improved data characterization method and its application in classification algorithm recommendation. *Appl. Intell.* **43**(4), 892–912 (2015)
9. Kotthoff, L.: Algorithm selection for combinatorial search problems: a survey. In: Bessiere, C., De Raedt, L., Kotthoff, L., Nijssen, S., O’Sullivan, B., Pedreschi, D. (eds.) *Data Mining and Constraint Programming. LNCS (LNAI)*, vol. 10101, pp. 149–190. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50137-6_7
10. Kalousis, A., Theoharis, T.: NOEMON: design, implementation and performance results of an intelligent assistant for classifier selection. *Intell. Data Anal.* **3**(5), 319–337 (1999)
11. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(3), 289–300 (2002)
12. Cano, J.R.: Analysis of data complexity measures for classification. *Expert Syst. Appl.* **40**(12), 4820–4831 (2013)
13. Cortes, C., Mohri, M., Rostamizadeh, A.: Algorithms for learning kernels based on centered alignment. *J. Mach. Learn. Res.* **13**(2), 795–828 (2012)
14. Nguyen, C.H., Tu, B.H.: An efficient kernel matrix evaluation measure. *Pattern Recogn.* **41**(11), 3366–3372 (2008)
15. Chudzian, P.: Evaluation measures for kernel optimization. *Pattern Recogn. Lett.* **33**(9), 1108–1116 (2012)
16. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**(1), 1–30 (2006)



Gradient-Based Representational Similarity Analysis with Searchlight for Analyzing fMRI Data

Xiaoliang Sheng, Muhammad Yousefnezhad, Tonglin Xu,
Ning Yuan, and Daoqiang Zhang^(✉)

College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
dqzhang@nuaa.edu.cn

Abstract. Representational Similarity Analysis (RSA) aims to explore similarities between neural activities of different stimuli. Classical RSA techniques employ the inverse of the covariance matrix to explore a linear model between the neural activities and task events. However, calculating the inverse of a large-scale covariance matrix is time-consuming and can reduce the stability and robustness of the final analysis. Notably, it becomes severe when the number of samples is too large. For facing this shortcoming, this paper proposes a novel RSA method called gradient-based RSA (GRSA). Moreover, the proposed method is not restricted to a linear model. In fact, there is a growing interest in finding more effective ways of using multi-subject and whole-brain fMRI data. Searchlight technique can extend RSA from the localized brain regions to the whole-brain regions with smaller memory footprint in each process. Based on Searchlight, we propose a new method called Spatiotemporal Searchlight GRSA (SSL-GRSA) that generalizes our ROI-based GRSA algorithm to the whole-brain data. Further, our approach can handle some computational challenges while dealing with large-scale, multi-subject fMRI data. Experimental studies on multi-subject datasets confirm that both proposed approaches achieve superior performance to other state-of-the-art RSA algorithms.

Keywords: RSA · Gradient · Searchlight · Whole-brain fMRI data

1 Introduction

One of the most significant challenges in brain decoding is finding some more effective ways of using multi-subject and whole-brain fMRI data. Representational Similarity Analysis (RSA) is one of the fundamental approaches in fMRI analysis and evaluates similarities between different cognitive tasks [1–3]. Here, one subject is scanned while watching different visual stimuli. With different pairs of stimuli, the brain generates corresponding patterns of neural activities, and then the RSA calculates the similarities between the neural activity patterns of different stimuli. This process obtains Representational Similarity Matrix (RSM), and the matrix encodes the similarity structure. The goal of the method is to explore the correlation between different cognitive tasks. Figure 1 shows the computation of the representational similarity matrix (RSM).

RSA can be casted as a multi-task regression problem. Classical RSA is based on basic linear approaches, e.g., Ordinary Least Squares (OLS) or General [1, 2]. Indeed, these methods are restricted to a linear model, each data contains a large number of voxels, and the number of voxels far exceeds the time points. The methods mentioned cannot obtain satisfactory results on fMRI datasets. Moreover, the data is difficult to be converted into a matrix by this method [4], and it could reduce the stability and robustness of the final analysis when the Signal-to-Noise Ratio (SNR) is low [7].

For OLS and GLM, they face a problem of overfitting. The current approaches consider that the regularization can avoid overfitting. For example, Least Absolute Shrinkage and Selection Operator (LASSO) method employs norm ℓ_1 to address the regression problem [9], whereas Ridge Regression method uses the norm ℓ_2 to deal with the mentioned problem [8]. As an alternative approach, the Elastic Net method handle above issue by employing ℓ_1 and ℓ_2 norms [10].

In general, The RSA provides a way to compare different representational geometries across subjects, brain regions, measurement modalities, and even species. Since the similarity structure can be estimated from the imaging data even if the coding model is not constructed, RSA is suitable not only for model testing but also for exploratory research [3]. Indeed, RSA is initially used as a tool to study visual representations [2, 5, 6], semantic representations [12, 13], and lexical representations [14]. Further, RSA is utilized to reveal the network about dimensions of social-information representations [15, 16].

As an alternative to region-of-interest based analysis, researchers introduce the ‘searchlight’ approach that performs multivariate analysis on sphere-shaped groups of voxels centered on each brain voxel one by one [1]. Nowadays, fMRI brain image datasets have a large number of subjects. Thus the whole-brain datasets are high-dimensional. In the current general RSA algorithm, the data is difficult to be converted into a matrix by this method and the inverse of the voxel matrix cannot be avoided. Besides, the optimization of RSA is difficult when the number of voxels is too large. Fortunately, modern RSA algorithm can optimize the solution process in comparison to traditional RSA method [17]. One of the modern RSA methods utilizes the searchlight technique, which is applied to MEG [14]. As a novel application, the searchlight RSA method can be utilized to analyze the structure of moral violations space [11].

In this paper, we propose a new RSA method based on gradient descent called Gradient Representational Similarity Analysis (GRSA). The Gradient RSA algorithm can handle the RSA problem by calculating the solution of LASSO using stochastic gradient descent. It can solve the mapping feature matrix by using stochastic gradient descent method with iteration to obtain an optimal result and explore the similarity between different neural activity patterns. Another key contribution of this paper is a novel application for Searchlight. GRSA is a tool for analyzing whether localized brain regions encode cognitive similarities. Using searchlight, we propose a new method called spatiotemporal searchlight GRSA (SSL-GRSA). In Sect. 3.2, we focus on this approach with an aim to link searchlight analysis with GRSA. We develop this model by using a spatiotemporal searchlight GRSA algorithm which can generalize our ROI-based GRSA algorithm to the whole-brain data.

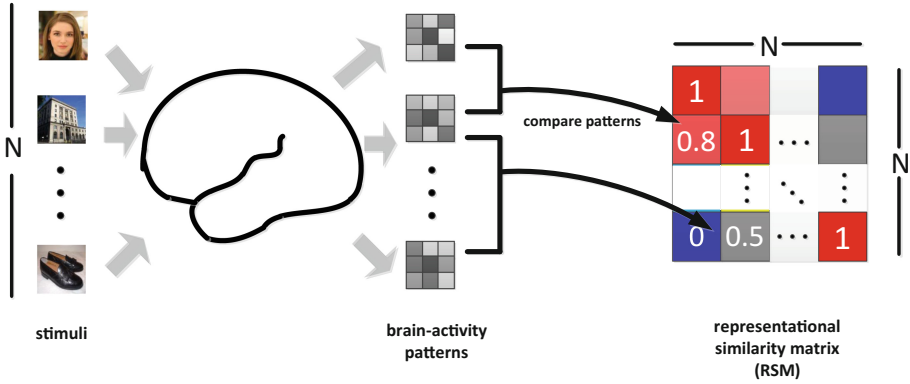


Fig. 1. Computation of the representational similarity matrix (RSM). The matrix encodes the similarity structure. Each block in the RSM is a correlation distance between activation patterns of a pair of experimental conditions (or stimuli). The elements on the main diagonal of the matrix are one by definition. In the non-diagonal part of RSM, a larger value indicates that two stimuli have a high similarity, and the small value implies that the two stimuli are not similar.

2 Representational Similarity Analysis (RSA)

The application of RSA is based on a general linear model (GLM). This method assumes that the neural pattern of fMRI responses is related to stimuli events.

$$Y^{(\ell)} = X^{(\ell)}B^{(\ell)} + \epsilon^{(\ell)} \tag{1}$$

where $Y^{(\ell)} = \{y_{ij}\} \in \mathbb{R}^{T \times V}$, $1 \leq i \leq T$, $1 \leq j \leq V$ denotes the fMRI time series from ℓ -th subject, T is the number of time points and V is the number of brain voxels. Design matrix is denoted by $X^{(\ell)} = \{x_{ik}\} \in \mathbb{R}^{T \times P}$, $1 \leq i \leq T$, $1 \leq k \leq P$. The design matrix is obtained by the convolution of the time series of the stimuli with a typical hemodynamic response function (HRF). Here, P denotes the number of distinct categories of stimuli, $B^{(\ell)} = \{\beta_{kj}\} \in \mathbb{R}^{P \times V}$, $\beta_{kj} \in \mathbb{R}$, $1 \leq k \leq P$, $1 \leq j \leq V$ denotes the matrix of estimated regressors, and β_{kj} is an amplitude reflecting the response of j -th voxel to the k -th stimulus. This paper assumes that the neural activities of each subject are column-wise standardized, i.e., $Y^{(\ell)} \sim \mathcal{N}(0, 1)$. Indeed, RSA method is looking for the following objective function:

$$\min_{B^{(\ell)}} \|Y^{(\ell)} - X^{(\ell)}B^{(\ell)}\|_F^2 - r(B^{(\ell)}) \tag{2}$$

where $r(B^{(\ell)})$ is the regularization term for ℓ -th subject. Notably, the regularization term is zero ($r(B^{(\ell)}) = 0$) for non-regularized methods, including OLS and GLM. The term $r(B^{(\ell)})$ is $\alpha \|B\|_F^2$ for Ridge Regression, $\alpha \|B\|_1$ for LASSO method, $\alpha \rho \|B\|_1 + \frac{\alpha(1-\rho)}{2} \|B\|_F^2$ for Elastic Net method.

In order to generalize RSA for multi-subject fMRI datasets, we calculate the mean of the regressors matrices across subjects:

$$\mathbf{B}^* = \frac{1}{S} \sum_{\ell=1}^S \mathbf{B}^{(\ell)} \quad (3)$$

where S denotes the number of subjects, and each row of $\mathbf{B}^* \in \mathbb{R}^{P \times V} = \{\beta_1^*, \dots, \beta_p^*\}$, $\beta_k^* \in \mathbb{R}^V$ illustrates the extracted neural signature belonging to k -th category of cognitive tasks.

Three metrics will be used to evaluate the performance of RSA methods. As the first metric, we calculate the mean of square error for analyzing the accuracy of regression:

$$MSE = \frac{1}{TSV} \sum_{\ell=1}^S \sum_{i=1}^T \sum_{j=1}^V \left(x_{ij}^{(\ell)} - \sum_{k=1}^P d_{ik}^{(\ell)} \beta_{kj}^{(\ell)} \right)^2 \quad (4)$$

The next two techniques evaluate between-class correlation and between-class covariance of the regressors matrices:

$$CR = \frac{1}{S} \sum_{\ell=1}^S \max_{\substack{1 \leq i \leq P \\ i < j \leq P}} \left\{ Corr\left(\beta_i^{(\ell)}, \beta_j^{(\ell)}\right) \right\} \quad (5)$$

$$CV = \frac{1}{S} \sum_{\ell=1}^S \max_{\substack{1 \leq i \leq P \\ i < j \leq P}} \left\{ Cov\left(\beta_i^{(\ell)}, \beta_j^{(\ell)}\right) \right\} \quad (6)$$

where $\beta_i^{(\ell)}, \beta_j^{(\ell)}$ are rows of $\mathbf{B}^{(\ell)}$, function $Corr$ is the Pearson correlation, and function Cov calculates the covariance between two vectors. All of these three metrics must be minimized for an ideal solution [7, 17].

3 Gradient Representational Similarity Analysis (GRSA)

fMRI brain data is high-dimensional. In fMRI, each data contains a large number of voxels, and the number of voxels far exceeds the time points. Meanwhile, the presence of similarity of different features leads to some redundant information. Feature selection can solve this problem. Therefore, we use the ℓ_1 norm here. The objective function is optimized as follows:

$$J\left(\mathbf{B}^{(\ell)}\right) = \min_{\mathbf{B}^{(\ell)}} L\left(\mathbf{B}^{(\ell)}\right) + r\left(\mathbf{B}^{(\ell)}\right) \quad (7)$$

where the typical loss functions considered here are squared Frobenius error, i.e., $L(\mathbf{B}^{(\ell)}) = \|\mathbf{Y}^{(\ell)} - \mathbf{X}^{(\ell)} \cdot \mathbf{B}^{(\ell)}\|_F^2$, and $r(\mathbf{B}^{(\ell)})$ is the ℓ_1 norm defined as $\alpha\|\mathbf{B}\|_1$. The problem of this approach is that the computation complexity is tremendous when there are a large number of features. And this method is merely applies to the linear model.

3.1 Optimization

In this section, we attempt to propose a method that is not restricted to a linear model and can reduce the time complexity on high-dimensional data. Here, we propose an effective approach that utilizes Stochastic Gradient Descent (SGD) for optimizing the LASSO objective function. In order to efficiently optimize (7), one solution is to calculate the gradient of (7) which is needed in Stochastic Gradient Descent (SGD) algorithm. The step of gradient optimization is as follows:

$$\nabla J(\mathbf{B}_t^{(\ell)}) = \frac{\partial}{\partial \mathbf{B}_t^{(\ell)}} J(\mathbf{B}^{(\ell)}) \quad (8)$$

$$\mathbf{B}_{t+1}^{(\ell)} = \mathbf{B}_t^{(\ell)} - \alpha^t \nabla J(\mathbf{B}_t^{(\ell)}) \quad (9)$$

where $\nabla J(\mathbf{B}_t^{(\ell)})$ denotes the gradient of $J(\mathbf{B}^{(\ell)})$ from t -th iteration. The step of iteration of $\mathbf{B}^{(\ell)}$ denoted as (9). α^t is the self-adaptive learning rate, which is defined as follows:

$$\alpha^t = \frac{\alpha}{\sqrt{t+1}} \quad (10)$$

Here, $t \in \mathbb{R}$ is the number of iterations. α^t denotes the updated learning rate of t -th iteration. Since different features have different ranges of values, the iteration could be very slow. In order to apply this algorithm to fMRI brain datasets, the SGD algorithm randomly selects a batch of the time points instead of the whole time points to update the model parameters. So each time of learning is fast and the model parameters can be updated online. This paper uses GRSA approach for estimating the optimized solution. GRSA can reduce the time complexity when applied to fMRI brain datasets, and explore the similarity between different neural activity patterns by iterative optimal algorithm. Our method can rapidly reduce the time complexity and have smaller memory footprint in each process. This application of GRSA could be used not only in the linear model but also in the non-linear model.

3.2 Spatiotemporal Searchlight GRSA (SSL-GRSA)

Finding the most effective method for analyzing multi-subject fMRI data is a long-standing and challenging problem. Since the scarcity of data for each subject and the differences of brain anatomy and functional response between different subjects, researchers have an increasing interest in human cognitive fMRI research.

Multi-subject fMRI datasets contain two group datasets, i.e., Region of Interests (ROI) based datasets, and whole-brain datasets. The ROI-based method analyzes the representation structure in a set of predefined brain regions. However, other brain regions also have representational structures that are suitable for the prediction of our model. Whole-brain data can be used to figure out what information is represented in a region of the human brain. People want to find some more effective ways to analyze whole-brain data. Searchlight analysis provides a way to map cube-shaped groups of voxels across the whole brain continuously [1]. Therefore, we propose a method that combines the ideas of the GRSA model and searchlight-based technique to analyze multi-subject whole-brain fMRI data. A searchlight version of GRSA is conceptually new. Therefore, we refer to our method as Searchlight GRSA (SSL-GRSA).

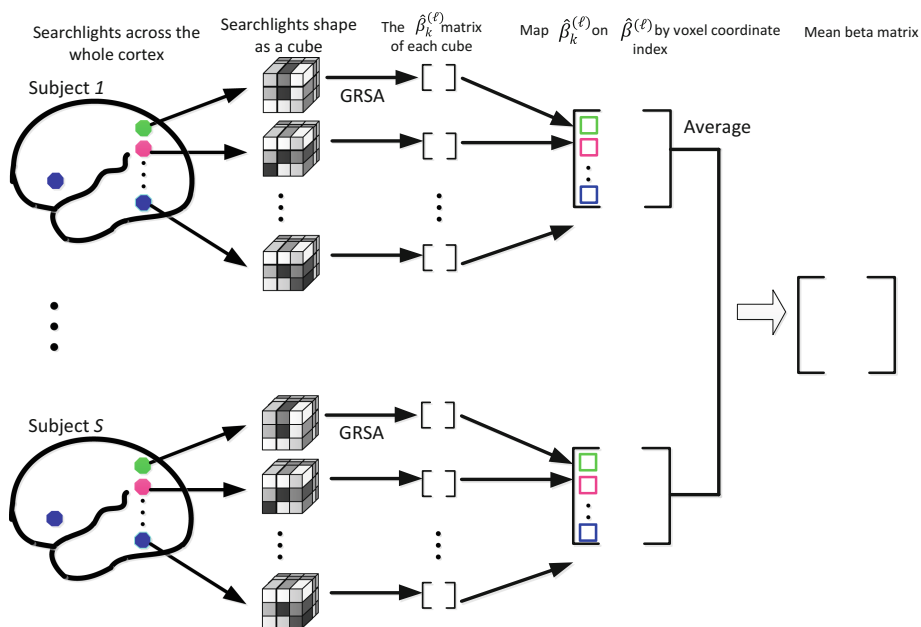


Fig. 2. Process of Spatiotemporal Searchlight GRSA (SL-GRSA). The whole-brain data of each subject is divided into K cubes (searchlights) with a specified size. Here, this size is fixed as $3 \times 3 \times 3$. Then, the GRSA approach applies to each cube to generate K local matrices denoted by $\hat{\beta}_k^{(\ell)}$. In the end, we splice those K local $\hat{\beta}_k^{(\ell)}$ matrices into a complete $\hat{\beta}^{(\ell)}$ matrix according to the coordinates of voxels. The mean matrix is obtained by averaging over all matrices $\hat{\beta}^{(\ell)}$.

$\hat{Y}^{(\ell)} \in \mathbb{R}^{v_x \times v_y \times v_z \times T}$ of four dimension is fMRI time series data from ℓ -th subject where $1 \leq \ell \leq S$ and S is the number of subjects. The tuple (x, y, z) refers to the standard axes, whereas v_x, v_y, v_z refer to the number of voxels along the corresponding axis respectively, and T is the number of time samples in units of repetition time (TR). The process of our searchlight method is as follows: Firstly, a sliding cube is selected and

the cube at a specific time covers a contiguous region of voxels. The selected snapshots of the cube need to be adjacent and avoid overlapping. Then, the voxels of the whole-brain is then analyzed by spatial local analysis in each cube. GRSA method is applied to cube groups of voxels in a line. Therefore, the ROI method can be extended to the whole-brain data. The process of our method is depicted in Fig. 2.

Table 1. The datasets.

Title	ID	Task type	S	P	T	Scan	TR	TE
Visual object recognition	R105	Visual	6	8	121	G3T	2500	30
Word and object processing	R107	Visual	49	4	164	S3T	2000	28
Weather prediction without feedback	W011	Decision	14	4	236	S3T	2000	25
Selective stop signal task	W017	Decision	8	6	546	S3T	2000	25
Weather prediction	W052	Decision	13	2	450	S3T	2000	20

This paper utilizes five datasets, shared by Open fMRI (<http://openfmri.org>). S is the number of subject, P denotes the number of stimulus categories, T is the number of scans in unites of scans in unites of TRs (Time of Repetition), V_{ROI} denotes the number of voxels in ROI. In the column of Scan, G = General Electric, or S = Siemens in 3 T. TR is Time of Repetition in millisecond and TE denotes Echo Time in millisecond.

For standard Searchlight-based RSA method, the study first used the scene image as task stimuli for experiment, and then used the Searchlight method to find brain regions related to the perception of human brain. The results show that using the searchlight method, we can find the active brain regions in the fMRI data related to scene recognition of each subject. Compared with standard searchlight RSA, our method is competitive and performs better with the same cube size. It’s worth mentioning that we only load necessary data according to the mini batch to maintain a reduced memory footprint in each process. We extend the application of GRSA from ROI to the whole-brain. Further, we create a novel approach that addresses some computational challenges while dealing with large-scale, multi-subject fMRI data.

4 Experiments

4.1 Datasets

This paper utilizes five datasets, shared by Open fMRI (<http://openfmri.org>), for running empirical studies. All datasets are separately preprocessed by FSL 5.0.10 (<https://fsl.fmrib.ox.ac.uk>), i.e., slice timing, anatomical alignment, normalization, smoothing. Here, we use two groups of datasets, i.e., Region of Interests (ROI) based datasets, and whole-brain datasets. Here, we analyze some specific parts of brain images in ROI-based data, where these parts are manually selected based on the original papers of each data. In this paper, we use ‘R’ prefix for the ROI-based dataset and a ‘W’ prefix is used for denoting the whole-brain data.

Technically, the whole-brain datasets include all of the neural activities which are registered to a standard space, i.e., Montreal Neurological Institute (MNI) 152 space $T1$ with voxel size 4 mm. Before applying our approach to each fMRI dataset, the dataset

is normalized, i.e., $Y^{(\ell)} \sim \mathcal{N}(0, 1)$, which allows us to obtain desirable experiment result. The technical information of these datasets is shown in Table 1.

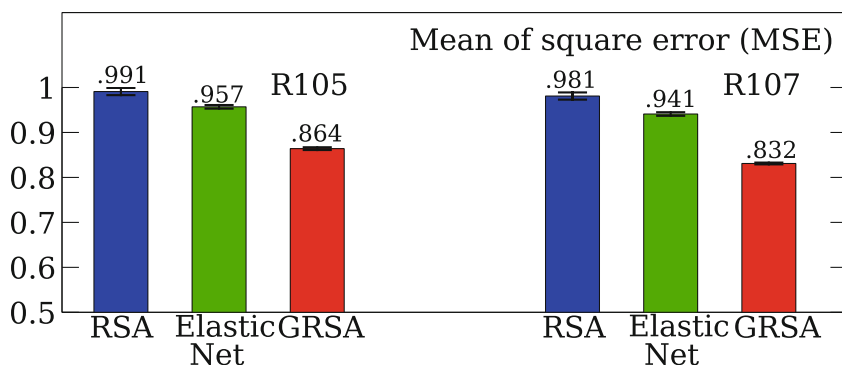


Fig. 3. The standard deviation of MSE for all RSA methods in the Fig. 3 is lower than 10^{-2} .

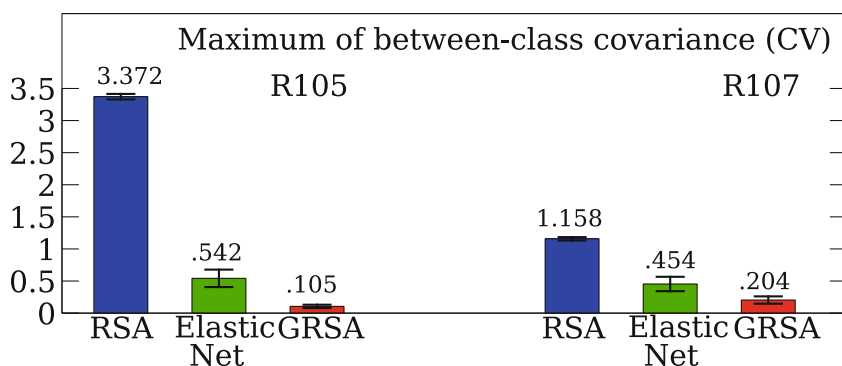


Fig. 4. Maximum of between-class covariance (CV) across subjects.

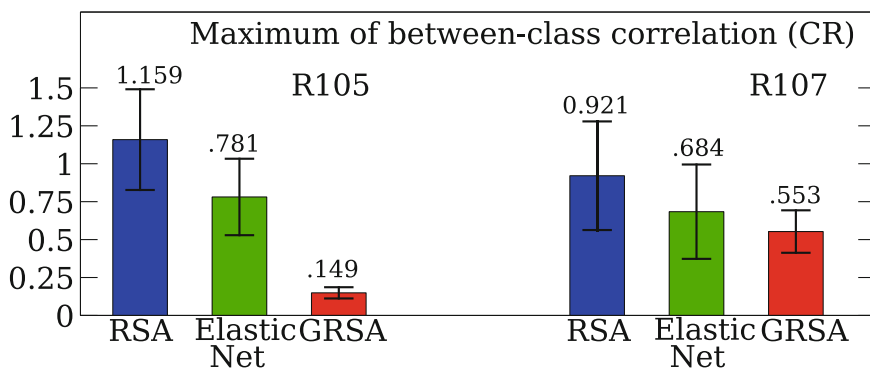


Fig. 5. Maximum of between-class correlation (CR) across subjects.

4.2 ROI Data Analysis

In this section, we analyze the performance of our method results by calculating three metrics, including mean of square error (MSE), the maximum of between-class covariance (CV), and the maximum of between-class correlation (CR). We use the ROI data in each experiment, thus R105 dataset and R107 dataset are selected from five different datasets. In order to create the comparative experiments, we use the classical RSA based on GLM as a baseline. Elastic Net is employed for the empirical research. In this method, the best results are obtained when the parameters are $\alpha = 1.0$ and $\rho = 0.5$. Moreover, GRSA generates the results by setting $\alpha = 0.9$. The number of iterations for our method is considered 1000. The batch size is set 50 and learning rates of normalized datasets is 10^{-3} .

Figure 3 shows the test results of MSE, which is non-negative. MSE is an indicator used to reflect the quality of the estimator. the smaller the MSE is, the better the method is. Further, MSE is calculated by Formula (4). The results of our method in comparison to other methods are shown in Fig. 3. GRSA has the best results compared to other RSA methods. The standard deviation of MSE for all RSA methods in the Table 2 is lower than 10^{-2} .

Figure 4 has analyzed the maximum of between-class covariance by using (6). The maximum of between-class covariance can be calculated as the maximum value ranging over all different pairs of stimuli. Moreover, Fig. 5 has evaluated the maximum of between-class correlation by employing (5) in which it searches the maximum Pearson correlation coefficient amongst different pairs of stimuli. For those indicators, the smaller they are, the better the method analyzes the similarity between different neural activity patterns. Compared with other RSA methods in Fig. 4 or Fig. 5, GRSA has the best results.

4.3 Whole-Brain Data Analysis

ROI is a manually selected area based on anatomical images of the brain. We analyze the potential information of the data through the ROI based method. However, a certain type of information is not necessarily confined to only one specific brain region, and could be included in several areas. Therefore, the analysis of the whole-brain data becomes more important. The GRSA method is applied to whole-brain data and this approach can explore the relationship between different cognitive tasks. In this paper, the whole-brain datasets are used in our method, i.e., W011 dataset, W017 dataset and W052 dataset.

In this section, we implement the comparative experiments by some traditional methods. We use the ordinary Spatiotemporal Searchlight RSA (SSL-RSA) as the baseline. For the empirical study, Spatiotemporal Searchlight Elastic Net (SSL- Elastic Net) is utilized. As mentioned before, both SSL-RSA and RSA share the same parameters. And so do SSL- Elastic Net and Elastic Net. Previously mentioned, the main challenges are the high dimension of data and the issue of memory footprint.

Our approach can address these challenges and has good performance. The cube size can be set arbitrarily. Thus, all Searchlight RSA methods take the same cube size

set as $3 \times 3 \times 3$. In fact, the best result is obtained by using this cube size. The result of each contrast experiment is showed in Tables 2 and 3.

In each comparative experiment, we evaluate all the methods by using CV and CR. The formulas of these two indicators have already been mentioned in the previous section. Table 2 has analyzed the maximum of between-class covariance whereas.

Table 2. Maximum of between-class covariance (CV) across subjects (max±std)

Datasets	SSL-RSA	SSL-elastic net	SSL-GRSA
W011	0.415 ± 0.125	0.265 ± 0.046	0.208 ± 0.042
W017	0.462 ± 0.062	0.237 ± 0.186	0.143 ± 0.143
W052	1.831 ± 0.184	0.396 ± 0.143	0.237 ± 0.052

Table 3. Maximum of between-class correlation (CR) across subjects (max±std)

Datasets	SSL-RSA	SSL-elastic net	SSL-GRSA
W011	0.785 ± 0.033	0.507 ± 0.042	0.609 ± 0.202
W017	0.849 ± 0.124	0.441 ± 0.052	0.358 ± 0.082
W052	0.866 ± 0.071	0.471 ± 0.104	0.407 ± 0.151

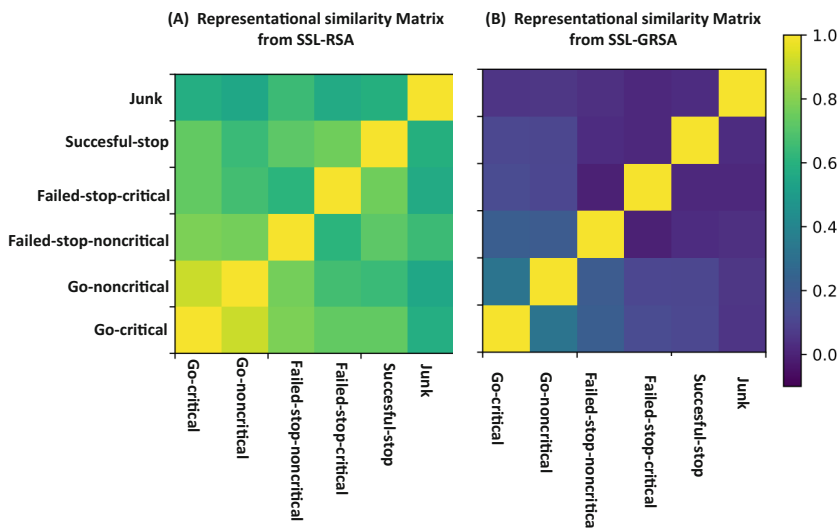


Fig. 6. Comparing correlation of a traditional method and SSL-GRSA method by using W017

Table 3 evaluated the maximum of between-class correlation. As depicted in the result Table 2, SSL-GRSA has generated better performance in comparison with other methods. Further, as Table 3 demonstrates, the performance of the maximum of between-class correlation is significantly lower except for W011, which confirms that our method is better.

Base on W017 data, Fig. 6 depicts the comparison of correlation of a traditional method and SSL-GRSA method. Each small block shows the similarity of the related category of stimuli with respect to the corresponding row and column. Therefore, we compare the between-class correlation of SSL-GRSA with the traditional methods. SSL-GRSA provides the best similarity analysis compared with other methods.

4.4 Runtime Analysis

This section analyzes the runtime of the proposed method and compares it to the runtime of other RSA methods. Here, the analysis is based on the ROI datasets. For convenience, the runtime of other methods is scaled based on GRSA, that is, the runtime of GRSA is regarded as a unit. As illustrated in Fig. 7, the Elastic Net is the slowest one whereas traditional RSA beats others. Since GRSA utilizes a min-batch of time-points, it runs faster than the regularized method. As a conclusion, the performance of GRSA is more efficient. It is worth mentioning that the runtime of the whole brain dataset has the same tendency.

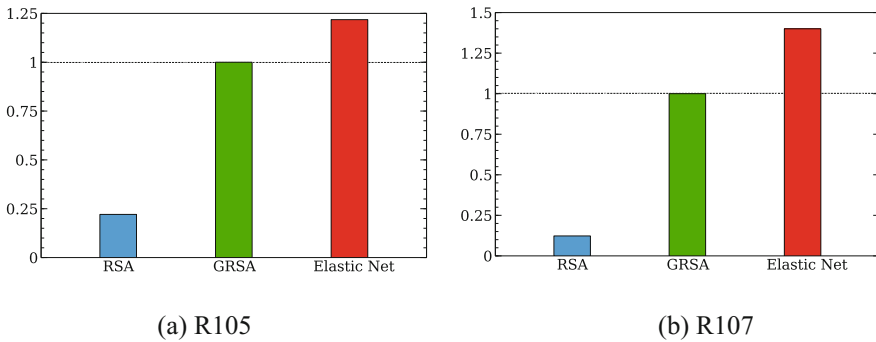


Fig. 7. Runtime analysis

5 Conclusion

In this paper, we explored the method of Representational Similarity Analysis. we propose a novel RSA method called Gradient descent RSA. The Gradient-RSA algorithm handles the RSA problem by calculating the solution of LASSO using stochastic gradient descent, which is novel to RSA study. For the whole-brain data, the primary challenges are the high dimension of data and the issue of memory footprint. Another primary contribution of this paper is a new application in Searchlight. Based on Searchlight, the application of our GRSA method is extended from the localized brain regions to the whole-brain region. Further, Our methods show improved results over standard competing methods. In the future work, our method can be applied to more large-scale, multi-subject fMRI datasets, and further optimized by other new approaches to obtain better performance.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant (61876082, 61861130366, 61703301, and 61473149), the Fundamental Research Funds for the Central Universities and the Foundation of Graduate Innovation Center in NUAA (kfjj20171609).

References

1. Kriegeskorte, N., Goebel, R., Bandettini, P.: Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U. S. A.* **103**(10), 3863–3868 (2006)
2. Connolly, A.C., et al.: The representation of biological classes in the human brain. *J. Neurosci.* **32**(8), 2608–2618 (2012)
3. Kriegeskorte, N., Mur, M., Bandettini, P.A.: Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008)
4. Yousefnezhad, M., Zhang, D.: Anatomical pattern analysis for decoding visual stimuli in human brains. *Cogn. Comput.* **10**(2), 284–295 (2018)
5. Peelen, M.V., Caramazza, A.: Conceptual object representations in human anterior temporal cortex. *J. Neurosci.* **32**(45), 15728–15736 (2012)
6. Kravitz, D.J., Peng, C.S., Baker, C.I.: Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *J. Neurosci.* **31**(20), 7322–7333 (2011)
7. Cai, M.B., Schuck, N.W., Pillow, J.W., Niv, Y.: A Bayesian method for reducing bias in neural representational similarity analysis. In: *Advances in Neural Information Processing Systems*, pp. 4951–4959 (2016)
8. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970)
9. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **68**(1), 49–67 (2006)
10. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **67**(2), 301–320 (2005)
11. Wasserman, E.A., Chakroff, A., Saxe, R., Young, L.: Illuminating the conceptual structure of the space of moral violations with searchlight representational similarity analysis. *NeuroImage* **159**, 371–387 (2017)
12. Handjaras, G., et al.: How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage* **135**, 232–242 (2016)
13. Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L.: A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**(6), 1210–1224 (2012)
14. Su, L., Fonteneau, E., Marslen-Wilson, W., Kriegeskorte, N.: Spatiotemporal searchlight representational similarity analysis in EMEG source space. In: *2012 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pp. 97–100. IEEE (2012)
15. Tamir, D.I., Thornton, M.A., Contreras, J.M., Mitchell, J.P.: Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. *Proc. Natl. Acad. Sci.* **113**(1), 194–199 (2016)
16. Chavez, R.S., Heatherton, T.F.: Representational similarity of social and valence information in the medial pFC. *J. Cogn. Neurosci.* **27**(1), 73–82 (2015)
17. Oswal, U., Cox, C., Lambon-Ralph, M., Rogers, T., Nowak, R.: Representational similarity learning with application to brain networks. In: *International Conference on Machine Learning*, pp. 1041–1049 (2016)



Feature Aggregation Tree: Capture Temporal Motion Information for Action Recognition in Videos

Bing Zhu^(✉)

Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology (BIT),
Beijing 100081, People's Republic of China
zhubing@bit.edu.cn

Abstract. We propose a model named Feature Aggregation Tree to capture the temporal motion information in videos for action recognition. Feature Aggregation Tree constructs a logical motion sequence by considering the concrete semantics of features and mining feature combinations in a video. It will save different feature combinations and then use the bayesian model to calculate the conditional probabilities of frame-level features based on the previous features to aggregate features. It doesn't matter about the length of the video. Compared with the existing feature aggregation methods that try to enhance the descriptive capacity of features, our model has the following advantages: (i) It considers the temporal motion information in a video, and predicts the conditional probability by using the bayesian model. (ii) It can deal with arbitrary length of the video, rather than uniform sampling or feature encoding. (iii) It is compact and efficient compared to other encoding methods, with significant results compared to baseline methods. Experiments on the UCF101 dataset and HMDB51 dataset demonstrate the effectiveness of our method.

Keywords: Action recognition · Feature learning
Feature aggregation

1 Introduction

Human action recognition [1] is one of the fundamental researches in the field of computer vision, which has great significance and application prospects in video retrieval, video recommendation and video surveillance. In recent years, many researches mainly focus on two aspects. One is how to extract a more discriminative spatio-temporal description for the video. The other is how to aggregate frame-level features to a video-level feature, which gives more attention to efficient feature organization strategies.

In terms of feature description, most of the existing video feature representations for action recognition are mainly learned by two different types of networks: one is two-stream network [2,3] and the other is 3D convolutional neural network [4–6]. The trend of networks is to learn better video features which can capture both spatial and temporal information in videos. And we need a strategy to handle long videos with arbitrary frames, which can aggregate frame-level features to a representation for the whole video.

In terms of feature aggregation, one strategy is selecting a key frame or several key frames to represent the entire action video [7–9]. This strategy can achieve satisfactory results when a video contains only one action instance, but it is not so useful in the videos containing multiple categories action instances. Another common strategy is to encode frame features, such as vectors of locally aggregated descriptors (VLAD) [10], fisher vectors (FV) [11,12] and bag of words (BoW) [13,14]. While these strategies cannot capture the temporal information of the entire video. In addition, in the neural network methods, the temporal pooling operation is usually used to compress the features of a video [3,15,16], e.g. the mean and the max pooling. There are also some recent works trying to modify the traditional pooling strategies to further improve the recognition performance, such as adascan [17] and ActionVLAD [18], which attaches frame features to different wight values. However, the pooling strategies don't consider the order of frames, which ignore the temporal information. Besides the CNNs, the LSTM network is also considered to use attention mechanism to learn the weight of different each frame [19–21]. But because of the complexity of the training process, LSTM doesn't become a mainstream method.

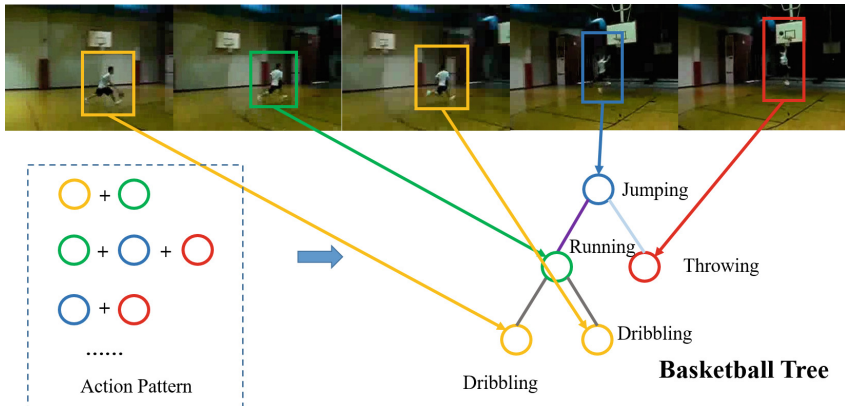


Fig. 1. We propose Feature Aggregation Tree to represent actions in videos. For example, Action “Basketball” can be grouped by “running”, “dribbling”, “jumping” and “throwing”. We construct a “basketball” tree to record action primitives nodes and excavate the action pattern between two action primitives.

To better understand what an action is performing, temporal information is as important as spatial information. However, no matter hand-crafted features or deep features, are all frame-level, which don't make full use of the temporal information of the entire video. To better use the temporal information, we need to understand the component of actions firstly. The hierarchical definition proposed by Moeslund [22] divides the actions into three levels, e.g. the lower level definition is "action primitive", the middle level definition is "action" which is an ordered combination of primitives, while the concept of high-level definition of "behavior" is the logical combination of actions, which is a complex advanced semantics. Taking the action of basketball as an example, shown in Fig. 1, playing basketball can be broken down into several action primitives like "running", "dribbling", "jumping" and "throwing" and these primitives are organized in a temporal order. Actions have different meanings in different orders, such as "running-throwing" means playing basketball, while "running-jumping" means high jump or long jump. And these temporal information involved in the patterns will be helpful in action recognition. The method ActionVLAD with the similar idea proves the effectiveness. In this work, we propose a novel method named Feature Aggregation Tree (FA-Tree) to learn video features for action recognition, which is based on the knowledge of frequent patterns and association rules in the field of data mining [23].

The main contribution of this work is that we propose a novel FA-Tree for action recognition, which has the following advantages: (i) The method treats frame-level features as action primitives, and aggregate them into action patterns. Taking the temporal information of primitives into account, FA-Tree organizes patterns with different orders to better represent a complete action, and then calculate the precise conditional probability of an action. (ii) The method can deal with arbitrary length of the video, rather than uniform sampling or feature encoding. (iii) The model is compact and efficient, and has achieved good results on two datasets.

2 Related Work

Action Feature Representation. In recent years, more and more researchers want to extract more discriminative features to represent a video, which should contain temporal information as well as spatial information. Some hand-crafted traditional features [1, 24, 25] are proposed from 2D to 3D, and their description ability has been significantly improved. It is worthy mentioning that Wang et al. [26] proposed improved Dense Trajectories (iDT), which is the best hand-crafted feature at present but it is computationally intensive. Simonyan and Zisserman [3] proposed the two-stream network, which decomposed a video into appearance and motion streams, and trained two networks respectively. Considering that the input of 2D convolutional neural networks is always an image so it lacks the temporal information, the 3D neural network uses the video segment as the input [4-6].

Video Feature Aggregation. One approach is to select a key frame or a key segment to replace the entire video when predicting the action category. Cao et al. [7] extracted the key frame with manifold learning based on the optical flow graph for action recognition. Liu et al. [8, 9] used supervised learning and unsupervised clustering methods to extract key segments in action videos.

Another approach is feature encoding. Some methods use the bag of words model (BoW) [14] to extract some local spatio-temporal descriptor, and encode them into dictionaries to make templates [13, 15, 27, 28]. Latev et al. [27] described a video with BoW that encoded HoG and HoF features. Ji et al. [5] also used BoW in their method. Similar to BoW are the methods such as VLAD [10, 18] and Fisher Vector [11–13]. Wang et al. [15] proposed the improved Dense Trajectories(iDT) approach, which combined dense trajectories, histogram by using Fisher Vector to encode. By combining iDT [26] features and Fisher Vector [29] algorithm, Peng et al. [13] discussed fusing first and then encoding or encoding first and then fusing, and finally found the latter method is better. Tang et al. [30] proposed a more flexible approach using a variable duration HMM [31] that factored each video into latent states with variable durations.

Now the popular strategy in the neural network is to compress the information of different frames in a video into a fixed summary vector by using pooling operation [3, 4, 15, 16]. The mean pooling and the max pooling are common choices, i.e. taking average or maximum values of each feature vector, such as C3D [4] adopts the average value of each feature in every dimension. However, these pooling methods consider each frame equally, which is not robust to the noisy information. As there may be some noisy frames in the video, these noisy frames will cause some losses and ultimately lead to error judgments. Some recent works try to modify pooling strategy for action recognition, such as ActionVLAD [18] and adascan [17].

Frequent Pattern Tree. Our Feature Aggregation Tree, which want to mine action pattern in a video, is inspired by Frequent Pattern Tree. Han et al. [32] introduced the Frequent Pattern Tree structure for storing crucial information about mining frequent patterns in transaction and time-series databases. They also developed the FP-Growth algorithm for efficient and scalable mining on both long and short frequent patterns. Chang et al. [33] proposed an incremental data mining algorithm based on FP-Growth using the concept of heap tree to address the issue of incremental updating of frequent itemsets. Aditya and Pradana [34] leveraged the FP-Growth algorithm to find the customer buying habits on market basket in organic medicine store. Dharmaraajan and Dorairangaswamy [35] utilized the FP-Growth algorithm to classify user behavior in identifying the patterns of the browsing and navigation data of web users.

3 Approach

In this section, we will describe the details of Feature Aggregation Tree. As is outlined in Fig. 2, we extract frame-level features by the C3D network and then regard these features as action primitives, which are the results of the softmax

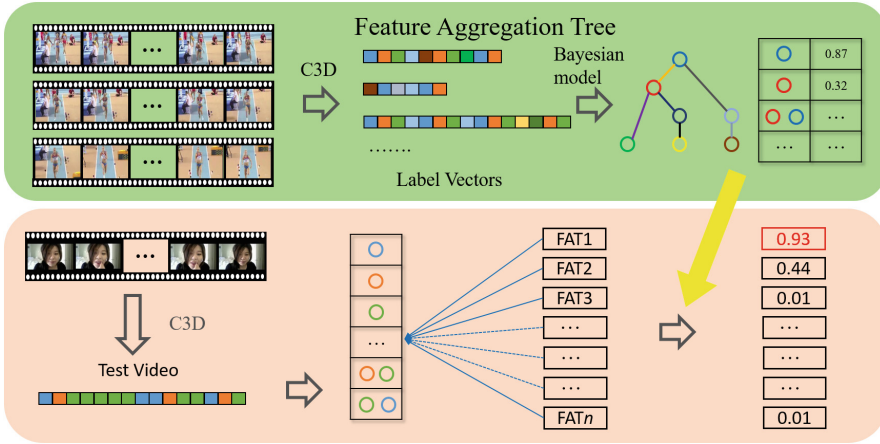


Fig. 2. When constructing the Feature Aggregation Tree (above), we extract the frame-level features by C3D network, and then get feature labels to make up label vectors. We use these label vectors to construct Feature Aggregation Tree model and calculate the probabilities of each node and each pattern. When testing (below), we predict the test video label by matching each FA-Tree and calculating the probability.

layer. The next step is aggregating primitives into patterns to construct Feature Aggregation Tree. In a FA-Tree, each node corresponds to the conditional probability that the node appears, which we use the Bayesian model to calculate. In the following we first describe how to construct Feature Aggregation Tree (Sect. 3.1) and then discuss the strategies for calculating the probability of each action pattern (Sect. 3.2).

3.1 How to Construct FA-Tree

In this part, we will give some definitions about the FA-Tree firstly. Specifically, divide the entire dataset D into different subsets, such as $D = \{S_1, S_2, \dots, S_k\}$. Videos in every subset S_i have the same category label i , i is from 1 to k . And every subset S_i will generate one FA-Tree. Let $S_i = \{v_1, v_2, \dots, v_j\}$, where v means a label vector, as every video corresponds to a label vector by C3D network, and j is the number of videos in subset S_i . Just like what is shown in Fig. 2. For each video, each frame in the video is regarded as an element in the label vector. Here we name one label in the vector as “item”, and two different label pairs as “pattern”. Item set is $F = \{f_1, f_2, \dots, f_m\}$ and pattern set is $P = \{p_1, p_2, \dots, p_n\}$.

The first step is using the unique operation to deal with the same consecutive items. Because in our approach, we just consider different item pairs to mine association rules. The second step is to set support and confidence thresholds. Because there will be some noisy labels in the vector after the softmax layer, we set the minimum item support threshold (MIST) to remove these noisy labels

when the frequency of one item is lower than the threshold. The other threshold is named minimum pattern confidence threshold (MPCT), which is set to choose the root node of a FA-Tree. When we construct a FA-Tree, the root node must be the actual label of this category. So we need to sort items by MIST, and MPCT determines that in top 0.05 or 0.1 rate of all items, we can choose the actual label as the root node. In Sect. 4.2, the data were uniformly sampled in experiments to help set the thresholds.

In addition, when constructing a FA-Tree, we have fully considered the temporal information in a video. Because in the processing step, we have not changed the positions of items. So the remaining items are organized in the order as the original video. The construction of a FA-Tree is divided into three steps. First, those items whose frequency is higher than MPCT are selected as the root node. Second, each label vector is divided into patterns to generate frequent pattern set. Third, for each root node, we connect the items that appear before root node in the left branch, and those after the root node in the right branch. The specific algorithm is shown as below.

Algorithm 1. Pseudo-code of the Construction of Feature Aggregation Tree

Input: Action label vector subset $S_i = \{v_1, v_2, \dots, v_j\}$, *MIST*, *MPCT*

Output: Feature Aggregation Tree *FA-Tree*

- 1 Scan S_i once. Collect items higher than *MIST* to group F . Construct the pattern set P . Sort F by support frequency in the descending order, and choose items higher than *MPCT* to be the *Root* of a FA-Tree ;
 - 2 Scan the pattern set P ;
 - 3 **for** each vector in V_j **do**
 - 4 **for** each pattern in P **do**
 - 5 **if** item p appears before *Root* **then**
 - 6 **if** *Root* has a left child node p **then**
 - 7 | the frequency of p add 1;
 - 8 **else**
 - 9 | reach to the left child node of *Root* recursively, create a new node p , and let its frequency be 1, linked to its parent node and recorded in the list;
 - 10 **else**
 - 11 | the same step as before except right instead of left;
 - 12 **if** there is no p in the pattern **then**
 - 13 | create new *Root* and repeat step 2
 - 14 **final ; return** *FA-Tree*;
-

Given a simple FA-Tree as an example in Fig. 3. The letter ‘a’ means ‘action’ while the subscript of ‘a’ is the result of the softmax layer. The item set is $\{a_2, a_1, a_{20}\}$ and the pattern set is $\{[a_{20}, a_2], [a_2, a_1], [a_2, a_{20}], [a_1, a_{20}]\}$. When the root is a_2 , we make a_{20} to be its left child node and a_1 to be its right child

node. When we extract the pattern $[a_2, a_{20}]$, we find a_2 already has the right child node, so let a_{20} be the right child node of a_1 .

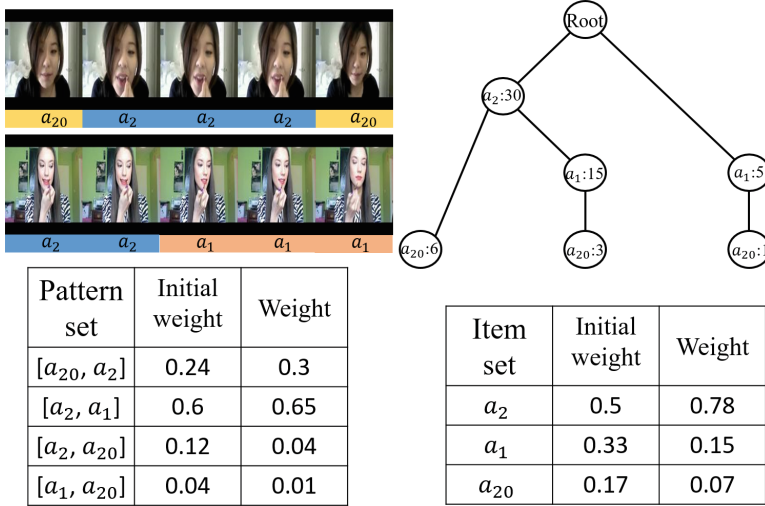


Fig. 3. We construct a simple AF-Tree as an example. Initial weights are calculated by the statistical approach while weights are calculated by the bayesian model.

FA-Tree is a highly compressed structure that stores all the information of action patterns, and the memory space occupied by FA-Tree is proportional to the depth and width of the tree. For the depth of the tree, it generally depends on the complexity of label vectors, as well as the quality of classifier. For example, the more chaotic the label vector is, the deeper the tree will be. The width of the tree indicates that there are not only one root node. FA-Tree is compact because the size of the tree is usually much smaller than the original label vector set.

3.2 How to Design FA-Tree Probability Estimation

After constructing a FA-Tree, we initialize the weights of each node with the simple statistical approach. The definition of weights is shown as below. For each single item, its weight means the probability that it belongs to one action. For each pattern, its weight means the probability product of a two-item combination and this combination belongs to one action. The weight can be thought as the contribution of items and patterns to the whole video. However, simple statistical approach can not get precise weights in our experiments, Table 1. So we use the bayesian model to predict weights, the formula is shown as:

$$P(C_{cls}|l_i) = \frac{P(l_i|C_{cls})P(C_{cls})}{P(l_i)},$$

where C_{cls} means the category of one action, l_i means the i^{th} label in a video label vector. $P(l_i)$ is given by softmax classifier. $P(C_{cls})$ and $P(l_i|C_{cls})$ are calculated by data statistics. So we can update $P(C_{cls}|l_i)$ to have a more precise weight. And the weight will be saved in the node of a FA-Tree.

The FA-Tree is used to compute the probability of the whole video by finding the matched patterns in the test video. We will introduce the probability formula for calculating the video probability, which is as follows:

- Set the node weight parameter μ , pattern weight parameter γ .
- The patterns extracted from a label vector has N nodes and M patterns, referred to as p_{node} and $p_{pattern}$ respectively.

The probability that a test label vector passed by a FA-Tree can be expressed as:

$$P(v, FA - Tree) = \sum_{i=1}^N \mu_i p_i^{node} + \sum_{j=1}^M \gamma_j p_j^{pattern} + c,$$

where v represents the test video label vector; μ and γ are the weight parameters corresponding to p ; c is a penalty, which plays a similar role as bias.

As is shown in Fig. 3, we get the initial weights with the statistical approach. Given that this AF-Tree belongs to action ‘‘ApplyLipstick’’, which is label 2. If the assumption is $P(C_{cls=2}) = 0.76$, and $P(a_2), P(a_1), P(a_{20})$ are given by softmax, we can calculate the weights as the figure.

4 Experiments

4.1 Dataset

UCF101. UCF101 [36] is a dataset which is cut from real action videos in YouTube. It contains a total of 101 action categories and 13320 videos. We use split 1 for the experiment, including 9537 training videos and 3783 test videos, whose total hours up to 27 h.

HMDB51. HMDB51 [37] is collected from a variety of sources, most of which come from movies, and a small percentage from public databases such as Prelinger files, YouTube and Google Video. The dataset contains 6849 segments, which are divided into 51 action categories with at least 101 segments for each category.

4.2 FA-Tree Construction

In the experiment, we use the first split of HMDB51 dataset to show the process of parameter setting. Each video is divided into segments with the length of 16 frames and 50% overlap between segments. We use these video segments as the input of the 3D convolutional neural network [4] and we will get the classification result of each feature after the softmax layer. Therefore, for each action video, we

can get a label vector which is made up by some different labels. We accumulate all the vectors of the same category action in one subset.

To get the item set and pattern set, first, we select 80% training data to predict the remaining 20% and we repeat this step 5 times. We randomly select some data in HMDB51, and finally select about 4500 videos to construct Feature Aggregation Trees. When observing these label vectors, some noisy data need to be removed. We just set the threshold MIST, and items below the thresholds are all excluded. The MIST is set to be 0.05 and the MPCT is set to be 0.1.

In the process of probability estimation, we set c as a penalty coefficient which is shown in the formula of the Sect. 3.2. We also test whether we should set the penalty factor c , which is shown in Table 1. The table (left) records accuracies without the bayesian model and the penalty coefficient c . While the first three columns in the table (right) record accuracies without the bayesian model but with c . And the last column in the table (right) records accuracies with the bayesian model and c .

Table 1. Accuracy (%) comparison between FA-Tree with PN (right) and without PN (left) on the HMDB51 dataset

	Rank-1	Rank-2	Rank-3		Rank-1	Rank-2	Rank-3	Bayesian
Split 1	56.9	69.9	75.0	Split 1	57.0	69.8	75.4	67.7
Split 2	53.3	67.5	72.3	Split 2	53.5	68.2	73.8	63.4
Split 3	55.4	69.2	74.7	Split 3	55.5	69.7	74.8	66.8

When experimenting on the HMDB51 dataset, if we only use the C3D features and all weight of items and patterns are initialized, we can calculate the accuracy of Rank-2 is 69.80%. This shows that the Feature Aggregation Tree can really capture the latent motion information in the video. The reason why these segments can not achieve the highest score is that the predictions are mainly limited to using only the simple softmax. So after using the bayesian model we get the result lower than Rank-2 but higher than Rank-1.

4.3 FA-Tree Comparison Experiment

In this part, we consider Fisher Vector [11, 13] and VLAD [10] to be the baseline method. In addition, we also consider the mean pooling and the max pooling, as well as RNN-FV [38] and ST-VLMPF [39]. The experimental results are in Table 2.

The result of FA-Tree is better than the baseline methods, which proves the effectiveness of our method. It is worthy mentioning that, compared with the improvement on the UCF101 dataset, the result is more obvious on the HMDB51 dataset because the labels in the UCF101 dataset are more ordered. However in the HMDB51 dataset, the FA-Tree can find enough action patterns from chaotic labels to represent the actions and ultimately improve the accuracy.

Table 2. Accuracy (%) comparison between mean pooling, max pooling and FA-Tree on the UCF101 dataset and the HMDB51 dataset

Strategies	UCF101	HMDB51
iFV [11]	79.8	49.0
VLAD [10]	81.4	49.1
RNN-FV [38]	82.3	52.9
Mean pooling	82.7	51.6
Max pooling	83.3	52.5
ST-VLMPF [39]	86.2	56.3
FA-Tree	86.9	66.2

4.4 Comparison with the State-of-the-Art

In Table 3, we show a comparison of our FA-Tree with the state-of-the-art methods on both datasets. Our method with MIFS feature achieves 94.6% on the UCF101 dataset and 74.2% on the HMDB51 dataset.

Table 3. Accuracy (%) comparison of our method with the state-of-the-art methods

Approach	UCF101	HMDB51
Wang et al. [26]	85.9	57.2
Tran et al. [4]	82.6	52.5
Simonyan et al. [3]	88.0	59.4
Peng et al. [13]	87.9	61.1
Wang et al. [16]	90.3	63.2
Wang et al. [2]	94.2	69.4
Kar et al. [17]	93.2	66.9
Girdhar et al. [18]	93.6	69.8
Duta et al. [39]	93.6	69.5
Our Method + MIFS [40]	94.6	74.2

5 Conclusion

We propose a novel model - the Feature Aggregation Tree to capture the temporal motion information in action videos. The FA-Tree connects frame-level features with the specific meanings of action primitives, and mines action patterns in the action sequence. We use the bayesian model to calculate the conditional probability of patterns. The experimental results on the UCF101 dataset and HMDB51 dataset demonstrate the effectiveness of our method.

References

1. Laptev, I., Lindeberg, T.: On space-time interest points. *Int. J. Comput. Vis.* **64**(2–3), 107–123 (2005)
2. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. *ACM Trans. Inf. Syst.* **22**(1), 20–36 (2016)
3. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: *International Conference on Neural Information Processing Systems*, pp. 568–576 (2014)
4. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks, pp. 4489–4497 (2014)
5. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 221–231 (2013)
6. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**(99), 1 (2016)
7. Cao, X., Ning, B., Yan, P., Li, X.: Selecting key poses on manifold for pairwise action recognition. *IEEE Trans. Ind. Inform.* **8**(1), 168–177 (2012)
8. Liu, L., Shao, L., Zhen, X., Li, X.: Learning discriminative key poses for action recognition. *IEEE Trans. Cybern.* **43**(6), 1860–1870 (2013)
9. Jiang, Z., Lin, Z., Davis, L.S.: Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(3), 533–547 (2012)
10. Jegou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation, pp. 3304–3311 (2010)
11. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010*. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_11
12. Sydorov, V., Sakurada, M., Lampert, C.H.: Deep fisher kernels - end to end learning of the fisher kernel GMM parameters, pp. 1402–1409 (2014)
13. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. *Comput. Vis. Image Underst.* **150**(C), 109–125 (2016)
14. Li, F.F., Perona, P.: A Bayesian hierarchical model for learning natural scene categories, pp. 524–531 (2005)
15. Wang, H., Dan, O., Verbeek, J., Schmid, C.: A robust and efficient video representation for action recognition. *Int. J. Comput. Vis.* **119**(3), 219–238 (2016)
16. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors, pp. 4305–4314 (2015)
17. Kar, A., Rai, N., Sikka, K., Sharma, G.: AdaScan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3376–3385 (2017)
18. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: ActionVLAD: learning spatio-temporal aggregation for action classification, pp. 3165–3174 (2017)
19. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. *arXiv preprint arXiv:1511.04119* (2015)
20. Ng, Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification, vol. 16, no. 4, pp. 4694–4702 (2015)

21. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description, pp. 677–691 (2015)
22. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **104**(2), 90–126 (2006)
23. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques, Data Mining Concepts Models Methods & Algorithms*, 2nd edn, vol. 5, no. 4, pp. 1–18 (2011)
24. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition, pp. 357–360 (2007)
25. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: *British Machine Vision Conference 2008*, Leeds, September 2008
26. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: *IEEE International Conference on Computer Vision*, pp. 3551–3558 (2014)
27. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies, pp. 1–8 (2008)
28. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: *BMVC 2009-British Machine Vision Conference*, p. 124:1. *BMVA Press* (2009)
29. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization, pp. 1–8 (2007)
30. Tang, K., Fei-Fei, L., Koller, D.: Learning latent temporal structure for complex event detection, pp. 1250–1257 (2012)
31. Vezzani, R., Baltieri, D., Cucchiara, R.: HMM based action recognition with projection histogram features. In: Ünay, D., Çataltepe, Z., Aksoy, S. (eds.) *ICPR 2010. LNCS*, vol. 6388, pp. 286–293. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-17711-8_29
32. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. *Data Min. Knowl. Discov.* **8**(1), 53–87 (2004)
33. Chang, H.-Y., Lin, J.-C., Cheng, M.-L., Huang, S.-C.: A novel incremental data mining algorithm based on FP-growth for big data. In: *2016 International Conference on Networking and Network Applications (NaNA)*, pp. 375–378. IEEE (2016)
34. Aditya, P.: Market basket analysis using FP-growth algorithm in organic medicine store. *Skripsi, Fakultas Ilmu Komputer* (2016)
35. Dharmaraajan, K., Dorairangaswamy, M.: Analysis of FP-growth and Apriori algorithms on pattern discovery from weblog data. In: *IEEE International Conference on Advances in Computer Applications (ICACA)*, pp. 170–174. IEEE (2016)
36. Soomro, K., Zamir, A.R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint [arXiv:1212.0402](https://arxiv.org/abs/1212.0402) (2012)
37. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition, pp. 2556–2563 (2011)
38. Lev, G., Sadeh, G., Klein, B., Wolf, L.: RNN fisher vectors for action recognition and image annotation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9910, pp. 833–850. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_50
39. Duta, I.C., Ionescu, B., Aizawa, K., Sebe, N., et al.: Spatio-temporal vector of locally max pooled features for action recognition in videos. In: *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pp. 3205–3214. IEEE (2017)
40. Lan, Z., Lin, M., Li, X., Hauptmann, A.G., Raj, B.: Beyond Gaussian pyramid: multi-skip feature stacking for action recognition, pp. 204–212 (2015)



Adaptive Ensemble Probabilistic Matrix Approximation for Recommendation

Xingxing Li, Liping Jing^(✉), and Huafeng Liu

Beijing JiaoTong University, Beijing 100044, China
lpjing@bjtu.edu.cn

Abstract. Matrix approximation has been increasingly popular for recommender systems, which have achieved excellent accuracy among collaborative filtering methods. However, they do not work well especially when there are a large set of items with various types and a huge number of users with diverse interests. In this case, the complicated structure of sparse rating matrix introduces challenges to the single global or local matrix approximation. In this paper, we propose an **Adaptive Ensemble Probabilistic Matrix Approximation** method (**AEPMA**), which can potentially alleviate the data sparsity and improve the recommendation accuracy. By integrating the global information over the entire rating matrix and local information on subsets of user/item ratings in a stochastic gradient boosting framework, **AEPMA** has the ability to capture the overall structures information and local strong associations in an adaptive weight strategy. A series of experiments on three real-world datasets (Ciao, Epinions and Douban) have shown that **AEPMA** can effectively improve the recommendation accuracy and scalability.

Keywords: Adaptive · Ensemble
Global and Local Matrix Approximation · Matrix approximation

1 Introduction

The variety and number of products provided by companies have increased dramatically. Companies produce a large number of products to meet the needs of customers. Although this gives more options to customers, Customers are facing more and more information, and how to obtain information accurately and effectively has become a dilemma. Recommender systems are becoming more important due to the increasing challenge-information overload. Recommender systems provide users with personalized recommendation service based on their preferences, needs, and past behaviors.

Till now, the widely-used historical data is user-item rating matrix which describes the user's observed preference. Most popular recommendation techniques (e.g., matrix approximation-based (MA) collaborative filtering) are proposed on rating matrix. In order to predict the rating accurately, many global-based methods have been proposed. The traditional matrix ratings prediction

based on global information [2,5,9,24] works by studying the latent feature matrix of users/items. Although this method has the advantages of prediction simple and easy to understand the method from math, the interpretability of the recommendation results is low and these methods failed to detect strong associations among a small set of items/users.

In order to solve the problem of the subsets of users' unique interests, researchers adopted local methods [1,25] to predict the missing values of rating matrix. They apply matrix clustering and community detection to matrix approximation methods. The main idea is to partition the large user-item matrix into a set of smaller submatrices, and the usual method for partition is to consider user-based clustering or item-based clustering. However, sub-matrix may appear over-fitting in this local method, and ignore the overall structure on the rating matrix. Now we proposed the new model AEPMA (Adaptive Ensemble Probabilistic Matrix Approximation), which help us sift through all the available global and local information to make accurate matrix rating prediction. The intuition is that, weaker between correlation of two models, more accurate the prediction values for missing value are. So we take both the global and local information into consideration. Simultaneously, we apply a gradient-boosting framework to learn the more accurate values and not sensitive to abnormal points. We learn the weight of different components in the model, which plays an important role in adaptive and effective prediction. More importantly, there is no manual setting of the parameters, both the weight and learning rate.

2 Related Work

Matrix approximation-based collaborative filtering methods have been proposed to alleviate the data missing issue. Some is from the overall structure, RSVD [9] is a standard matrix factorization method inspired by the effective to the domain of collaborative filtering, which is from the domain of natural language processing. Then NMF [24] view the recommendation task as a actual situation, so the components are non-negative and NMF assume the ratings follows the Poisson distribution. Then the Gaussian distribution assumption has been attempted, PMF [2] is a Probabilistic Matrix Factorization model, which define the conditional distribution over the ratings as Gaussian distributions. And later BPF [5] – a Bayesian extension of PMF, in which the model is using Markov chain Monte Carlo (MCMC) methods for approximate inference.

Although these methods work well, these methods still limited in detecting the overall structure. More recently, model such as ACCAMS [1] focused on local strong correlation. ACCAMS [1] is an additive model of co-clustering, which can partition rating matrix into blocks that are highly similar through a clustering of the rows and columns. SIACC [25] is a extension of ACCAMS, and has a better effect on co-clustering by using a social influence. WEMAREC [4] takes the rating distribution into consideration. And as a weighted and ensemble model, the submatrix is generated using different co-clustering constraints in WEMAREC. Furthermore, LLORMA [3], SMA [25] also focused on using ensembles of factorization to exploit local structure. But these ensembles models only focused on

ratings inside clusters and ignore the majority of user ratings outside clusters. Since training data are often insufficient in the detected clusters, the performance of local ensemble models may degrade due to overfitting. To tackle this problem, we address these issues of ratings prediction by applying an ensemble approach, which can incorporate both global and local information.

In this paper, we unify localized relationships in user-item subgroups and common associations among all users and items to improve the recommendation accuracy. The most related works are Probabilistic Matrix Factorization (PMF) and ACCAMS. In AEPMA, the proposed method can learn global information and local information simultaneously, since we can alternate optimization iteration to obtain of the adaptive sample weight. We use stochastic gradient boosting framework to learn more hidden information of the complex rating matrix. More importantly, In the boosting framework, the ensemble models can enhance the recommendation accuracy and stability.

3 The Proposed AEPMA Model

The structure of rating matrix is more and more complicated. The single framework such as PMF can not accurately predict the rating. So we propose a boosting-based matrix approximation for describing the different information of the rating matrix. Because the user-item rating matrix is represented in a global strategy by PMF, such as the whole rating matrix, which ignore the local structure among rating information. In AEPMA, We can capture sufficient information by combining global rating predictions and local rating predictions. Then a stochastic gradient boosting framework is adopted to produce accurate ratings prediction and enhance the recommendation stability. More importantly, we learn adaptive weight for each predictive rating matrix. Which can sufficiently prevent overfitting. Similar to shrinkage in XGBOOST, the learned weights reduce the influence of prediction in each stage and leave space for finer prediction.

3.1 Global and Local Matrix Approximation

We exploit Global and Local Matrix Approximation (GLMA) which is a new probabilistic model which combined global and local information. More importantly, the user-item rating weight can be learned adaptively. And the rating with most suitable global or local model for each user/item should be with large weights. The conditional distribution over the observed ratings for the global and local model can be given as follows:

$$p(X|\mathbf{U}, \mathbf{V}, \sigma, \alpha, \beta) = \prod_{X_{ij} \in \Omega} [\alpha_i^1 \beta_j^1 N(X_{i,j}|S_{ij}, \sigma^2) + \alpha_i^2 \beta_j^2 N(X_{i,j}|U_i^T V_j, \sigma^2)] \quad (1)$$

Where S is the prediction rating by local method ACCAMS, And U, V are the global user, item latent feature vectors, which is inferred from all user-item

rating matrix. And α^1, β^1 are the weight vectors of the local model for all user-item ratings, respectively, and accordingly α^2, β^2 are the weight vectors of the global model for all user-item ratings, Thus, α_i^1, β_j^1 reflect the weights of the local model for the i^{th} user and j^{th} item. The local predictions that reflected the unique interests shared among only subsets of users/items should be with large weights, α_i^2, β_j^2 denote the weights of the globally optimized model, the ratings that reflect the overall structures should be with large weight.

For $\alpha^1, \beta^1, \alpha^2, \beta^2$, we choose a Laplacian prior here, because the models with most suitable global or local model for user-item ratings should be with large weight, the variable should be sparse. More importantly, the adaptive weight can make the model learn useful information and avoid overfitting. Thus the log of the posterior distribution over the user and item features and weights can be given as follows:

$$\ln p(U, V, \alpha, \beta | X, \sigma_U, \sigma_V, \sigma, u_\alpha, u_\beta, b_\alpha, b_\beta) \quad (2)$$

$$\propto \ln [p(X | U, V, \sigma, \alpha, \beta) p(U | \sigma_U) p(V | \sigma_V) p(\alpha | u_\alpha, b_\alpha) p(\beta | u_\beta, b_\beta)]$$

Where u_α, u_β are the location parameter of the Laplacian distribution, and accordingly b_α, b_β are the scale parameter of the Laplacian distribution. Unfortunately, it is very difficult to solve the above optimization problem directly. In order to simplify the model, we try to obtain the approximate solution using Jensen's inequality, the lower bound of Eq. (2) can be obtained as follows:

$$l = \sum_{i=1}^n \sum_{j=1}^m I_{ij} [\ln \alpha_i^1 \beta_j^1 N(X_{i,j} | S_j, \sigma^2) + \ln \alpha_i^2 \beta_j^2 N(X_{i,j} | U_i^T V_j, \sigma^2)] \quad (3)$$

$$- \frac{1}{2\sigma_u^2} \|U\|_F^2 - \frac{1}{2\sigma_v^2} \|V\|_F^2 - n \ln \sigma_u^2 - m \ln \sigma_v^2$$

$$- \frac{1}{b_\alpha} \sum_{k=1}^2 \sum_{i=1}^n |\alpha_i^k - u_\alpha| - \frac{1}{b_\beta} \sum_{k=1}^2 \sum_{j=1}^m |\beta_j^k - u_\beta| - n \ln b_\alpha^2 - m \ln b_\beta^2$$

If we keep the hyperparameters of the prior distribution fixed may easily lead to overfitting. And we want to obtain the adaptive weight of the model, so we estimate the parameters and hyperparameters simultaneously during model training. In order to estimate the hyperparameters, while fixed the rest variables and then iterate until convergence. The hyperparameters can be given as:

$$\sigma^2 = \frac{\sum_{X_{ij} \in \Omega} \alpha_i \beta_j (X_{ij} - R_{ij})^2}{\sum_{X_{ij} \in \Omega} 1} \quad (4)$$

$$\sigma_u^2 = \frac{1}{n} \sum_{X_{ij} \in \Omega} (U_i)^2 \quad \sigma_v^2 = \frac{1}{m} \sum_{X_{ij} \in \Omega} (V_j)^2$$

$$u_\alpha = \frac{1}{n} \sum_{X_{ij} \in \Omega} \alpha_i \quad u_\beta = \frac{1}{m} \sum_{X_{ij} \in \Omega} \beta_j$$

$$b_\alpha = \frac{1}{n} \sum_{X_{ij} \in \Omega} |\alpha_i - u_\alpha| \quad b_\beta = \frac{1}{m} \sum_{X_{ij} \in \Omega} |\beta_j - u_\beta|$$

3.2 Boosting-Based Matrix Approximation

The structure of rating matrix is more and more complicated, the single framework has trouble discovering abundant hidden information of the rating matrix.

Thus, we propose a boosting-based mixture matrix approximation model- Adaptive Ensemble Probabilistic Matrix Approximation (AEPMA).

In order to describe the different information of the rating matrix, We propose an ensemble mixture matrix approximation approach for rating prediction. In AEPMA model, we learn an additive model X , with K products $\omega_k * R^k$. Thus the prediction rating matrix \hat{X} is presented:

$$\hat{X} = \sum_{k=1}^K \omega_k * R^k \tag{5}$$

Where K is the number of individual learner, R^k is the prediction rating matrix of k^{th} individual learner. $R^k = (U^k)^T V^k$, and ω_k is the weight of prediction rating matrix R^k . And (U^k, V^k) is the pair of user, item latent factor vectors.

In order to discover the global structure information and detect local strong association. We let the first learner is GLMA, so we can get the prediction rating S^1 , and the other individual learner corresponds to PMF. Thus the optimal prediction rating value is then equal to:

$$\hat{X} = \omega_1 * S^1 + \sum_{k=2}^K \omega_k * R^k \tag{6}$$

To achieve the rating matrix approximation, we use the Frobenius norm-based objective function as follows:

$$\min_{U_k, V_k} \left\| X - \omega_1 * S^1 - \sum_{k=2}^K \omega_k * U_k^T V_k \right\|_F^2 \tag{7}$$

Residual Matrix Update. AEPMA solves this problem in a gradient boosting manner, which iteratively adds a new individual learner to better approximate the true rating matrix. The partial residual rating matrix is learn from the negative gradient of the loss function. In AEPMA, the negative gradient of the loss function is the difference of the true ratings and the prediction ratings.

To fit the $k - 1$ learner PMF, $R^{k-1} = \omega_{k-1} * U_{k-1}^T V_{k-1}$, with rank r_{k-1} to the residual matrix X^{k-1} . Where the matrix rank r_{k-1} is adaptive, because the distribution of the residual is different. Then The specific residual matrix X^k calculation method is shown in Fig. 1:

Due to the forward stage-wise manner, We constantly iterative add a new model to better approximate rating matrix X . The prediction from previously learned $k - 1$ models is fixed, Thus the k^{th} residual rating matrix can be indicated by the previously learned $k - 1$ models. Thus, we can define the residual rating matrix at stage k as:

$$X^k = \begin{cases} X & \text{if } i = 1 \\ \frac{X - \omega_1 S^1}{\omega_2} & \text{if } i = 2 \\ \frac{X^{k-1} - \omega_{k-1} R^{k-1}}{\omega_k} & \text{if } i \geq 3 \end{cases} \tag{8}$$

<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>0.1</td><td>4.8</td><td>3.3</td><td>4.6</td></tr> <tr><td>1.2</td><td>2.7</td><td>0.8</td><td>4.9</td></tr> <tr><td>2.5</td><td>1.3</td><td>2.9</td><td>1.5</td></tr> <tr><td>3.9</td><td>0.9</td><td>3.4</td><td>3.1</td></tr> <tr><td>4.5</td><td>3.1</td><td>4.1</td><td>2.1</td></tr> </table>	0.1	4.8	3.3	4.6	1.2	2.7	0.8	4.9	2.5	1.3	2.9	1.5	3.9	0.9	3.4	3.1	4.5	3.1	4.1	2.1	-	<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>0.1</td><td>4.3</td><td>2.7</td><td>2.8</td></tr> <tr><td>0.9</td><td>1.9</td><td>0.2</td><td>4.5</td></tr> <tr><td>1.2</td><td>1.2</td><td>2.9</td><td>1.3</td></tr> <tr><td>3.3</td><td>0.9</td><td>3.0</td><td>2.5</td></tr> <tr><td>3.9</td><td>2.5</td><td>2.1</td><td>2.0</td></tr> </table>	0.1	4.3	2.7	2.8	0.9	1.9	0.2	4.5	1.2	1.2	2.9	1.3	3.3	0.9	3.0	2.5	3.9	2.5	2.1	2.0	=	<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>0.0</td><td>0.5</td><td>0.5</td><td>1.8</td></tr> <tr><td>0.3</td><td>0.8</td><td>0.6</td><td>0.4</td></tr> <tr><td>1.3</td><td>0.1</td><td>2.9</td><td>0.2</td></tr> <tr><td>0.6</td><td>0.0</td><td>0.4</td><td>0.6</td></tr> <tr><td>0.6</td><td>0.6</td><td>2.0</td><td>0.1</td></tr> </table>	0.0	0.5	0.5	1.8	0.3	0.8	0.6	0.4	1.3	0.1	2.9	0.2	0.6	0.0	0.4	0.6	0.6	0.6	2.0	0.1
0.1	4.8	3.3	4.6																																																													
1.2	2.7	0.8	4.9																																																													
2.5	1.3	2.9	1.5																																																													
3.9	0.9	3.4	3.1																																																													
4.5	3.1	4.1	2.1																																																													
0.1	4.3	2.7	2.8																																																													
0.9	1.9	0.2	4.5																																																													
1.2	1.2	2.9	1.3																																																													
3.3	0.9	3.0	2.5																																																													
3.9	2.5	2.1	2.0																																																													
0.0	0.5	0.5	1.8																																																													
0.3	0.8	0.6	0.4																																																													
1.3	0.1	2.9	0.2																																																													
0.6	0.0	0.4	0.6																																																													
0.6	0.6	2.0	0.1																																																													
X^{k-1}		R^{k-1}		X^k																																																												

Fig. 1. The residual matrix determined by rating matrix and predictive rating matrix in previous stage.

where ω_k is the weight of k^{th} prediction rating matrix R^k . S^1 is the prediction rating matrix which fitting the GLMA model. And accordingly R^{k-1} is the prediction rating matrix fitting the PMF model, $R^{k-1} = (U^{k-1})^T V^{k-1}$. And (U^{k-1}, V^{k-1}) is the pair of user, item factor latent vectors. Then the input residual rating matrix X^k of the k^{th} individual learner PMF can be written as;

$$X^k = \frac{[[[X - \omega_1 S^1] - \omega_2 R^2] \cdots - \omega_{k-1} R^{k-1}]]}{\omega_k} \tag{9}$$

In the k^{th} epoch, according to the Probabilistic Matrix Factorization(PMF) model, we can obtain the user/item factor latent vectors. In our proposed method solves each model of R in a greedy sequential manner, which means that once the solution for R^k is obtained at stage k, it is fixed during the remaining iterations. And in our model, we want to consider the local and global information simultaneously, so the general we choose more than three models.

Adaptive Weight. One important step in the approximate algorithm is to propose adaptive weight. In AEPMA, we assign smaller weight to those components R less explained (large residuals). Let us define the residual probability distribution $P_{ij}^k R_{ij}^k - \hat{X}_{ij}^k \sim N(0, \sigma_u^2)$, Then large residuals is far from the mean, in which the corresponding probability is relatively small. Thus the weight of each prediction rating matrix is given by;

$$w_k = \frac{1}{N} \sum_{i,j} \frac{P_{ij}^k}{\sum_{s=1}^K P_{ij}^s} \tag{10}$$

In the above equation, higher weight values is assigned to components with smaller residual. In other word, The better the fitting rating matrix, the greater the corresponding weight.

In APEMA, each user-item rating is characterized by a mixture model, and then to predict user-item ratings by the mixture components and the weight of each model. We can predict the user-items ratings as follows:

$$\hat{X} = \omega_1 * R^1 + \omega_2 * R^2 + \omega_3 * R^3 \cdots + \omega_k * R^k \tag{11}$$

4 Experiments

4.1 Experiment Setup

In the following, we introduce our experimental setup include dataset, baseline methods, and evaluation measures.

Datasets. We selected the following three real-world datasets that has widely used for evaluating recommendation algorithm – Ciao, Epinions, and Douban which are usually used in literatures. The rating score is from 1 to 5 score. For each datasets, we randomly split it into five equal sized subsets. Four subsets are used as training set and the left one as testing set in each fold. In the five-fold cross-validation, the result are represented by averaging the results over five different train-test splits. These datasets are summarized in Table 1.

Table 1. Summary of experimental datasets

Dataset	Ciao	Epinions	Douban
#users	7,375	49,290	129,490
#items	106,797	139,738	58,541
#ratings	284,086	284,086	16,830,839
Rating density	0.036%	0.010%	0.222%

Baselines. We compared the recommendation accuracy of our proposed method against various state-of-the-art methods, including PMF [2], BPMF [5], LLORMA [3], WEMAREC [4], ACCAMS [1], SMA [25]. Because in the paper (Low-Rank Matrix Approximation with Stability), the author proposed that the performance of SMA is better than BPMF, LLORMA and WEMAREC. Thus the proposed method (AEPMA) is compared against three state-of-the-art matrix approximation based CF models, which are described as follows:

- **PMF:** A probabilistic matrix factorization, which define the conditional distribution of the observed ratings as Gaussian distribution.
- **RSVD:** A global-based matrix factorization method, in which user/item features are estimated by minimizing the sum-squared error.
- **ACCAMS:** An additive co-clustering model to approximate rating matrix, which can partition rating matrix into blocks that are highly similar through a clustering of the rows and columns. Then using the mean of the values to represent the block missing ratings.
- **SMA:** An low-rank matrix approximation framework, which achieving high stability.

Metrics. The root mean square error (RMSE) and Mean Absolute Error (MAE) is adopted as the evaluation metric for recommendation accuracy. The RMSE is

$$\text{defined as } RMSE = \sqrt{\frac{\sum_{x_{ij} \in T} (X_{ij} - \hat{R}_{ij})^2}{|T|}}.$$

where T is the set of ratings in the testing set and $|T|$ is the size of the test ratings. \hat{R}_{ij} is the predicted rating X_{ij} is represented the true rating value from i^{th} user to j^{th} item in the testing set. The MAE is defined by $MAE = \frac{1}{|T|} \sum_{X_{ij} \in T} |X_{ij} - \hat{R}_{ij}|$.

4.2 Recommendation Performance

Table 2 compares RMSE and MAE in our method with classic matrix approximation method. We can see our method can achieve both lower generalization error and lower expected risk than other methods.

In this experiment, we compare the recommendation accuracy of AEPMA against various state-of-the-art methods, including PMF, RSVD, ACCAMS and SMA. In most of these methods, we use the same parameters values provided in the original papers, and for ACCAMS, we tuned its parameters including the number of users clusters and item cluster, and the number of stencils. In PMF, we set the max-number of iterations as 300 in our experiment. And the regularization parameter on latent is 0.01. In AEPMA, in order to reduce the manual setting of the learning rate, we use adam for stochastic optimization. And we choose 0.001 as stepsize, 0.9, 0.999 as the exponential decay rates for the moment estimates. Then we set the number of individual learners as three. The relative improvements that AEPMA achieves relative to four state-of-the-art methods on three datasets are calculated. As shown in Fig. 2. Obviously, AEPMA performs better than ACCAMS and SMA, which demonstrates that the model with global structure information is better than the only local ensemble matrix approximation methods. Simultaneously, Our method is much better than the only global method PMF and RSVD. From the relative improvements, we can see the SMA is better on the dense dataset (Douban) and perform poor on the sparse datasets (Ciao and Epinions). More importantly, In order to prove that the importance of GLMA method, we compare the performance in terms of MAE, RMSE for PMF+ (global) and ACCAMS+ (local). In the boosting framework, PMF+ is fitting PMF then get S^1 , accordingly ACCAMS+ is fitting ACCAMS to get S^1 . In additional, our method which using global and local information is better than the method only global on local. And we also find that the model can achieve relatively stable prediction accuracy due to the framework of boosting. A smaller RMSE or MAE value indicate better performance. Because there are too many ratings, a small improvement in RMSE or MAE can have a significant impact on the recommendation result. As shown in Table 2, It can be seen that AEPMA consistently outperforms the global method (PMF, RSVD) and the local method (ACCAMS, SMA), Which means that considering both local and global information is more useful than only considering unilateral influence.

The true datasets have different rating density, For example, The Ciao and Epinions (the rating density is 0.036% and 0.010%) is sparse. Simultaneously, We can see in the Table 2, the recommendation accuracy on sparse dataset is worse than the dense datasets. Thus how to improve the recommendation accuracy of sparse data, is the challenge of the recommendation system. More importantly, ACCAMS is better than SMA on the Ciao and Epinions, but worse than SMA on

Table 2. RMSE and MAE comparison of different methods

Datasets	Metrics	PMF	RSVD	ACCAMS	SMA	PMF+	ACCAMS+	AEPMA
Ciao	RMSE	1.1146	1.4268	1.0540	1.0746	1.0642	1.0339	1.0121
	MAE	0.8256	1.0745	0.8084	0.8175	0.8130	0.7846	0.7788
Epinions	RMSE	1.3203	1.4772	1.1689	1.1847	1.1710	1.1406	1.1118
	MAE	1.1206	1.1411	0.8971	0.9157	0.9112	0.8875	0.8597
Douban	RMSE	0.7699	0.7360	0.7309	0.7092	0.7098	0.7261	0.7038
	MAE	0.6230	0.5752	0.5818	0.5594	0.5635	0.5779	0.5574

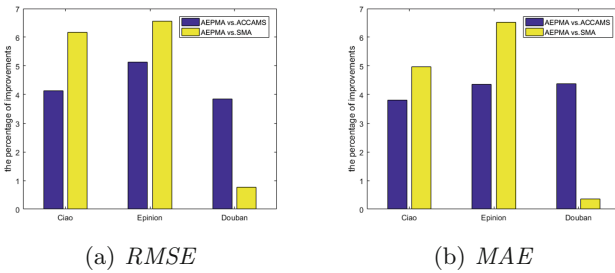


Fig. 2. The relative improvements of AEPMA vs. ACCAMS and SMA in three datasets in terms of (a) RMSE and (b) MAE

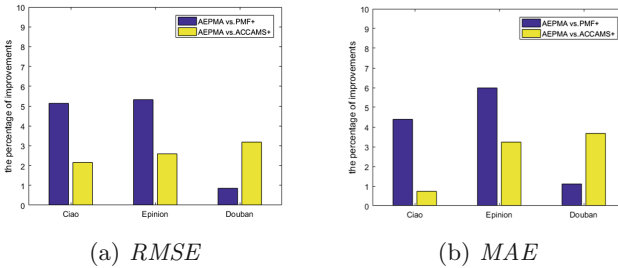


Fig. 3. The relative improvements of AEPMA vs. PMF+ and ACCAMS+ in three datasets in term of (a) RMSE and (b) MAE

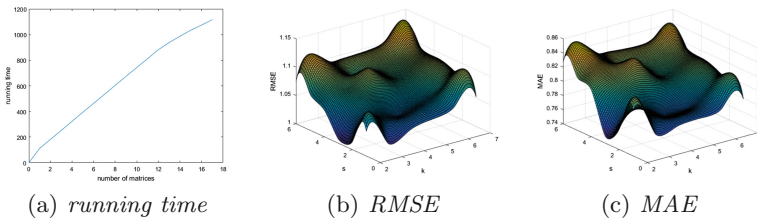


Fig. 4. Effect of Parameters k , s and matrices number on AEPMA

the Douban dataset. In AEPMA, we exploit Global and Local information, and use the boosting framework to learn the hidden information. Thus the proposed AEPMA can outperform on both sparse and dense datasets. Table 2 show that the ensemble-based local methods (ACCAMS, SMA) especially outperforms the global method (PMF, RSVD). Thus we pay attention to the relative improvements that AEPMA achieves to two local baselines on three datasets, as shown in Fig. 2. Obviously, AEPMA performs better than ACCAMS and SMA, which demonstrates the global information benefit the AEPMA model. Figure 2 also reflects that the sparsity of data influence the recommendation accuracy. The relative improvements that AEPMA achieves relative to SMA on sparse datasets (Ciao, Epinions), is superior to the relative improvement to ACCAMS. In other word, The ACCAMS performs better than SMA on sparse dataset, but worse on dense dataset. Because the ACCAMS use the mean of values to the block can lead to overfitting on the dense datasets. In AEPMA, we exploit Global and Local information can fully learn the complicated ratings. More importantly, the boosting framework can improve the model stability and robustness. In the global model such as PMF and RSVD, the same vectors of latent factors inferred from all user-item rating matrix is adopted to describe all users and items, However in many real-world user-item rating matrices, if we think of the global latent factors as “common interests”, then subset of users may share “unique interests” that are not reflected by the “common interest”. Thus, Fig. 3 investigates the effect of global and local information in our model. We fix the boosting framework and change the S^1 . We can see AEPMA is better than PMF+ and ACCAMS+, and ACCAMS+ performs better than PMF+. Because ACCAMS+ trained the model by both global (boosting) and local (S^1) information. But AEPMA can learn the sample weight adaptively to learn sufficient information. Thus AEPMA is superior to ACCAMS+.

Figure 4(a) analyzes the running time with the number of matrices increases. The method AEPMA based on boosting can reduce the bias. And with the number of iterations increasing, the RMSE and MAE can decrease gradually, but running time increases. So in this experiment, we choose a compromise method, the number of matrices is smaller than five. Figure 4(b, c) analyzes the impact of clustering method with different numbers of clusters k and stencils s on Ciao dataset. From Fig. 4(b, c), it can be seen that the performances is destroyed when s is large. The main reason is that large stencils will make overfitting. Meanwhile, we discover that AEPMA is stable under varying k with fixing s . Thus small k is enough to approximate the rating matrix.

5 Conclusions

Traditional matrix approximation based collaborative filtering methods have a major drawback that they perform poorly at detecting strong associations among a small set of closely related items. In this paper, we can capture sufficient information by combining global and local information. More importantly, by placing a Laplacian prior on the user and item weight vectors, we can adaptively

learn the sample weight. In the stochastic gradient boosting framework, we can learn the hidden information and enhance the recommendation accuracy and scalability. Experimental study on three real-world datasets demonstrates that proposed AEPMA method can outperform several state-of-art ensemble matrix approximation methods.

References

1. Beutel, A., Ahmed, A., Smola, A.J.: ACCAMS: additive co-clustering to approximate matrices succinctly. In: Proceedings of International Conference on World Wide Web, pp. 119–129 (2016)
2. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: Proceedings of International Conference on Machine Learning, pp. 880–887 (2007)
3. Lee, J., Kim, S., Lebanon, G., Singer, Y.: Local low-rank matrix approximation. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 82–90 (2013)
4. Chen, C., Li, D., Zhao, Y., Lv, Q., Shang, L.: WEMAREC: accurate and scalable recommendation through weighted and ensemble matrix approximation. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015), pp. 303–312 (2015)
5. Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In: Proceedings of the 25th International Conference on Machine Learning (ICML 2008), pp. 880–887. ACM (2008)
6. Chen, C., Li, D., Lv, Q., Yan, J., Chu, S.M., Shang, L.: MPMA: mixture probabilistic matrix approximation for collaborative filtering. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), pp. 1382–1388 (2016)
7. Li, D., Chen, C., Lv, Q., Yan, J., Shang, L., Chu, S.: Low-rank matrix approximation with stability. In: Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), pp. 295–303 (2016)
8. Srebro, N., Jaakkola, T.: Weighted low-rank approximations. In: Proceedings of the 20th International Conference on Machine Learning (ICML 2003), pp. 720–727 (2003)
9. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **54**(11), 4311–4322 (2006)
10. Jing, L., Wang, P., Yang, L.: Sparse probabilistic matrix factorization by Laplace distribution for collaborative filtering. In: Proceedings of the International Conference on Artificial Intelligence, pp. 1771–1777 (2015)
11. Lee, J., Kim, S., Lebanon, G., Singer, Y.: Local low-rank matrix approximation. In: Proceedings of the 30th International Conference on Machine Learning, pp. 82–90 (2013)
12. Mackey, L.W., Jordan, M.I., Talwalkar, A.: Divide-and-conquer matrix factorization. In: Advances in Neural Information Processing Systems, pp. 1134–1142 (2011)
13. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: Knowledge Discovery and Data Mining KDD, pp. 426–434 (2008)
14. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999), pp. 230–237 (1999)

15. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web (WWW 2001), pp. 285–295 (2001)
16. Zhang, Y., Zhang, M., Liu, Y., Ma, S.: Improve collaborative filtering through bordered block diagonal form matrices. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013), pp. 313–322 (2013)
17. Lawrence, N.D., Urtasun, R.: Non-linear matrix factorization with Gaussian processes. In: Proceedings of the International Conference on Machine Learning (2009)
18. Mirbakhsh, N., Ling, C.X.: Clustering-based matrix factorization. ArXiv Report [arXiv:1301.6659](https://arxiv.org/abs/1301.6659) (2013)
19. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159 (2011)
20. Moulines, E., Bach, F.R.: Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Advances in Neural Information Processing Systems, pp. 451–459 (2011)
21. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. *SIAM J. Control. Optim.* **30**(4), 838–855 (1992)
22. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 1139–1147 (2013)
23. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. arXiv preprint [arXiv:1212.5701](https://arxiv.org/abs/1212.5701) (2012)
24. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562 (2001)
25. Li, D., Chen, C., Lv, Q., Yan, J., Shang, L., Chu, S.M.: Low-rank matrix approximation with stability (2017)



A Deep Structure-Enforced Nonnegative Matrix Factorization for Data Representation

Yijia Zhou¹ and Lijun Xu²(✉)

¹ Dalian Neusoft University of Information,
Dalian 116023, Liaoning, People's Republic of China
zhouyijia@neusoft.edu.cn

² Dalian Maritime University,
Dalian 116026, Liaoning, People's Republic of China
lijun_xu@dlmu.edu.cn

Abstract. In this paper, we focus on a deep structure-enforced non-negative matrix factorization (DSeNMF) which represents a large class of deep learning models appearing in many applications. We present a unified algorithm framework, based on the classic alternating direction method of multipliers (ADMM). For updating subproblems, we derive an efficient updating rule according to its KKT conditions. We conduct numerical experiments to compare the proposed algorithm with state-of-the-art deep semi-NMF. Results show that our algorithm performs better and our deep model with different sparsity imposed indeed results in better clustering accuracy than single-layer model. Our DSeNMF can be flexibly applicable for data representation.

Keywords: Deep matrix factorization · Alternating direction method
Data representation

1 Introduction

Matrix factorization techniques have found great utility in various data-related applications, such as in signal and image processing and in machine learning tasks, primarily because they often help reveal latent features in a dataset. In recent years, Non-negative Matrix Factorization (NMF) is a widely-used method for finding meaningful representations of nonnegative data and has been proven useful in dimension reduction of images, text data and signals, for example. The family of NMF algorithms has been successfully applied to a variety of areas, like environmetrics [1], microarray data analysis [2, 3], document clustering [4], face recognition [5, 6], speech recognition [7], hyperspectral image unmixing [8, 9], blind audio source separation [10], etc. Moreover, NMF has been extended into

Supported by the Fundamental Research Funds for the Central Universities (3132018218).

a number of variant forms, allowing for various structures or regularized models, most of which demonstrate distinct advantages in local feature extraction or data representation learning.

The work of Lee and Seung [11] demonstrates that NMF models tend to return part-based sparse representations of data, which has popularized the use of and research on NMF-related techniques. In particular, various NMF-inspired formulations add different regularization or penalty terms to promote desired properties, such as sparsity patterns or orthogonality in addition to nonnegativity (see [12–18], for example). Besides, graph-regularized NMF versions have also been explored. For example, Cai et al. [19] proposed a graph-regularized NMF by incorporating prior information of samples into the typical NMF. This helps to keep the original topological structure of data after being projected into a subspace and usually leads to better clustering results.

Semi Non-negative Matrix Factorization (Semi-NMF) [20], as one of the most popular variants of NMF, was proposed to extend NMF by relaxing the factorized basis matrix to be real values. This practice allows Semi-NMF to learn new lower-dimensional features from the data that have a convenient clustering interpretation and have a wider application in the real world than traditional NMF. Moreover, it has shown that it is equivalent to k-means clustering, and that in fact, this NMF variants are expected to perform better than k-means clustering particularly when the data is not distributed in a spherical manner.

Although there have been extensive variants of NMF, most of them remain to be single-layer models, hence can only capture one level of data features. Most recently, deep learning is becoming increasingly popular and has been demonstrated to be powerful in learning data representation. Inspired by the success of training deep architectures, Multi-layer NMF (see [21, 22] for example), Deep Semi-NMF [23], Deep Orthogonal NMF [24], Sparse Deep NMF [25], Deep Non-smooth NMF [26], etc. have been proposed by stacking one-layer variants of NMF into multiple layers to learn hierarchical relationships among features or hierarchical projections. Since these deep (multi-layer) models can extract high level data representations and yield intuitive interpretations for features generated in each layer, they have been successfully applied to many areas, such as recommender systems [27], image clustering [28], neural network [29], speech separation [30], matrix completion [31], for example. However, these models are only designed for specific problems with certain intuitive structures. In this paper, we focus on a unified deep structure-enforced NMF in data representation, which imposing desired properties (like sparsity, orthogonality, for example) in addition to nonnegativity. A specific algorithmic approach to solve the deep structure-enforced NMF is further studied and can be applicable to a range of easily projectable structures.

This paper is organized as follows. In Sect. 2, we introduce the deep structure-enforced NMF (DSeNMF) and propose a new ADMM-based algorithm framework for solving DSeNMF. Section 3 contains several numerical experiments comparing the proposed algorithm with Deep Semi-NMF and single-layer matrix factorization on MNIST digit dataset. Finally, we conclude this paper in Sect. 4.

2 Deep Structure-Enforced Nonnegative Matrix Factorization Model

The general structured-enforced matrix factorization (SeMF) model (1) is firstly proposed in the earlier work in [32]. That is, decomposing a given data matrix $M \in \mathbb{R}^{p \times n}$ into two factors $Z \in \mathbb{R}^{p \times k}$ and $H \in \mathbb{R}^{k \times n}$ which belong to \mathcal{Z} and \mathcal{H} , respectively,

$$\min_{Z, H} \frac{1}{2} \|M - ZH\|_F^2 \quad \text{s.t. } Z \in \mathcal{Z}, H \in \mathcal{H}, \tag{1}$$

where $\|\cdot\|_F$ is Frobenius norm, and \mathcal{Z} and \mathcal{H} are subsets of $\mathbb{R}^{p \times k}$ and $\mathbb{R}^{k \times n}$, respectively. Obviously, the model (1) is a single-layer matrix factorization. Thus, it can only do one-layer feature extraction even utilizing more structures. In practice, it is common that complex data objects have hierarchical features, each of which denotes a different level of abstract understanding of the objects. It is therefore meaningful to develop corresponding models with a deep architecture, which allows to discover the hierarchy of data. It is well known that NMF is widely used both in single-layer and in multi-layer data representation. To this end, we propose a deep structure-enforced version for nonnegative matrix factorization by extending model (1).

Similar to the general multi-layer framework, the Deep Structure-enforced NMF (DSeNMF) model is presented to factorize $M \in \mathbb{R}^{p \times n}$ into the multiplier of $m + 1$ nonnegative matrices, as follows:

$$\min_{\{Z_i \geq 0\}_{i=1}^m, H_m \geq 0} \frac{1}{2} \|M - Z_1 Z_2 \cdots Z_m H_m\|_F^2 \quad \text{s.t. } Z_i \in \mathcal{Z}_i, H_m \in \mathcal{H}, \tag{2}$$

where $Z_1 \in \mathbb{R}^{p \times k_1}$, $\{Z_i \in \mathbb{R}^{k_{i-1} \times k_i}\}_{i=2}^m$, $H_m \in \mathbb{R}^{k_m \times n}$, $\{\mathcal{Z}_i\}_{i=1}^m$ and \mathcal{H} are structure subsets with proper dimensions. In our model, prior knowledge are explicitly enforced as constraint sets $\{\mathcal{Z}_i\}_{i=1}^m$ and \mathcal{H} whose members possess desirable matrix structures allowing “easy projection”. In practice, the most useful structures of this kind include, but are not limited to, nonnegativity, normality and various sparsity patterns. Many deep NMF models can be represented by the DSeNMF (2) with different structure constraints, see Sparse Deep NMF, Deep Orthogonal NMF, Deep Semi-NMF as mentioned above, for example.

To make it more intuitive, one can split the model (2) into the following factorizations:

$$\begin{aligned} M &\approx Z_1 H_1, \\ H_1 &\approx Z_2 H_2, \\ &\vdots \\ H_{m-1} &\approx Z_m H_m, \end{aligned} \tag{3}$$

where $\{Z_i\}_{i=1}^m$ and $\{H_i\}_{i=1}^m$ satisfy proper constraints, respectively. This formulation can intuitively illustrate that deep model (2) allows for a hierarchy of m layers of implicit representations of data. In other words, not only most multi-layer and deep matrix factorizations is derived from the formulation (3), but

also most algorithms for (2) are designed by solving (3) layer by layer. In the beginning of approaches, the objective data matrix are multi-factorized only by solving (3) one round layer by layer. Obviously, these approaches are inefficient since the factor matrices in former layers are useless for subsequent layer factorizations. Therefore, the popular scheme is utilizing the layer by layer technique as initialization or pre-training, then fine-tuning all layers by alternating updating factor matrices one by one. Now, we propose a novel approach based on alternating direction algorithm framework to solve the non-convex problem (2).

2.1 An Alternating Direction Algorithm for the Proposed DSeNMF

As introduced in the work [32,33], an alternating direction and projection method solves single layer structure-enforced matrix factorization (SeMF) efficiently. Motivated by the algorithms in [32,33], we propose a novel way to tackle multi-layer or deep matrix factorizations. To facilitate an efficient use of alternating minimization, we introduce auxiliary variables $\{U_i\}_{i=1}^m$ and V_m in order to separate $\{Z_i\}_{i=1}^m$ and H_m from structure constraints $\{Z_i\}_{i=1}^m$ and \mathcal{H} , respectively. Consider the following model equivalent to (2),

$$\begin{aligned} \min_{\{Z_i \geq 0, U_i\}_{i=1}^m, H_m \geq 0, V_m} & \frac{1}{2} \|M - Z_1 Z_2 \cdots Z_m H_m\|_F^2 \\ \text{s.t. } & Z_i - U_i = 0, U_i \in \mathcal{Z}_i, i = 1, \dots, m, \\ & H_m - V_m = 0, V_m \in \mathcal{H}, \end{aligned} \quad (4)$$

where $\{U_i\}_{i=1}^m$ and V_m have the same dimension size with $\{Z_i\}_{i=1}^m$ and H_m , respectively. The augmented Lagrangian function of (4) is

$$\begin{aligned} & \mathcal{L}_A(\{Z_i, U_i, A_i\}_{i=1}^m, H_m, V_m, \Pi) \\ &= \frac{1}{2} \|M - Z_1 Z_2 \cdots Z_m H_m\|_F^2 + \\ & \quad \sum_{i=1}^m A_i \bullet (Z_i - U_i) + \Pi \bullet (H_m - V_m) \\ & \quad + \sum_{i=1}^m \frac{\alpha_i}{2} \|Z_i - U_i\|_F^2 + \frac{\beta}{2} \|H_m - V_m\|_F^2, \end{aligned} \quad (5)$$

where $\{A_i\}_{i=1}^m, \Pi$ are Lagrangian multipliers with equal-size of $\{Z_i\}_{i=1}^m, H_m$, respectively, and $(\{\alpha_i\}_{i=1}^m, \beta) \geq 0$ are penalty parameters for equality constraints, respectively. Note that the scalar product “ \bullet ” of two equal-size matrices X and Y is the sum of all element-wise products, i.e., $X \bullet Y = \sum_{i,j} X_{ij} Y_{ij}$.

The alternating direction method of multiplier (ADMM) [34,35] for (4) is derived by successively minimizing the augmented Lagrangian function \mathcal{L}_A with respect to $\{Z_i\}_{i=1}^m, H_m, \{U_i\}_{i=1}^m$ and V_m , one at a time while fixing others at their most recent values, and then updating the multipliers after each sweep of such alternating minimization. The introduction of the auxiliary variables $\{U_i\}_{i=1}^m$ and V_m makes it easy to carry out each of the alternating minimization steps. Specifically, these steps can be written in the following forms,

$$Z_j^+ \approx \arg \min_{Z_j \geq 0} \mathcal{L}_A(\{Z_i, U_i, \Lambda_i\}_{i=1}^m, H_m, V_m, \Pi), j = 1, 2, \dots, m, \tag{6a}$$

$$H_m^+ \approx \arg \min_{H_m \geq 0} \mathcal{L}_A(\{Z_i^+, U_i, \Lambda_i\}_{i=1}^m, H_m, V_m, \Pi), \tag{6b}$$

$$U_j^+ = \mathcal{P}_{Z_j}(Z_j^+ + \Lambda_j/\alpha_j), j = 1, 2, \dots, m, \tag{6c}$$

$$V_m^+ = \mathcal{P}_{\mathcal{H}}(H_m^+ + \Pi/\beta), \tag{6d}$$

$$\Lambda_j^+ = \Lambda_j + \alpha_j(Z_j^+ - U_j^+), j = 1, 2, \dots, m, \tag{6e}$$

$$\Pi^+ = \Pi + \beta(H_m^+ - V_m^+). \tag{6f}$$

where \mathcal{P}_{Z_j} ($\mathcal{P}_{\mathcal{H}}$) stands for the projection onto the set Z_j (\mathcal{H}) in Frobenius norm, and the superscript “+” is used to denote iterative values at the new iteration.

Updating Rule for Z_j . We fix the rest of the factor matrices and minimize the cost function with respect to Z_j . The Z_j -updating subproblem (6a) actually can be rewritten as

$$\begin{aligned} \min_{Z_j} \quad & \frac{1}{2} \|M - \Phi_j Z_j \Psi_j\|_F^2 + \Lambda_j \bullet (Z_j - U_j) + \frac{\alpha_j}{2} \|Z_j - U_j\|_F^2 \\ \text{s.t.} \quad & Z_j \geq 0, \end{aligned} \tag{7}$$

where $\Phi_j = Z_1 Z_2 \dots Z_{j-1}$ and $\Psi_j = Z_{j+1} \dots Z_m H_m$. Let Γ be the lagrangian multiplier for constraint $Z_j \geq 0$, the Lagrangian function of (7) is

$$\mathcal{L} = \frac{1}{2} \|M - \Phi_j Z_j \Psi_j\|_F^2 + \Lambda_j \bullet (Z_j - U_j) + \frac{\alpha_j}{2} \|Z_j - U_j\|_F^2 + \Gamma \bullet Z_j.$$

The partial derivative of \mathcal{L} with respect to Z_j is

$$\frac{\partial \mathcal{L}}{\partial Z_j} = \Phi_j^T \Phi_j Z_j \Psi_j \Psi_j^T + \alpha_j Z_j - \Phi_j^T M \Psi_j^T - \alpha_j U_j + \Lambda_j + \Gamma.$$

Using the Karush-Kuhn-Tucker (KKT) conditions $\Gamma_{ik} Z_{j ik} = 0$, we get the following equations respect to the (i, k)-th element:

$$(\Phi_j^T \Phi_j Z_j \Psi_j \Psi_j^T + \alpha_j Z_j - \Phi_j^T M \Psi_j^T - \alpha_j U_j + \Lambda_j)_{ik} (Z_j)_{ik} = 0.$$

This equation leads to the following updating rule:

$$(Z_j^+)_{ik} = (Z_j)_{ik} \frac{(\Phi_j^T M \Psi_j^T + \alpha_j U_j - \Lambda_j)_{ik}}{(\Phi_j^T \Phi_j Z_j \Psi_j \Psi_j^T + \alpha_j Z_j)_{ik}}, \tag{8}$$

and it can be rewritten as

$$Z_j^+ = Z_j \odot [(\Phi_j^T M \Psi_j^T + \alpha_j U_j - \Lambda_j) \oslash (\Phi_j^T \Phi_j Z_j \Psi_j \Psi_j^T + \alpha_j Z_j)], \tag{9}$$

where \odot and \oslash denote component multiplications and divisions, respectively.

Updating Rule for H_m . We can derive the H_m -updating rule of (6b) in a similar way. We omit the derivative procedure and directly write updating rule for (i, k)-th component of H_m :

$$(H_m^+)_{ik} = (H_m)_{ik} \frac{(\Phi^T M + \beta V_m - \Pi)_{ik}}{(\Phi^T \Phi H_m + \beta H_m)_{ik}}, \tag{10}$$

where $\Phi = Z_1^+ Z_2^+ \cdots Z_m^+$. Namely,

$$H_m^+ = H_m \odot [(\Phi^T M + \beta V_m - \Pi) \oslash (\Phi^T \Phi H_m + \beta H_m)], \tag{11}$$

where \odot and \oslash denote component multiplications and divisions, respectively.

Since we update Z_j and H_m by component multiplications and divisions instead of involving inverse matrices, the dominant computational tasks at each iteration are the matrix multiplications. Therefore, our updating scheme poses much lower complexity than inverting matrices.

Based on the formulas in (6), (9) and (11), we can implement the following ADMM algorithmic framework so long as we can compute the projections in steps (6c) and (6d).

Algorithm 1. ADMM Framework for DSeNMF

Input: M , each layer dimension $k_i, i = 1, \dots, m$, $maxiter > 0$ and $tol > 0$.

Output: $\{Z_i\}_{i=1}^m$ and H_m .

Set $\{\alpha_i\}_{i=1}^m, \beta > 0$.

$H_0 = M$;

for $i = 1$ **to** m **do**

$Z_i, H_i \leftarrow \text{SeMF}(H_{i-1}, k_i)$ \\ Initialization.

end

for $k = 1$ **to** $maxiter$ **do**

Update $(\{Z_i, U_i, A_i\}_{i=1}^m, H_m, V_m, \Pi)$ by the formulas in (6), (9) and (11).

if stopping criterion (12) is met **then**

output $\{Z_i\}_{i=1}^m$ and H_m , and exit.

end

end

We use the following practical stopping criterion: for given tolerance $tol > 0$,

$$\frac{|f_k - f_{k+1}|}{|f_k|} \leq tol, \tag{12}$$

where $f_k = \|X - Z_1^k Z_2^k \cdots Z_m^k H_m^k\|_F$, Z_i^k is the k -th iterate for the variable Z_i , and so on. For the sake of robustness, in our implementation we require that the above condition be satisfied at three consecutive iterations. In other words, we stop the algorithm when data fidelity does not change meaningfully in three consecutive iterations.

3 Experimental Results

In this section we test the proposed model on MNIST dataset to show that our Deep SeNMF is able to learn better high-level representations of data than a single one-layer structure-enforced NMF. In addition, we compare the performance of the proposed DSeNMF with recently Deep Semi-NMF on the task of clustering analysis and consuming time. Note that we consider to impose several sparse constraints on our DSeNMF model (2).

To better understand the proposed model, we introduce three way to impose sparsity on H_m . One is adding sparsity not only during initialization but also in subsequential updating and denote this case as DSeNMF(**sparse**). The other way is imposing sparsity only in step (6d), that is, using standard NMF to initialize each layer matrix, and is denoted as DSeNMF(**semi-sparse**). The last one will not impose sparsity and denote this case as DSeNMF(**no sparse**). To illustrate deep model and single-layer factorization distinct, we also consider single-layer structure-enforced matrix factorization and denote as **SingleSeMF**.

Next, we apply models to the testing data in an unsupervised way to clustering. We opt the digits from 0 to 4 in MNIST which constitute a 784×5139 matrix M . In this test, we choose the number of layers to be 3 and dimension size of each layer is 300, 15 and 50, respectively. Besides, set the maximum number of iteration $maxiter = 500$ and tolerance $tol = 1e-6$. We factorize data matrix M using Deep Semi-NMF (DSemiNMF) in [23], DSeNMF(**sparse**), DSeNMF(**semi-sparse**) and **SingleSeMF**, respectively. Then we cluster columns of the final H_m according to the approach in [23] and output the clustering accuracy as AC.

Table 1. Results comparison with different deep NMF models

Method	DSeNMF (sparse)	DSeNMF (semi-sparse)	DSeNMF (no sparse)	DSemiNMF [23]	SingleSeMF [32]
AC	0.57	0.68	0.48	0.40	0.33
Time(s)	64.18	64.38	66.57	292.54	29.39
RMSE	37.3688	37.3693	37.3676	37.4621	24.0117

In Table 1, we tabulate the average clustering accuracy (AC), average running time (in second) and average root mean square error (RMSE). We see from the table that our deep structure-enforced NMF performs well both in accuracy and in time consuming. It should be note that our algorithm only need about one fifth running time comparing with deep semi-NMF algorithm. In addition, note that the last column in Table 1, we use the SeMF algorithm in [32] to decompose M into multiplication of $Z \in \mathbb{R}^{784 \times 50}$ and $H \in \mathbb{R}^{50 \times 5139}$ which is indeed a single-layer nonnegative matrix factorization. Obviously, **SingleSeMF** obtain the best data fidelity, but get the worst clustering accuracy meanwhile. It confirms that all the DSeNMF models are able to learn better high-level representations of data than a single one-layer structure-enforced NMF. Among

results of our proposed model with three different structure constraints, we note that DSeNMF(**sparse**) and DSeNMF(**semi-sparse**) obtain better clustering results than DSeNMF(**no sparse**) since imposing sparsity on H_m . More interestingly, comparing DSeNMF(**sparse**) with DSeNMF(**semi-sparse**), the former gets lower clustering accuracy even though considering sparsity in initialization. It demonstrates that imposing structure constraints earlier could not obtain a better initialization. It makes sense that some properties in real data should be considered step by step rather than completely utilized at the beginning.

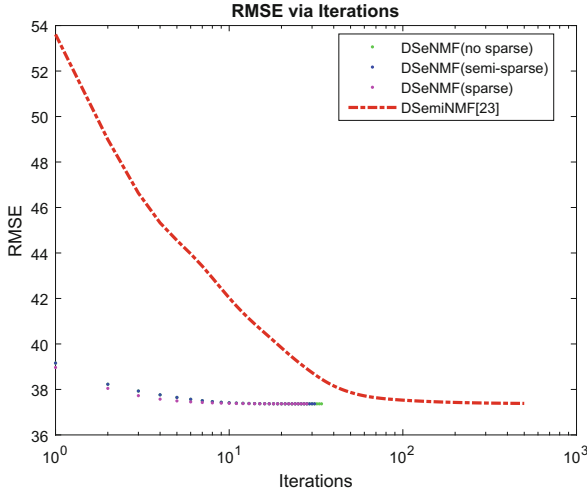


Fig. 1. RMSE comparison with different deep NMF models

Figure 1 presents RMSE curves of four deep models. It shows that our algorithm for solving deep NMF models needs much less (about 50) iterations than the algorithm in [23] (around 500 iterations). It will be evident that our proposed model and algorithm are efficient for the class of deep structured NMF.

4 Conclusion and Future

We have introduced a kind of deep structure-enforced nonnegative matrix factorization and proposed a novel framework for solving the unified model. Although the proposed framework introduces many auxiliary variables, these variables aim to separate from complex structure constraints and split original factor matrices. Further, it can facilitate the obtained model equivalently transformed to an ADMM-applicable model which is easy implemented. Numerical experiments also show the efficiency of the proposed algorithm and the applicable of our deep model for data representing problems.

Although deep structured matrix factorization problems are generally highly nonconvex, they widely and variously exist in real-world applications. Our next step is testing the proposed model and algorithm on more datasets and comparing it with other deep NMF algorithms. Another work will be focusing on how different decomposed dimension would affect clustering performance of deep non-negative matrix factorization.

References

1. Paatero, P., Tapper, U.: Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**(2), 111–126 (1994)
2. Brunet, J.-P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. *PNAS* **101**(12), 4164–4169 (2004)
3. Devarajan, K.: Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol.* **4**(7), e1000029 (2008)
4. Berry, M.W., Browne, M.: Email surveillance using nonnegative matrix factorization. *Comput. Math. Organ. Theory* **11**(3), 249–264 (2005)
5. Zafeiriou, S., Tefas, A., Buciu, I., Pitas, I.: Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *TNN* **17**(3), 683–695 (2006)
6. Kotsia, I., Zafeiriou, S., Pitas, I.: A novel discriminant nonnegative matrix factorization algorithm with applications to facial image characterization problems. *TIFS* **2**(3–2), 588–595 (2007)
7. Zdunek, R., Cichocki, A.: Non-negative matrix factorization with quasi-newton optimization. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) *ICAISC 2006. LNCS (LNAI)*, vol. 4029, pp. 870–879. Springer, Heidelberg (2006). https://doi.org/10.1007/11785231_91
8. Wang, W., Li, S., Qi, H., Ayhan, B., Kwan, C., Vance, S.: Identify anomaly component by sparsity and low rank. In: *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensor (WHISPERS)*, Tokyo, Japan (2015)
9. Qu, Y., Guo, R., Wang, W., Qi, H., Ayhan, B., Kwan, C., Vance, S.: Anomaly detection in hyperspectral images through spectral unmixing and low rank decomposition. In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Beijing, pp. 1855–1858 (2016)
10. Weninger, F., Schuller, B.: Optimization and parallelization of monaural source separation algorithms in the openBliSSART toolkit. *J. Signal Process. Syst.* **69**(3), 267–C277 (2012)
11. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999)
12. Hoyer, P.O.: Non-negative sparse coding. In: *IEEE Workshop on Neural Networks for Signal Processing*, Martigny, Switzerland, pp. 557–565 (2002)
13. Feng, T., Li, S.Z., Shum, H.Y., Zhang, H.J.: Local non-negative matrix factorization as a visual representation. In: *Proceedings of the 2nd International Conference on Development and Learning*, pp. 178–183 (2002)
14. Hoyer, P.O., Dayan, P.: Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* **5**, 1457–1469 (2004)

15. Montano, A.P., Carazo, J.M., Kochi, K., Lehmann, D., Pascual-Marqui, R.D.: Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans. Pattern Anal.* **28**(3), 403–415 (2006)
16. Jenatton, R., Obozinski, G., Bach, F.: Structured sparse principal component analysis. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)* (2010)
17. Peharz, R., Pernkopf, F.: Sparse nonnegative matrix factorization with ℓ_0 -constraints. *Neurocomputing* **80**, 38–46 (2012)
18. Zheng, W.S., Lai, J.H., Liao, S.C., He, R.: Extracting non-negative basis images using pixel dispersion penalty. *Pattern Recogn.* **45**(8), 2912–2926 (2012)
19. Cai, D., He, X., Han, J.: Locally consistent concept factorization for document clustering. *IEEE Trans. Knowl. Data Eng.* **23**(6), 902–913 (2011)
20. Ding, C.H., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(1), 45–55 (2010)
21. Ahn, J.H., Choi, S., Oh, J.: A multiplicative up-propagation algorithm. In: *Proceedings of the 21st International Conference on Machine Learning*, p. 3 (2004)
22. Song, H.A., Kim, B.K., Xuan, T.L., Lee, S.Y.: Hierarchical feature extraction by multi-layer non-negative matrix factorization network for classification task. *Neurocomputing* **165**, 63–74 (2015)
23. Trigeorgis, G., Bousmalis, K., Zafeiriou, S., Schuller, B.W.: A deep matrix factorization method for learning attribute representations. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(3), 417–429 (2017)
24. Lyu, B., Xie, K., Sun, W.: A deep orthogonal non-negative matrix factorization method for learning attribute representations. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S. (eds.) *ICONIP 2017. LNCS*, vol. 10639, pp. 443–452. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70136-3_47
25. Guo, Z., Zhang, S.: Sparse deep nonnegative matrix factorization (2017). <http://arxiv.org/abs/1707.09316>
26. Yu, J., Zhou, G., Cichocki, A., Xie, S.: Learning the hierarchical parts of objects by deep non-smooth nonnegative matrix factorization (2018). <http://arxiv.org/abs/1803.07226>
27. Xue, H., Dai, X., Zhang, J., Huang, S., Chen, J.: Deep matrix factorization models for recommender systems. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, pp. 3203–3209 (2017)
28. Zhao, H., Ding, Z., Fu Y.: Multi-view clustering via deep matrix factorization. In: *Thirty-First AAAI Conference on Artificial Intelligence*, pp. 2921–2927 (2017)
29. Flenner, J., Hunter, B.: A deep non-negative matrix factorization neural network (2017)
30. Le Roux, J., Hershey, J.R., Wenginger, F.: Deep NMF for speech separation. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, Australia, pp. 66–70 (2015)
31. Fan, J., Cheng, J.: Matrix completion by deep matrix factorization. *Neural Netw.* **98**, 34–41 (2017)
32. Xu, L., Yu, B., Zhang, Y.: An alternating direction and projection algorithm for structure-enforced matrix factorization. *Comput. Optim. Appl.* **68**(2), 333–362 (2017). <https://doi.org/10.1007/s10589-017-9913-x>

33. Xu, L., Zhou, Y., Yu, B.: Classification and clustering via structure-enforced matrix factorization. In: Sun, Y., Lu, H., Zhang, L., Yang, J., Huang, H. (eds.) IScIDE 2017. LNCS, vol. 10559, pp. 403–411. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67777-4_35
34. Glowinski, R., Marroco, A.: Sur l'approximation, par elements finis d'ordre un, et la resolution, par penalisation-dualite d'une classe de problemes de dirichlet non lineaires. *Revue francaise d'automatique, informatique, recherche operationnelle. Analyse numerique* **9**(2), 41–76 (1975)
35. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* **2**(1), 17–40 (1976)



An Embedded Method for Feature Selection Using Kernel Parameter Descent Support Vector Machine

Haiqing Zhu¹, Ning Bi¹, Jun Tan^{1(✉)}, and Dongjie Fan²

¹ School of Mathematics, Sun Yat-sen University, Guangzhou 510275, China
mcstj@mail.sysu.edu.cn

² Center for Urban Science and Progress, New York University,
New York 10012, USA

Abstract. We introduce a novel embedded algorithm for feature selection, using Support Vector Machine (SVM) with kernel functions. Our method, called Kernel Parameter Descent SVM (KPD-SVM), is taking parameters of kernel functions as variables to optimize the target functions in SVM model training. KPD-SVM use sequential minimal optimization, which breaks the large quadratic optimization problem into some smaller possible optimization problem, avoids inner loop on time-consuming numerical computation. Additionally, KPD-SVM optimize the shape of RBF kernel to eliminate features which have low relevance for the class label. Through kernel selection and execution of improved algorithm in each case, we simultaneously find the optimal solution of selected features in the modeling process. We compare our method with algorithms like filter method (Fisher Criterion Score) or wrapper method (Recursive Feature Elimination SVM) to demonstrate its effectiveness and efficiency.

Keywords: Feature selection · Support vector machine
Kernel function

1 Introduction

Feature Selection is a vital issue in machine learning. It is common to apply feature selection methods to classification problems, especially when those original data sets have redundant features [1].

According to [2], there are three main directions for feature selection: filter, wrapper, and embedded methods.

Filter takes statistical analysis to filter out poorly informative features, it is usually done before the samples taken into a classifier. Relief [3] is a typical filter method which is statistically relevant to the target concept and feeds features into the classifier.

Wrapper approach searches the whole set of samples to score feature subset, therefore it naturally entails training and implementation of learning algorithms

during the procedure of feature selection, wrappers use different classifier such as naive Bayes [4], neural networks [5] and nearest neighbor [6]. The random forests based wrapper approaches [7,8] are widely used to identify important features from feature subset.

In embedded method, feature selection is embedded into the classifier [9], feature is selected by the internal function of an algorithm such as least absolute shrinkage and selection operator (LASSO) [10] and decision tree [11].

Above methods have their limitation, wrapper algorithms are complex in computation, but usually obtain more accurate results than filter methods [12], the problem of a wrapper is high computational cost because it involves repeated training. The robustness of above methods in high dimension data set is a crucial problem. Therefore some features select approaches constructed by combining multiple classifiers, their robust more than the approaches with a single classifier [13]. In addition, support vector machines (SVM) have been proposed as a wrapper classifier for feature selection [14].

Although standard implementation of SVM shows good performance in classification prediction, it cannot rank each features' importance for feature elimination. Thus we introduce a novel approach which selects features according to the descent path of kernel parameters, indirectly figuring out the importance of each features as well as optimizing the model predicting ability. The method we called Kernel Parameter Descent Support Vector Machine (KPD-SVM), the approach not only optimizes the parameter of SVM, but also obtains a subset of features for specific objective. KPD-SVM will be talked in detail and be compared with other characteristic approaches of feature selection in SVM.

2 Related Works

2.1 Support Vector Machine

In this section, we will simply review the development of SVM method.

Support Vector Machine (SVM) is a strictly math-based machine learning model, raised by Vapnik [15]. The principle of SVM classifier is obvious. It tries to find out the optimal hyperplane for the optimization problem with "soft margin" as follows:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i \cdot (\mathbf{w}^T \cdot \phi(\mathbf{x}_i) + b) \geq (1 - \xi_i) \quad i = 1, \dots, n, \\ & \xi_i \geq 0, \quad i = 1, \dots, n \end{aligned} \quad (1)$$

Here we denote ξ_i as slack variable. The training data can be transformed into higher dimensional space through kernel function $x \rightarrow \phi(x)$. So the decision function can be rewritten as:

$$f(x) = \sum_{i=1}^n \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle + b \quad (2)$$

Since the scalar products $\langle \phi(x), \phi(y) \rangle$ are the only value to be calculated, kernel function

$$K(x, y) = \langle \phi(x), \phi(y) \rangle \tag{3}$$

is used to solve them. As result the optimization problem can be rewritten as:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \quad i = 1, \dots, n \\ & \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned} \tag{4}$$

2.2 Feature Selection in SVMs

Typically, there are three methods in SVM based feature selection process, Filter, Wrapper and Embedded [1]. Here we review each of them briefly and stress one representative algorithm of each method, for experimental comparison in next section.

- **Filter Method:** Among all the measurement in Filter method, Fisher Criterion Score (F-Score) is one of the most common indicator to use. It computes the significance of each feature independently of the other feature by comparing that feature’s correlation to the output labels. The respective score $F(j)$ of feature j is given by:

$$F(j) = \left| \frac{\mu_j^+ + \mu_j^-}{(\sigma_j^+)^2 + (\sigma_j^-)^2} \right| \tag{5}$$

Where μ_j^+ (μ_j^-) is the mean value for the j th feature in positive(negative) class. And σ_j^+ (σ_j^-) is the standard deviation. When the $F(j)$ is large, it means j th feature has much more information to discriminate itself from other features, which suggests it ranks top of the feature list and would be more likely not to be eliminate and vice versa. The disadvantage of filter method is time consuming and skillful because you need to choose a suitable measurement method.

- **Wrapper Method:** One representative wrapper method is Recursive Feature Elimination SVM (RFE-SVM), which is raised by Guyon [16]. RFE-SVM aims to find out the r -feature subset among the original n -feature set through backward greedy algorithm, which build model by the whole feature at the beginning then cut off one feature according the ranking order. The disadvantage of Wrapper method is that it is more time consuming than filter method because it need to train models on different feature subsets.

- **Embedded Method:** The last method for feature selection is embedded method. The most different novelty between embedded and others is that it conducts the selection in the process of model training. One common embedded method is to add a penalty item to the target function which limits the model complexity [17]. Compared with filter method and wrapper method, we choose embedded method in our model because it is less time consuming.

3 The Proposed Method: KPD-SVM

The principle of proposed method aims to improve the classification performance as well as to eliminate less important features by optimizing parameter/parameters in kernel function. This method use penalty item like $L0 - norm$ or $L1 - norm$ of the parameter to punish the large number of feature we consider in modeling which is more likely to cause over-fitting problems. Through gradient descent algorithm, we can find out the best solution (which means the best classification performance) of the vector of kernel parameters. During this iteration process, we set the parameters whose values are lower than a small criterion as 0. Thus we can deal with the feature selection task.

3.1 Kernel Function

Among the kernel function SVM commonly uses, we pay attention to the following mostly-used kernels:

Gaussian Kernel function we write the kernel function in the form of the summation in each feature:

$$K(x, y) = \exp\left(-\sum_{j=1}^d \frac{(x_j - y_j)^2}{2\sigma_j^2}\right) \quad (6)$$

where $\sigma = [\sigma_1, \sigma_2, \sigma_3 \dots, \sigma_n]$ indicates the width of the kernel and determines the kernel shape. d is the number of features. For better demonstration, we denote:

$$\gamma = \left[\frac{1}{2\sigma_1^2}, \frac{1}{2\sigma_2^2}, \frac{1}{2\sigma_3^2}, \dots, \frac{1}{2\sigma_d^2}\right] \quad (7)$$

which leads to

$$K(x, y) = \exp\left(-\sum_{j=1}^d \gamma_j (x_j - y_j)^2\right) \quad (8)$$

Exponential kernel (Laplace) Similar with Gaussian kernel, it is shown as:

$$K(x, y) = \exp\left(-\sum_{j=1}^d \gamma_j (x_j - y_j)\right) \quad (9)$$

Polynomial kernel its function as:

$$K(x, y) = (\alpha x^T y + c)^D \tag{10}$$

Here we fix D and let $c = 1$ in our proposed method, hence we only need to consider the vector of α :

$$K(x, y) = \left(\sum_{j=1}^d \alpha_j x_j y_j + 1 \right)^D \tag{11}$$

3.2 Target Function in KPD-SVM

According the previous definition, the set of Lagrange multipliers α is considered, and adding the new parameter γ in kernel function and penalty item of model complexity, therefore the optimization problem $\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$ is minimized with a penalty function and some constrains. Our target function G is as follows:

$$\min_{\alpha, \gamma} G(\alpha, \gamma) = \min_{\alpha} \Psi(\alpha) + \min_{\alpha, \gamma} \Phi(\alpha, \gamma) \tag{12}$$

where the $\Psi(\alpha)$ are transformed from the target optimization function (4) of the standard SVM:

$$\begin{aligned} \min_{\alpha} \Psi(\alpha) &= \min_{\alpha} - \sum_{i=1}^n \alpha_i + \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{13} \\ \text{s.t.} \quad &0 \leq \alpha_i \leq C \quad i = 1, \dots, n \\ &\sum_{i=1}^n \alpha_i y_i = 0 \end{aligned}$$

and $\Phi(\alpha, \gamma)$ is penalized function, the first item of Eq. (14) is transformed from the second item of Eq. (13), the second item of Eq. (14) is penalized item:

$$\begin{aligned} \min_{\alpha, \gamma} \Phi(\alpha, \gamma) &= \min_{\alpha, \gamma} \frac{1}{2} \sum_{i,s=1}^n \alpha_i \alpha_s y_i y_s K(x_i, x_s, \gamma) + C_2 f(\gamma) \tag{14} \\ \text{s.t.} \quad &0 \leq \alpha_i \leq C \quad i = 1, \dots, n \\ &\sum_{i=1}^n \alpha_i y_i = 0 \\ &\gamma_j \geq 0 \quad i = 1, \dots, d \end{aligned}$$

where γ_j need to be non-negative and we use $L0 - norm$ as $f(\gamma)$, which is approximately equal to [9]:

$$f(\gamma) = \mathbf{e}^T (\mathbf{e} - \exp(-\beta\gamma)) = \sum_{j=1}^d [1 - \exp(-\beta\gamma_j)] \tag{15}$$

C_2 is the strength of the penalty of the complexity of our model which is different from C for penalty of training error (slack variable ξ). Also $L0 - norm$ can be replaced with $L1 - norm$ or $L2 - norm$ in our target function.

Because this optimization problem is not convex [17], it may be hard to search the globally optimal solution. So that we propose an algorithm to search a locally optimal solution. Then we use a method to solve this optimization problem in two step [17]:

[Step 1] Given a set of fixed kernel parameter γ , calculate the value of α in optimal function $\min_{\alpha} \Psi(\alpha)$, here sequential minimal optimization (SMO) [18] is a method to solve the SVM QP problem.

For convenience, all quantities that refer to the first multiplier will have a subscript 1, while the other refers to the second multiplier α_2 . Without loss of generality, the second multiplier α_2 will be computed firstly. The following bounds W, H apply to α_2 while the target y_1 does not equal the target y_2 :

$$W = \max(0, \alpha_2 - \alpha_1), H = \min(C, C + \alpha_2 - \alpha_1). \tag{16}$$

If the target $y_1 = y_2$, the bounds apply to α_2 is shown as:

$$W = \max(0, \alpha_2 + \alpha_1 - C), H = \min(C, \alpha_2 + \alpha_1). \tag{17}$$

The second derivative of the objective function $\min_{\alpha} \Psi(\alpha)$ along the diagonal line can be conducted as:

$$\eta = K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2). \tag{18}$$

Under the normal condition, the objective function is positive definite, there will be a minimum along the direction of the linear constraint, and η is greater than 0. The new minimum is computed along the direction of the constraint as follow:

$$\alpha_2^{opt} = \alpha_2 + \frac{y_2(E_1 - E_2)}{\eta} \tag{19}$$

where $E_i = u_i - y_i, i = 1, 2$ is the error on the i -th training example, as a next step, the constrained minimum is clipped by the bound W, H . Let $s = y_1 y_2$. The optimal α_1 is computed by the optimized and clipped α_2 :

$$\alpha_1^{opt} = \alpha_1 + s(\alpha_2 - \alpha_2^{opt}) \tag{20}$$

Under unusual condition, η will not be positive, which can cause the objective function to become indefinite.

[Step 2] Find out the best γ for given fixed α in step 1, solve the objective function $\min_{\alpha, \gamma} \Phi(\alpha, \gamma)$ using gradient descent algorithm. And if the renewed γ_j is below the criterion we set, eliminate the feature j and loop for next iteration until reaching the stop criterion. For given j the gradient of $F(\gamma_j^*)$ is:

Gaussian

$$\begin{aligned} \Delta_j \Phi(\gamma_j^*) &= \frac{1}{2} \sum_{i,s=1}^n \gamma_j^*(x_{i,j} - x_{s,j})^2 \alpha_i \alpha_s y_i y_s K(x_i, x_s, \gamma_j^*) \\ &+ C_2 \beta \exp(-\beta \gamma_j^*) \end{aligned} \tag{21}$$

Polynomial

$$\Delta_j \Phi(\gamma^{poly}) = \frac{1}{2} \sum_{i,s=1}^n Dx_{i,j}x_{s,j}\alpha_i\alpha_s y_i y_s K(x_i, x_s, \gamma^{poly}, D - 1) + C_2 \beta \exp(-\beta \gamma_j^{poly}) \tag{22}$$

To avoid misunderstandings of γ in polynomial kernel and target function, we set γ^{poly} in polynomial kernel. **Exponential Kernel (Laplace)**

$$\Delta_j \Phi(\gamma^*) = \frac{1}{2} \sum_{i,s=1}^n (x_{i,j} - x_{s,j})\alpha_i\alpha_s y_i y_s K(x_i, x_s, \gamma^*) + C_2 \beta \exp(-\beta \gamma_j^*) \tag{23}$$

The algorithm adjust the kernel components using gradient descent procedure, specially to parameter γ , which is set to be small to avoid negative at the first iterations.

3.3 Detailed Process of Proposed Algorithm

The pseudo code is shown as below:

Algorithm 1. KPD-SVM

kernel selection: we take **Gaussian kernel** as an example.

input:

parameter of gentle update strategy: d_1, d_2, θ ;

parameter of update: $\varepsilon_1, \varepsilon_2$

01 **start:** $stop = False, t = 0,$

$$\gamma^* = (\gamma^*)^{[0]}, \quad \alpha_1^{[0]}, \alpha_2^{[0]}$$

02 **WHILE** $stop \neq True$

03 train SVM for a given γ^* using SMO

04 **FOR** $i = 1, \dots, d_1$

05 compute E_1, E_2, η, s

$$06 \quad \alpha_2^{[i+1]} = \alpha_2^{[i]} + \frac{\eta(E_1 - E_2)}{\eta}$$

$$\alpha_1^{[i+1]} = \alpha_1^{[i]} + s * (\alpha_2^{[i]} - \alpha_2^{[i+1]})$$

07 **IF** $\|(\alpha_1)^{[t+1]} - (\alpha_1)^{[t]}\| < \varepsilon_1$

THEN $\alpha^* = (\alpha_1)^{[t+1]}$ **Break** **ENDIF**

08 **ENDFOR**

09 train SVM for a given α^*

10 **FOR** $j = 1, \dots, d_2$

$$11 \quad (\gamma_j^*)^{[t+1]} = (\gamma_j^*)^{[t]} - \theta \Delta_j \Phi((\gamma^*)^{[t]})$$

12 **IF** $(\gamma_j^*)^{[t+1]} < \varepsilon_2$

THEN $(\gamma_j^*)^{[t+1]} = 0$ **Break** **ENDIF**

13 **ENDFOR**

14 **IF** $(\gamma^*)^{[t]}, (\gamma^*)^{[t+1]}$ meet the requirements of $\zeta_{absolute}, \zeta_{relative}$

15 $stop = True$

16 **ENDIF**

17 **ENDWHILE**

where

$$\gamma^* = \sqrt{2\gamma} = \left[\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \dots, \frac{1}{\sigma_d} \right] \tag{24}$$

and

$$F(\gamma^*) = \sum_{i,s=1}^n \alpha_i \alpha_s y_i y_s K(x_i, x_j, \gamma^*) + C_2 f(\gamma^*) \tag{25}$$

In the algorithm, we may consider the following vital step, some details are given as follows:

Kernel Selection, Use the whole features to train model with different kernels (eg. Gaussian, Polynomial) and different parameter (γ, D, c) . Calculate the average accuracy of each model with different kernels by cross validations. Then select the kernel with the best performance which is the most appropriate kernel of this data set.

Set Original Value, At the start of algorithm, we give the initial value of α, γ , and some parameter for update.

Calculate α , Based on standard SVM training process and may take SMO algorithm [18] to quickly and efficiently find out the answer α^* .

Update σ and γ , Apply gradient descent algorithm to renew σ_i or γ_i^* , the lines 10–13 of the algorithm shows the iteration process, one by one for fixed the optimal α .

Step size of gradient descent, We set θ as the step size of gradient descend in each iteration.

Elimination criterion, ε is the eliminate threshold which means we eliminate the feature j by setting $\gamma_j^* = 0$ if value γ_j^* is below ε .

Stop criterion, For the stop criterion, we set a relative stop criterion $\zeta_{relative}$ and an absolute stop criterion $\zeta_{absolute}$ in order to balance the time of iterations and the performance of the model. $\zeta_{relative}$ is defined as the ratio $\frac{\|(\gamma^*)^{[t+1]} - (\gamma^*)^{[t]}\|_1}{\|(\gamma^*)^{[t]}\|_1}$ and $\zeta_{absolute}$ is set as $\|(\gamma^*)^{[t]}\|_1$.

3.4 Discussion of Parameter

Our discussion mainly concentrates on one issue: Selection of parameter values in proposed method. Basically, the proposed method outperforms in its process of feature selection and modeling. However, there are some parameters we need to tune for the optimal solution of classification. In [17], it has already concluded that β, ε and $\gamma^{[0]}$ have less influence in the final solution. In terms of the penalty for slack variables, C , we use Leave-One-Out Cross-Validation to find the best value of C in each case.

Complexity Penalty C_2 : C_2 is the coefficient of penalty item on the number of feature or model complexity. A large C_2 means a strict limitation to build greatly complicated model. We choose C_2 according to the balance of prediction performances and model complexity.

Step Size θ : θ represents the step size of gradient descend in each iteration.

We want to use an automatically adjusted step size in some cases. Hence, we denote θ_{auto} as $\frac{\varepsilon}{\text{median}\{\Delta_j F(\gamma^*)\}}$, $j = 1, \dots, d$. And we may take $\theta = \min\{\theta_{original}, \theta_{auto}\}$ as step size in each iteration.

Stop Criterion $\zeta_{absolute}, \zeta_{relative}$: With the increasing number of iterations, the $1 - norm$ difference of kernel parameter in t and $t + 1$ iteration goes to convergence, which shows the algorithm can find out the best kernel parameter in certain countable iterations.

4 Experiments

In this section, we apply the proposed method to do experiments in some real-world dataset. Also we will compare our method with F-score and RFE-SVM, which represents the filter and wrapper algorithm in feature selection. The measurements we make comparison are as follows: First, model prediction performance. Second, the number of features in the optimal solution.

4.1 Data Set

The data sets we selected are from UCI Machine Learning Database. Detailed information of each data set is shown as follows:

- **Sonar**: This is the data set used by Gorman and Sejnowski in their study of the classification of sonar signals.
The data set contains 111 patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions. And it contains 97 patterns obtained from rocks under similar conditions. The label associated with each record contains the letter “R” if the object is a rock and “M” if it is a mine (metal cylinder).
- **WBC**: The Wisconsin Breast Cancer data set has 569 observations and 30 features. All feature values are recoded with four significant digits. In addition, people who are diagnosed are labeled as M (*malignant tumor*) and the other are marked as B (*benign tumor*).

We basically consider the following three kernel functions: Gaussian, Polynomial, Laplace (Exponential). Then the values of parameters in each kernel function we used are as follows:

- $\sigma_{Gaussian} = (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 50, 100, 500, 1000)$
- $D = (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)$
- $\sigma_{Laplace} = (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 50, 100, 500, 1000)$

4.2 Case: Sonar

Basic information of this data set is shown in Table 1.

Table 1. Basic information of Sonar (mines vs. rocks) data set

	Features	Observations	Proportion	Predominant class prop.
Total	60	208	100%	53.4%
Train	60	145	70%	54.5%
Test	60	63	30%	50.8%

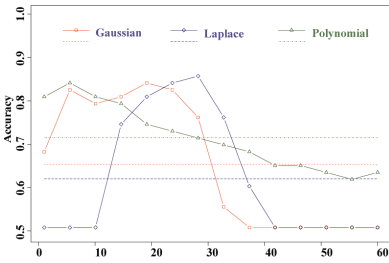


Fig. 1. The accuracy of Gaussian, Laplace and Polynomial in Sonar (horizontal axis represents feature numbers)

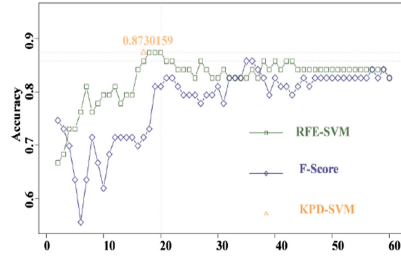


Fig. 2. The accuracy of KPD-SVM, F-Scores and RFE-SVM in Sonar (horizontal axis represents feature numbers)

First we carry out kernel selection. Fig. 1 shows the average performances of each kernel function applied in Sonar. Thus we choose Polynomial Kernel in this case.

Figure 2 shows the performance of proposed method KPD-SVM compared with F-Score and RFE-SVM. The optimal feature subsets are selected by each method, and the number of these subsets are shown below: Filter(F1-Scores):24, Wrapper(RFE-SVM):18-20, Embedded(KPD-SVM):20.

In conclusion, KPD-SVM outperforms F-Score and RFE-SVM in this Sonar case.

4.3 Case: WBC

Basic information of this data set is shown in Table 2.

First we carry out kernel selection. In WBC we choose Polynomial Kernel in this case. Figure 3 shows the average performances of each kernel function applied in WBC.

The performance of proposed method KPD-SVM compared with F-Score and RFE-SVM shown in Fig. 4. The optimal feature subsets are selected by each

Table 2. Basic information of WBC data set

	Features	Observations	Proportion	Predominant class prop.
Total	30	569	100%	62.7%
Train	30	512	90%	63.4%
Test	30	57	10%	52.6%

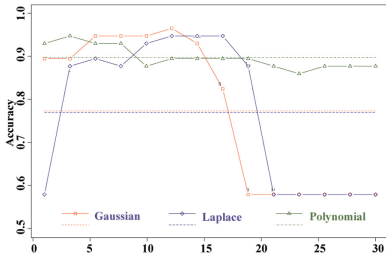


Fig. 3. The accuracy of Gaussian, Laplace and Polynomial in WBC (horizontal axis represents feature numbers)

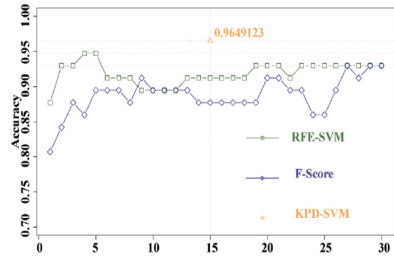


Fig. 4. The accuracy of KPD-SVM, F-Scores and RFE-SVM in WBC (horizontal axis represents feature numbers)

method, and the number of these subsets are shown below: Filter(F1-Scores):26, Wrapper(RFE-SVM):19, Embedded(KPD-SVM):15.

In conclusion, considering the model prediction accuracy and the model complexity (the number of features), we can say KPD-SVM outperforms in this WBC case.

5 Conclusion

In this paper, we have presented a novel method called Kernel Parameter Descent Support Vector Machine (KPD-SVM) for feature selection using kernel functions. Our embedded method can generalize a well-trained SVM classifier as well as a good solution for feature selecting. In addition, our KPD-SVM method outperforms other methods, like filter method (F-Score) and wrapper method (RFE-SVM). Besides, compared with former embedded algorithm by optimizing kernel parameters [1–4], our method has novelties in stop criterion and step size settings in executions, which performs better in time consuming.

Acknowledgements. We would like to acknowledge Professor Chih-Jen Lin from National Taiwan University for his research on Support Vector Machine and his work on software LIBSVM.

This work was supported by the Guangdong Provincial Government of China through the “Computational Science Innovative Research Team” program and Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University, and the National Science Foundation of China (11471012).

References

1. Chandrashekar, G., Sahin, F.: A Survey on Feature Selection Methods. Pergamon Press, Inc., Oxford (2014)
2. Cheriet, M., Kharma, N., Liu, C.L., et al.: Character Recognition Systems: A Guide for Students and Practitioners. Scitech Book News (2007)
3. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: International Workshop on Machine Learning, pp. 249–256. Morgan Kaufmann Publishers Inc., Burlington (1992)
4. Kohavi, R., John, G.H.: Wrappers for feature subset selection. *Artif. Intell.* **97**(1–2), 273–324 (1997)
5. Sesmero, M.P., Alonso-Weber, J.M., Ledezma, A., et al.: A new artificial neural network ensemble based on feature selection and class recoding. *Neural Comput. Appl.* **21**(4), 771–783 (2012)
6. Yang, J., Yao, D., Zhan, X., Zhan, X.: Predicting disease risks using feature selection based on random forest and support vector machine. In: Basu, M., Pan, Y., Wang, J. (eds.) ISBRA 2014. LNCS, vol. 8492, pp. 1–11. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08171-7_1
7. Anaissi, A., Kennedy, P.J., Goyal, M., et al.: A balanced iterative random forest for gene selection from microarray data. *Bmc Bioinform.* **14**(1), 1–10 (2013)
8. Swan, A.L., Mobasheri, A., Allaway, D., et al.: Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics J. Integr. Biol.* **17**(12), 595–610 (2013)
9. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. *Artif. Intell.* **97**(1–2), 245–271 (1997)
10. Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc.* **73**, 273–282 (2011)
11. Chan, H.P., Kim, S.B.: Sequential random k-nearest neighbor feature selection for high-dimensional data. *Expert. Syst. Appl.* **42**(5), 2336–2342 (2015)
12. Maldonado, S., Weber, R.: A wrapper method for feature selection using support vector machines. *Inf. Sci.* **179**(13), 2208–2217 (2009)
13. Tuv, E., Borisov, A., Runger, G., et al.: Feature selection with ensembles, artificial variables, and redundancy elimination. *J. Mach. Learn. Res.* **10**(3), 1341–1366 (2009)
14. Chen, P., Zhang, D.: Constructing support vector machines ensemble classification method for imbalanced datasets based on fuzzy integral. In: Ali, M., Pan, J.-S., Chen, S.-M., Horng, M.-F. (eds.) IEA/AIE 2014. LNCS (LNAI), vol. 8481, pp. 70–76. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07455-9_8
15. Vapnik, V.N., Vapnik, V.: *Statistical Learning Theory*, vol. 1. Wiley, New York (1998)
16. Guyon, I., Gunn, S., Nikravesh, M., et al.: *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing. Springer, New York (2006). <https://doi.org/10.1007/978-3-540-35488-8>
17. Maldonado, S., Weber, R., Basak, J.: *Simultaneous feature selection and classification using kernel-penalized support vector machines*. Elsevier Science Inc. (2011)
18. Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines. In: *Advances in Kernel Methods-Support Vector Learning*, pp. 212–223 (1998)



Multimodal Joint Representation for User Interest Analysis on Content Curation Social Networks

Lifang Wu¹, Dai Zhang¹, Meng Jian¹✉, Bowen Yang, and Haiying Liu

Faculty of Information Technology, Beijing University of Technology, Beijing, China
jianmeng648@163.com

Abstract. Content curation social networks (CCSNs), where users share interests by images and their text descriptions, are booming social networks. For the purpose of fully utilizing user-generated contents to analysis user interests on CCSNs, we propose a framework of learning multimodal joint representations of pins for user interest analysis. First, images are automatically annotated with category distributions, which benefit from the network characteristics and represent interests of users. Further, image representations are extracted from an intermediate layer of a fine-tuned multilabel convolutional neural network (CNN) and text representations are obtained with a trained Word2Vec. Finally, a multimodal deep Boltzmann machine (DBM) are trained to fuse two modalities. Experiments on a dataset from Huaban demonstrate that using category distributions instead of single categories as labels to fine-tune CNN significantly improve the performance of image representation, and multimodal joint representations perform better than either of unimodal representations.

Keywords: Multimodal · Content curation social networks
User modeling · Recommender systems

1 Introduction

Content curation social networks (CCSNs) are interest-driven social networks where users can organize and demonstrate multimedia contents they like. Since the most typical CCSN Pinterest became the fastest social network to reach 10M users [4], CCSNs have become popular worldwide. In China, more than 50 Pinterest-like websites such as Huaban, Duitang, Meilishuo, Mogujie and so forth have been published. The rapid development of CCSNs attracts much attention on different research topics, for example, network characteristic analysis [4], user

Supported by National Natural Science Foundation of China 61702022, Beijing Municipal Education Commission Science and Technology Innovation Project KZ201610005012, China Postdoctoral Science Foundation funded project 2017M610027 and 2018T110019.

behavior study [5], influence analysis [18], search engine [21], recommender systems [2, 9, 10, 19] and user modeling [1, 3, 20].

On CCSNs, the carrier of user interests is the basic unit of the network called “pin”, which comprises an image and its text description. Most prior works on CCSNs only focused on unimodal data. Yang et al. [19] modeled boards with text representations and recommended boards re-ranked with image representations. Cinar et al. [1] separately predicted categories of pins with either image representations or text representations and fused the results of two modalities by decision fusion. Liu et al. [10] used unimodal representations to respectively generate candidate pins and to re-rank all the candidates. All these methods are late fusion methods which cannot obtain multimodal joint representations.

Multimodal joint representation commonly consists of unimodal representation and multimodal fusion. With regard to image representation, convolutional neural networks (CNNs) have recently achieved many outstanding performances on computer vision. Some works have been done on employing CNNs to represent pins. A key to train CNNs is to create a large labelled dataset. Cinar et al. [1], and You et al. [20] directly used the category of a pin as its label, but this label may be inaccurate as different users may select different categories for a same image. Geng et al. [3] constructed an ontology in fashion domain and trained a multi-task CNN with concepts in ontology, but this methods is hard to be deployed in all domains. Zhai et al. [21] obtained more detailed labels by taking top text search queries on Pinterest, however, the quality and consumption of this annotation highly depends on the performance of the search engine. Inspired by the fact that categories predefined by CCSNs are not independent objects but related notions, we use category distributions based on statistics as labels and fine-tuned a multilabel CNN for image representation.

Many multimodal fusion studies have been carried out on classification and retrieval. Most existed methods are based on discriminative models such as latent Dirichlet allocation [15], CNN [11] and recurrent neural network [12]. Those methods mainly learn the consistency between modalities and can hardly deal with missing input modalities. On the generative side, restricted Boltzmann machine (RBM) [6], deep autoencoder (DAE) [14] and deep Boltzmann machine (DBM) [17] are proved to be feasible to learn both the consistency and complementarity between modalities and can easily deal with the absence of some modalities, however, limited works have been done on fusing features obtained by deep learning with these models. Zhang et al. [22] fused visual features extracted from the 6-th layer of AlexNet and textual features generated by sparse coding of word vectors from a Word2Vec [13] with a DAE. Since DAEs are deterministic models while DBMs are probabilistic models, we trained a multimodal DBM to improve generalization performance.

The proposed framework of learning multimodal joint representations of pins is shown in Fig. 1. For image representation, visual features are extracted from an intermediate layer of the fined-tuned CNN. For text representation, distributed representations of words are learned on corpora and are encoded to represent texts. As our choice, Word2Vec is a frequently used distributed representation

for capturing semantic and syntactic relations between words. Mean vector [8] of Word2Vec performs well on text representation and is unsupervised. Our multimodal joint representations is finally generated by a pretrained modified multimodal DBM.

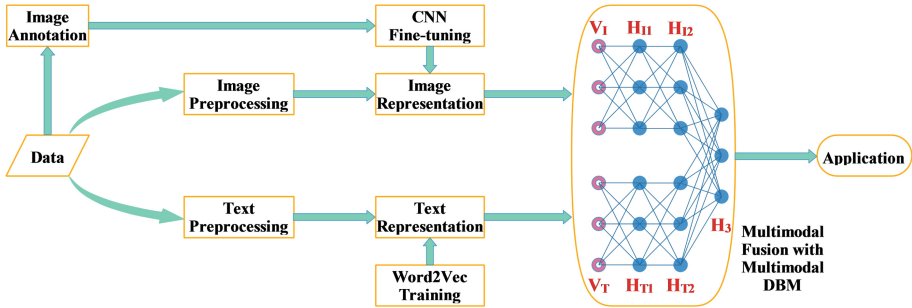


Fig. 1. Framework of learning multimodal joint representations of pins.

We believe that our research is the foundation of further researches on CCSNs such as board and user modeling, with the following contributions:

- We propose an easy-to-accomplish automatic annotate method that accumulate category selections of users to form category distributions of pins and fine-tune a multilabel CNN which significantly improves the category prediction performance.
- Multimodal joint representations of pins we get performs better than the unimodal representations.

The rest of the paper is organized as follows. Section 2 describes the proposed framework in details. Experiments and the corresponding analysis are provided in Sect. 3. And it is followed by conclusions in Sect. 4.

2 Multimodal Joint Representations of Pins

A pin comprise an image and its text description. As shown in Fig. 1, the whole process of multimodal joint representation can be roughly divided into three parts: image representation, text representation and multimodal fusion.

2.1 Image Representation

The aim of image representation is to learn features which not only maintain intrinsic characteristics of images but also relate to user interests on CCSNs. As supervised learning models, CNNs can certainly capture the relationships between images and user interests if user interests on CCSNs are used as labels during the learning process. Not to mention that top layers of CNNs can learn

high-level image features, which can be interpreted as color, material, texture, object, scene and so on by some means.

All pins on CCSNs are collected into boards. When a board is created, the owner must select one of categories predefined by CCSNs for it, and all pins in this board will have the same category as the board. Since the category can be considered as the theme of the board, it can be directly treated as a label, which describes a coarse-grained user interest. However, this label is probably weak and noisy, mainly because user preferences may lead to various category selections for a same image since it can be observed that categories in Table 1 are sometimes related notions. To put it in practical terms, the image in Fig. 2a may belong to photography, kids and pets on Huaban.

Table 1. List of all 33 predefined categories on Huaban.

Anime	Apparel	Architecture	Art	Beauty
Cars motorcycles	Data presentation	Design	Desire	DIY crafts
Education	Film music books	Fitness	Food drink	Funny
Games	Geek	Home	Illustration	Industrial design
Kids	Men	Modeling hair	People	Pets
Photography	Quotes	Sports	Tips	Travel places
Web app icon	Wedding events	Other		

The most frequently activity on CCSNs is called “re-pin”, which means a user collects an image and may add a text description for it from a board of another user into his or hers own board. A “re-pin path” is formed if users are interested in a same image and thus they re-pin it one by one, and all re-pin paths of an image form a “re-pin tree”, as illustrated in Fig. 2b. Because any one of the categories in the re-pin tree cannot decide what this image is about but describes a portion of it instead, we use a category distribution to represent interests of an image. The category distribution of a given image I can be computed after counting the categories in the re-pin tree as

$$Interest_I = \left(p_{C_i} = \frac{f_{C_i}}{\sum_{i=1}^{N_C} f_{C_i}} \right) \in [0, 1]^{N_C} \quad (1)$$

where f_{C_i} denotes the frequency of the i -th category C_i , N_C is the total number of categories on CCSNs. In practice we set

$$f_{C_i} = 0 \quad \text{if} \quad f_{C_i} < \frac{\sum_{i=1}^{M_C} f_{C_i}}{M_C} \quad (2)$$

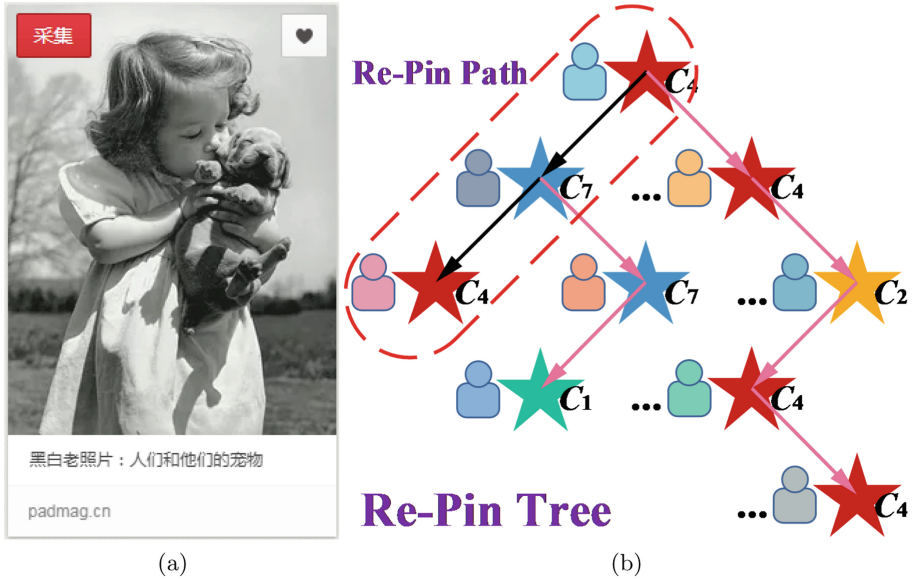


Fig. 2. (a): Example of a pin on Huaban. (b): Illustration of a re-pin tree composed of some re-pin paths. Each star represents a pin and C_i nearby is the category of this pin. All pins in the re-pin tree have a same image.

where M_C is the total number of categories occurred in the re-pin tree to filter out spam and make the sequence on behalf of majority opinion.

After automatic image annotation, we then choose a pretrained CNN model to fine-tune for the purpose of accelerate the training process. Most available pretrained CNNs are designed for classifying independent objects, while our model should be a multilabel regressor. Accordingly, we change the loss layer from softmax with logarithmic loss layer to sigmoid with cross entropy loss layer. The loss function is defined as

$$E = - \sum_{i=1}^{N_C} [p_{C_i} \ln \hat{p}_{C_i} + (1 - p_{C_i}) \ln (1 - \hat{p}_{C_i})] \quad (3)$$

where p_{C_j} is the percentage in Eq. (1), \hat{p}_{C_j} denotes the corresponding sigmoid output.

After fine-tuning, the weights of the CNN are stored for feature extraction. The activation values of an fully connected (FC) layer will be extracted as the image representations.

2.2 Text Representation

An important aim of text representation is also to discover the relationships between descriptions of pins and categories. However, it is difficult to create a

large labelled dataset on CCSNs for supervised learning as descriptions in the re-pin tree may be different.

Since there is no obvious difference between words used on CCSNs and those in common situations, we train a Word2Vec on some public corpora to encode words. Word2Vec is an efficient shallow model for learning distributed representations of words. Although the learning process of either of its two log-linear models, which are continuous bag-of-words (CBOW) and continuous skip-gram, is supervised, there is no need to annotate the training texts. Since the learned vectors capture a large number of meaningful semantic and syntactic word relationships, we make sure that the categories are in the training dictionary in order that the relationships between words and the categories can be considered as the relationships between words and user interests. In addition, distributed representations are scalable even though the vocabulary of natural language is extremely wide.

Owing to the fact that texts have diverse lengths, it is necessary to transform a set of word vectors into a single vector with a constant dimension for representing a complete text. For a text T , the text representation is the mean vector computed as

$$V_T = \frac{1}{M_T} \sum_{i=1}^{M_T} \text{KeyedVector}_{\text{Word}_i} \quad (4)$$

where $\text{KeyedVector}_{\text{Word}_i}$ denotes the word vector of the i -th word Word_i , M_T is the text length.

2.3 Multimodal Fusion

Different modalities typically have different statistical properties, which makes it difficult to learn a joint representation that capture both consistent and complementary relationships across modalities. A multimodal DBM which combines DBMs by adding a shared hidden layer on top of them can effectively solve this problem. A DBM is structured by stacking RBMs in a hierarchical manner. A RBM is an undirected graphical model with binary-valued visible layer V and binary-valued hidden layer H fully connected to each other defines the energy function

$$E(V, H; \theta) = -H^T W V - A^T V - B^T H \quad (5)$$

where $\theta = \{W, A, B\}$ denotes the model parameters including the symmetric interaction terms W between two layers, visible layer bias terms A and hidden layer bias terms B .

As illustrated in Fig. 1, we use two-layer DBMs with Gaussian-Bernoulli RBMs, which are a variant of RBMs that can model real-valued vectors, as bottom for both modalities. A Gaussian-Bernoulli RBM with visible units $V = \{v_i\} \in \mathbb{R}^D$ and hidden units $H = \{h_j\} \in \{0, 1\}^F$ defines the energy function

$$E(V, H; \theta) = \sum_{i=1}^D \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i=1}^D \sum_{j=1}^F \frac{v_i}{\sigma_i} w_{ij} h_j - \sum_{j=1}^F b_j h_j \quad (6)$$

where σ_i denotes the standard deviation of the i -th visible unit and $\theta = \{\{w_{ij}\} \in \mathbb{R}^{D \times F}\}, \{a_i\} \in \mathbb{R}^D, \{b_j\} \in \mathbb{R}^F, (\sigma_i) \in \mathbb{R}^D\}$. During the unsupervised training of the multimodal DBM, modalities can be thought of labels for each other. Since RBMs can be considered as autoencoders, each layer of the multimodal DBM makes a small contribution to eliminate modality-specific correlations. Consequently, the top layer can learn a relatively modality-free representation as opposed to the modality-full input layers. The joint distribution over the multimodal inputs can be written as

$$P(V_I, V_T; \theta) = \sum_{H_{I2}, H_{T2}, H_3} P(H_{I2}, H_{T2}, H_3) \left(\sum_{H_{I1}} P(V_I, H_{I1}, H_{I2}) \right) \left(\sum_{H_{T1}} P(V_T, H_{T1}, H_{T2}) \right) \quad (7)$$

where θ denotes all model parameters.

A pin may has no text description. The multimodal DBM can be used to generate missing text representation by sampling it from the conditional distribution with the standard Gibbs sampler. Finally, activation probabilities of H_3 are used as the multimodal joint representations of pins no matter they have text descriptions or not.

3 Experiment

3.1 Dataset and Implementation Details

All data used in experiments was crawled from Huaban, which is one of typical CCSNs in China. Huaban provides almost the same applications as Pinterest provides, while three main differences between them are: users can “like” pins or boards on Huaban while “like” has been removed by Pinterest; Huaban records both users from whom a pin re-pinned and by whom it initially created while Pinterest only records the direct source; some predefined categories are different and Pinterest has 5 more categories.

We first crawled pins without images of 5957 users and sampled 88 users according to pin counts and categories of their boards. To make our dataset diverse and real, a few cold start and extremely active users have been confirmed in it. We then downloaded the images of sampled users and pins of their like boards. In addition, top 1000 recommended pins of every category was crawled for fine-tuning the CNN, and re-pin paths of all recommended pins was crawled for automatic annotation. In total, the dataset includes 1694 boards and 167747 unique images. All pins was used as supplements for obtaining category distributions of all recommended pins. The average nodes of the incomplete re-pin trees is 47.57.

Labeled images was split into 80% for training and validating and the remaining 20% for testing after label balancing. AlexNet [7] with ImageNet [16] pre-trained weights was chosen as a basis. Because AlexNet requires a constant input dimension, the image was first rescaled such that the shorter side was of length 256 pixels, and then the central 256×256 patch of the resulting image was cropped out. As a comparison, we also used the most frequent category as label

to fine-tune an multiclass AlexNet. The dimensions of the fc8 layers of both CNNs were change to 33. Image representations was extracted from the FC7 layer of the fine-tuned CNN.

Word2Vec was trained on Wikipedia dumps and Sougou Lab dataset with CBOW and negative sampling. The vector dimension was set to 300. Words with total frequency lower than 5 are ignored. Preprocessing such as traditional Chinese and simplified Chinese conversion, removing punctuation, word tokenize, removing stop words and machine translation has been done on text descriptions of pins.

Image and text features were used for pretraining our multimodal DBM. Dimensions of H_{T_1} , H_{T_2} and H_{V_1} were equal to their corresponding visible inputs, and dimension of H_{V_2} and H_3 was set to 2048 for the purpose of compressing the vectors. DBM was pretrained using a greedy layer-wise strategy by learning a stack of modified RBMs. Finally, we ran Gibbs sampler to generate missing text representations and to infer multimodal join representations.

3.2 Analysis of Interests Represented by Pins

Analysis of interests represented by pins is the prerequisite of analysis of interests represented by boards and user interest analysis. The category distribution are interests of the image and can be approximate the interests of the pin, even though some of categories will be enhanced by the text description.

Table 2. Comparison on pin category prediction

Model	Dimension	Dominant category accuracy	Mean nonzero error	Mean error
AlexNet [1]	4096	57.53%	—	—
Word2Vec [1]	300	33.47%	—	—
AlexNet [20]	4096	43.1%	—	—
AlexNet	4096	45.85%	—	—
Multilabel	4096	82.71%	0.1320	0.0141
Word2Vec	300	42.88%	0.3249	0.0415
Multimodal	2048	84.13%	0.1181	0.0119

Multidimensional logical regressions (LRs) were trained on recommended pins for all unimodal and multimodal representations. The results are shown in Table 2, together with the result of the compared AlexNet. Relevant results on 32 [1] and 34 [20] Pinterest categories are also cited as references. Mean nonzero error is the average error between all nonzero categories and corresponding predictions. The dominant category accuracy checks the consistency of the most frequently category between predictions and labels. Comparison of two fine-tuned CNNs shows that our multilabel regressor significantly improves the accuracy.

It is because that category distributions can not only eliminate the interference of related categories but also provide more information to learn than only dominant categories. Although the performance of text representations is not comparable with those of image representations, the complementarity between two modalities helps the multimodal joint representations perform better than the unimodal representations. Our framework can also infer interests of images from other social networks.

3.3 Board Category Recommendation

Every board must be assigned a category nowadays, while some boards have no category as a result of that they were created before the constraint entered into force. However, it is illogical because even if it is hard to select a category for a board about wide interests, “other” in Table 1 can be selected. Consequently, Huaban offers a function that allows any user to select a category for a board which haven’t categorized. Board category recommendation will be useful on that occasion, and the first selection and further editing too.

Table 3. Comparison on board category recommendation

Model	Top-1 MRR	MRR
Random	3.03%	12.39%
Text + Cosine similarity	25.65%	38.78%
Image + Multidimensional LR	60.10%	73.41%
Text + Multidimensional LR	38.00%	54.30%
Multimodal + Multidimensional LR	62.35%	74.77%

Same as interests represented by pins, interests represented by boards should not be limited to one category. The interest distribution of a board can be computed by averaging all category distributions of its pins. As pins are accumulated, the category preference is reinforced due to the fact that the accumulation process of strong categories is faster than those of weak categories. Our recommended category is the max category in the interest distribution of the board, and the ground truth is the real category of the board. Mean reciprocal rank (MRR) are used as the performance metric. As board category recommendation actually has only one correct selection, we also give the top-1 MRR. Results are organized in Table 3. Cosine similarities between texts and categories are less effective than category distributions obtained with texts, this indicates that there is a gap between semantics and interests. The multimodal joint representations, which benefit from personalized texts, perform better than image representations. Notice that the recommendation dataset is different from the training dataset, it also proves that our framework has a good generalization ability.

The first selection is a cold start problem, as it only depends on one pin. We then evaluate the influence of pin counts on board category recommendation.

Table 4. Influence of pin count on board category recommendation based on multi-modal join representation

Pin count	Top-1 MRR	MRR
=1	7.69%	35.56%
≤ 4	45.57%	59.68%
≤ 30	56.53%	69.75%
≤ 100	59.06%	72.39%
> 100	67.11%	78.33%

As shown in Table 4, our recommendation suffers the cold start. However, the theory about preference reinforcement is proved as more pins lead to better performance. Although interests of users are more discrete than interests of boards, we infer that the accumulation process is still effective on user interest analysis.

3.4 Board Recommendation

Well organized boards can be high quality galleries, which makes it easier for users to collect pins. For this reason, CCSNs offer users a board recommendation function. Besides interest distributions, a board can be represent by the mean vector of representations of pins. And similarity between boards can be simply measured with some distance metrics, for example cosine similarity.

Table 5. Comparison on board recommendation

Model	Top-5 MRR	MRR
Category based	2.12%	3.85%
Image + Multidimensional LR	16.08%	18.61%
Text + Multidimensional LR	15.93%	17.95%
Multimodal + Multidimensional LR	17.58%	20.13%
Image + Mean vector	33.66%	35.97%
Text + Mean vector	25.96%	27.49%
Multimodal + Mean vector	35.76%	37.88%

We divided every board in half according to the order of pins, and each half must be similar board for another. The owner of each half will be interested in another half and further re-pin from or like or follow it beyond all doubt. On the basis of this, we consider half of the board as the only correct recommendation result and retrieve the index in the similarity sequence. As Huaban exhibit five pins at the top row of its waterfall flow for common resolution screens, we also demonstrate top-5 MRR. Table 5 shows that results of mean vectors is higher

than those of respective interest distributions, simply owing to the additional information. All of our methods significantly improve the results in comparison with the category based filtering. The results also show that multimodal joint representations can model boards better than either of unimodal representations.

4 Conclusion

We propose a framework of learning multimodal joint representations of pins on CCSNs. Experimental results show that multimodal joint representations performs better than either of unimodal representations on interpreting pin-level interests and board-level interests. The obtained representations can be easily used on user modeling and recommender systems for CCSNs. Future work will be focused on extending our framework to model boards and users. In addition, other effective feature extraction methods and multimodal fusion approaches may be taken into account.

References

1. Cinar, Y., Zoghbi, S., Moens, M.F.: Inferring user interests on social media from text and images. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1342–1347. IEEE (2015)
2. Geng, X., Zhang, H., Bian, J., Chua, T.: Learning image and user features for recommendation in social networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4274–4282. IEEE Computer Society (2015)
3. Geng, X., Zhang, H., Song, Z., Yang, Y., Luan, H., Chua, T.: One of a kind: user profiling by social curation. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 567–576. ACM (2014)
4. Gilbert, E., Bakhshi, S., Chang, S., Terveen, L.: “i need to try this!”: a statistical overview of pinterest. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2427–2436. ACM (2013)
5. Han, J., et al.: Sharing topics in pinterest: understanding content creation and diffusion behaviors. In: Proceedings of the 2015 ACM on Conference on Online Social Networks, pp. 245–255. ACM (2015)
6. Jia, X., Wang, A., Li, X., Xun, G., Xu, W., Zhang, A.: Multi-modal learning for video recommendation based on mobile application usage. In: 2015 IEEE International Conference on Big Data (Big Data) (BIG DATA), pp. 837–842. IEEE (2015)
7. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates Inc. (2012)
8. Lev, G., Klein, B., Wolf, L.: In defense of word embedding for generic text representation. In: Biemann, C., Handschuh, S., Freitas, A., Mezziane, F., Métais, E. (eds.) NLDB 2015. LNCS, vol. 9103, pp. 35–50. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19581-0_3
9. Li, Y., Mei, T., Cong, Y., Luo, J.: User-curated image collections: modeling and recommendation. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 591–600. IEEE (2015)

10. Liu, D., et al.: Related pins at pinterest: the evolution of a real-world recommender system. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 583–592. International World Wide Web Conferences Steering Committee (2017)
11. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015), pp. 2623–2631. IEEE (2015)
12. Mao, J., Xu, J., Jing, Y., Yuille, A.: Training and evaluating multimodal word embeddings with large-scale web annotated images. In: Advances in Neural Information Processing Systems 29, pp. 442–450. Curran Associates, Inc. (2016)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, pp. 3111–3119. Curran Associates Inc. (2013)
14. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, pp. 529–545. Omnipress (2011)
15. Qian, S., Zhang, T., Xu, C.: Multi-modal multi-view topic-opinion mining for social event analysis. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 2–11. ACM (2016)
16. Russakovsky, O., Salakhutdinov, R.: Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015)
17. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. *J. Mach. Learn. Res.* **15**, 2949–2980 (2014)
18. Venkatadri, G., Goga, O., Zhong, C., Viswanath, B., Gummadi, K., Sastry, N.: Strengthening weak identities through inter-domain trust transfer. In: Proceedings of the 25th International Conference on World Wide Web, pp. 1249–1259. ACM (2016)
19. Yang, X., Li, Y., Luo, J.: Pinterest board recommendation for twitter users. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 963–966. ACM (2015)
20. You, Q., Bhatia, S., Luo, J.: A picture tells a thousand words-about you! user interest profiling from user generated visual content. *Signal Process.* **124**, 45–53 (2016)
21. Zhai, A., et al.: Visual discovery at pinterest. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 515–524. International World Wide Web Conferences Steering Committee (2017)
22. Zhang, H., Yang, Y., Luan, H., Yan, S., Chua, T.: Start from scratch: towards automatically identifying, modeling, and naming visual attributes. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 187–196. ACM (2014)



LTSG: Latent Topical Skip-Gram for Mutually Improving Topic Model and Vector Representations

Jarvan Law, Hankz Hankui Zhuo^(✉), JunHua He, and Erhu Rong

Department of Computer Science, Sun Yat-Sen University, GuangZhou 510006, China
JarvanLaw@gmail.com, zhuohank@mail.sysu.edu.cn,
{hejunh, rongerhu}@mail2.sysu.edu.cn

Abstract. Topic models have been widely used in discovering latent topics which are shared across documents in text mining. Vector representations, word embeddings and topic embeddings, map words and topics into a low-dimensional and dense real-value vector space, which have obtained high performance in NLP tasks. However, most of the existing models assume the results trained by one of them are perfect correct and used as prior knowledge for improving the other model. Some other models use the information trained from external large corpus to help improving smaller corpus. In this paper, we aim to build such an algorithm framework that makes topic models and vector representations mutually improve each other within the same corpus. An EM-style algorithm framework is employed to iteratively optimize both topic model and vector representations. Experimental results show that our model outperforms state-of-the-art methods on various NLP tasks.

Keywords: Topic modeling · Polysemous-word · Word embeddings
Text mining

1 Introduction

Word embeddings, e.g., distributed word representations [16], represent words with low dimensional and dense real-value vectors, which capture useful semantic and syntactic features of words. Distributed word embeddings can be used to measure word similarities by computing distances between vectors, which have been widely used in various IR and NLP tasks, such as entity recognition [23], disambiguation [5] and parsing [21]. Despite the success of previous approaches on word embeddings, they all assume each word has a specific meaning and represent each word with a single vector, which restricts their applications in fields with polysemous words, e.g., “bank” can be either “a financial institution” or “a raised area of ground along a river”.

To overcome this limitation, [14] propose a topic embedding approach, namely Topical Word Embeddings (TWE), to learn topic embeddings to characterize various meanings of polysemous words by concatenating topic embeddings

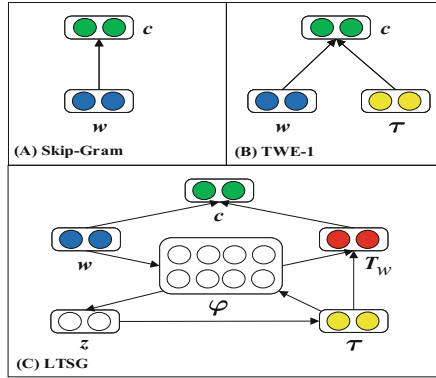


Fig. 1. Skip-Gram, TWE and LTSG models. Blue, yellow, green circles denote the embeddings of word, topic and context, while red circles in LTSG denote the global topical word. White circles denote the topic model part, topic-word distribution φ and topic assignment z . (Color figure online)

with word embeddings. Despite the success of TWE, compared to previous multi-prototype models [11,20], it assumes that word distributions over topics are provided by off-the-shelf topic models such as LDA, which would limit the applications of TWE once topic models do not perform well in some domains [19]. As a matter of fact, pervasive polysemous words in documents would harm the performance of topic models that are based on co-occurrence of words in documents. Thus, a more realistic solution is to build both topic models with regard to polysemous words and polysemous word embeddings simultaneously, instead of using off-the-shelf topic models. In this work, we propose a novel learning framework, called Latent Topical Skip-Gram (LTSG) model, to mutually learn polysemous-word models and topic models. To the best of our knowledge, this is the first work that considers learning polysemous-word models and topic models simultaneously. Although there have been approaches that aim to improve topic models based on word embeddings MRF-LDA [24], they fail to improve word embeddings provided words are polysemous; although there have been approaches that aim to improve polysemous-word models TWE [14] based on topic models, they fail to improve topic models considering words are polysemous. Different from previous approaches, we introduce a new node T_w , called *global topic*, to capture all of the topics regarding polysemous word w based on topic-word distribution φ , and use the global topic to estimate the context of polysemous word w . Then we characterize polysemous word embeddings by concatenating word embeddings with topic embeddings. We illustrate our new model in Fig. 1, where Fig. 1(A) is the skip-gram model [16], which aims to maximize the probability of context c given word w . Figure 1(B) is the TWE model, which extends the skip-gram model to maximize the probability of context c given both word w and topic t , and Fig. 1(C) is our LTSG model which aims to maximize the probability of context c given word w and global topic T_w . T_w is generated based on topic-distrib-

bution φ (i.e., the joint distribution of topic embedding τ and word embedding w) and topic embedding τ (which is based on topic assignment z). Through our LTSG model, we can simultaneously learn word embeddings w and global topic embeddings T_w for representing polysemous word embeddings, and topic word distribution φ for mining topics with regard to polysemous words. We will exhibit the effectiveness of our LTSG model in text classification and topic mining tasks with regard to polysemous words in documents.

In the remainder of the paper, we first introduce preliminaries of our LTSG model, and then present our LTSG algorithm in detail. After that, we evaluate our LTSG model by comparing our LTSG algorithm to state-of-the-art models in various datasets. Finally we review previous work related to our LTSG approach and conclude the paper with future work.

2 Preliminaries

In this section, we briefly review preliminaries of Latent Dirichlet Allocation (LDA), Skip-Gram, and Topical Word Embeddings (TWE), respectively. We show some notations and their corresponding meanings in Table 1, which will be used in describing the details of LDA, Skip-Gram, and TWE.

Table 1. Notations of the text collection.

Term	Notation	Definition or description
Vocabulary	\mathcal{V}	Set of words in the text collection, $ \mathcal{V} = W$
Word	w	A basic item from vocabulary indexed as $w \in \{1, 2, \dots, W\}$
Document	\mathbf{w}	A sequence of N words, $\mathbf{w} = (w_1, w_2, \dots, w_N)$
Corpus	\mathcal{D}	A collection of M documents, $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$
Topic-word	φ	K distributions over vocabulary ($K \times W$ matrix), $ \varphi = K, \varphi_k = W$
Word embedding	v	Distributed representation of <i>word</i> , denoted by v_w , $v \in \mathbb{R}^d$
Topic embedding	τ	Distributed representation of <i>topic</i> , denoted by τ_k , $\tau \in \mathbb{R}^d$

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [2], a three-level hierarchical Bayesian model, is a well-developed and widely used probabilistic topic model. Extending Probabilistic Latent Semantic Indexing (PLSI) [10], LDA adds Dirichlet priors to document-specific topic mixtures to overcome the overfitting problem in PLSI. LDA aims at modeling each document as a mixture over sets of topics, each associated with a multinomial word distribution. Given a document corpus \mathcal{D} , each document $\mathbf{w}_m \in \mathcal{D}$ is assumed to have a distribution over K topics. The generative process of LDA is shown as follows,

1. For each topic $k = 1 \rightarrow K$, draw a distribution over words $\varphi_k \sim Dir(\beta)$
2. For each document $\mathbf{w}_m \in \mathcal{D}, m \in \{1, 2, \dots, M\}$
 - (a) Draw a topic distribution $\theta_m \sim Dir(\alpha)$
 - (b) For each word $w_{m,n} \in \mathbf{w}_m, n = 1, \dots, N_m$
 - i. Draw a topic assignment $z_{m,n} \sim Mult(\theta_m), z_{m,n} \in \{1, \dots, K\}$.
 - ii. Draw a word $w_{m,n} \sim Mult(\varphi_{z_{m,n}})$

where α and β are Dirichlet hyperparameters, specifying the nature of priors on θ and φ . Variational inference and Gibbs sampling are the common ways to learn the parameters of LDA.

2.2 The Skip-Gram Model

The Skip-Gram model is a well-known framework for learning word vectors [16]. Skip-Gram aims to predict context words given a target word in a sliding window, as shown in Fig. 1(A).

Given a document corpus \mathcal{D} defined in Table 1, the objective of Skip-Gram is to maximize the average log-probability

$$\mathcal{L}(\mathcal{D}) = \frac{1}{\sum_{m=1}^M N_m} \sum_{m=1}^M \sum_{n=1}^{N_m} \sum_{-c \leq j \leq c, j \neq 0} \log \Pr(w_{m,n+j} | w_{m,n}), \tag{1}$$

where c is the context window size of the target word. The basic Skip-Gram formulation defines $\Pr(w_{m,n+j} | w_{m,n})$ using the softmax function:

$$\Pr(w_{m,n+j} | w_{m,n}) = \frac{\exp(\mathbf{v}_{w_{m,n+j}} \cdot \mathbf{v}_{w_{m,n}})}{\sum_{w=1}^W \exp(\mathbf{v}_w \cdot \mathbf{v}_{w_{m,n}})}, \tag{2}$$

where $\mathbf{v}_{w_{m,n}}$ and $\mathbf{v}_{w_{m,n+j}}$ are the vector representations of target word $w_{m,n}$ and its context word $w_{m,n+j}$, and W is the number of words in the vocabulary \mathcal{V} . Hierarchical softmax and negative sampling are two efficient approximation methods used to learn Skip-Gram.

2.3 Topical Word Embeddings

Topical word embeddings (TWE) is a more flexible and powerful framework for multi-prototype word embeddings, where topical word refers to a word taking a specific topic as context [14], as shown in Fig. 1(B). TWE model employs LDA to obtain the topic distributions of document corpora and topic assignment for each word token. TWE model uses topic $z_{m,n}$ of target word to predict context word compared with only using the target word $w_{m,n}$ to predict context word in Skip-Gram. TWE is defined to maximize the following average log probability

$$\mathcal{L}(\mathcal{D}) = \frac{1}{\sum_{m=1}^M N_m} \sum_{m=1}^M \sum_{n=1}^{N_m} \sum_{-c \leq j \leq c, j \neq 0} \log \Pr(w_{m,n+j} | w_{m,n}) + \log \Pr(w_{m,n+j} | z_{m,n}). \tag{3}$$

TWE regards each topic as a pseudo word that appears in all positions of words assigned with this topic. When training TWE, Skip-Gram is being used for learning word embeddings. Afterwards, each topic embedding is initialized with the average over all words assigned to this topic and learned by keeping word embeddings unchanged.

Despite the improvement over Skip-Gram, the parameters of LDA, word embeddings and topic embeddings are learned separately. In other word, TWE just uses LDA and Skip-Gram to obtain external knowledge for learning better topic embeddings.

3 Our LTSG Algorithm

Extending from the TWE model, the proposed Latent Topical Skip-Gram model (LTSG) directly integrates LDA and Skip-Gram by using topic-word distribution φ mentioned in topic models like LDA, as shown in Fig. 1(C). We take three steps to learn topic modeling, word embeddings and topic embeddings simultaneously, as shown below.

Step 1. Sample topic assignment for each word token. Given a specific word token $w_{m,n}$, we sample its latent topic $z_{m,n}$ by performing Gibbs updating rule similar to LDA.

Step 2. Compute topic embeddings. We average all words assigned to each topic to get the embedding of each topic.

Step 3. Train word embeddings. We train word embeddings similar to Skip-Gram and TWE. Meanwhile, topic-word distribution φ is updated based on Eq. (10). The objective of this step is to maximize the following function

$$\mathcal{L}(\mathcal{D}) = \frac{1}{\sum_{m=1}^M N_m} \sum_{m=1}^M \sum_{n=1}^{N_m} \sum_{-c \leq j \leq c, j \neq 0} \log \Pr(w_{m,n+j} | w_{m,n}) + \log \Pr(w_{m,n+j} | T_{w_{m,n}}), \quad (4)$$

where $T_{w_{m,n}} = \sum_{k=1}^K \tau_k \cdot \varphi_{k,w_{m,n}} \cdot \tau_k$ indicates the k -th topic embedding. $T_{w_{m,n}}$ can be seen as a distributed representation of global topical word of $w_{m,n}$.

We will address the above three steps in detail below.

3.1 Topic Assignment via Gibbs Sampling

To perform Gibbs sampling, the main target is to sample topic assignments $z_{m,n}$ for each word token $w_{m,n}$. Given all topic assignments to all of the other words, the full conditional distribution $\Pr(z_{m,n} = k | \mathbf{z}^{-(m,n)}, \mathbf{w})$ is given below when applying collapsed Gibbs sampling [9],

$$\Pr(z_{m,n} = k | \mathbf{z}^{-(m,n)}, \mathbf{w}) \propto \frac{n_{k,w_{m,n}}^{-(m,n)} + \beta}{\sum_{w=1}^w n_{k,w}^{-(m,n)} + W\beta} \cdot \frac{n_{m,k}^{-(m,n)} + \alpha}{\sum_{k'=1}^K n_{m,k'}^{-(m,n)} + K\alpha}, \quad (5)$$

where $-(m, n)$ indicates that the current assignment of $z_{m,n}$ is excluded. $n_{k,w}$ and $n_{m,k}$ denote the number of word tokens w assigned to topic k and the count of word tokens in document m assigned to topic k , respectively. After sampling all the topic assignments for words in corpus \mathcal{D} , we can estimate each component of φ and θ by Eqs. (6) and (7).

$$\hat{\varphi}_{k,w} = \frac{n_{k,w} + \beta}{\sum_{w'=1}^W n_{k,w'} + W\beta} \quad (6)$$

$$\hat{\theta}_{d,k} = \frac{n_{m,k} + \alpha}{\sum_{k'=1}^K n_{m,k'} + K\alpha} \quad (7)$$

Unlike standard LDA, the topic-word distribution φ is used directly for constructing the modified Gibbs updating rule in LTSG. Following the idea of DRS [7], with the conjugacy property of Dirichlet and multinomial distributions, the Gibbs updating rule of our model LTSG can be approximately represented by

$$\Pr(z_{m,n} = k | \mathbf{w}, \mathbf{z}^{-(m,n)}, \varphi, \alpha) \propto \varphi_{k,w_{m,n}} \cdot \frac{n_{m,k}^{-(m,n)} + \alpha}{\sum_{k'=1}^K n_{m,k'}^{-(m,n)} + K\alpha}. \quad (8)$$

In different corpus or applications, Eq. (8) can be replaced with other Gibbs updating rules or topic models, eg. LFLDA [18].

3.2 Topic Embeddings Computing

Topic embeddings aim to approximate the latent semantic centroids in vector space rather than a multinomial distribution. TWE trains topic embeddings after word embeddings have been learned by Skip-Gram. In LTSG, we use a straightforward way to compute topic embedding for each topic. For the k th topic, its topic embedding is computed by averaging all words with their topic assignment z equivalent to k , i.e.,

$$\tau_k = \frac{\sum_{m=1}^M \sum_{n=1}^{N_m} \mathbb{I}(z_{m,n} = k) \cdot \mathbf{v}_{w_{m,n}}}{\sum_{w=1}^W n_{k,w}} \quad (9)$$

where $\mathbb{I}(x)$ is indicator function defined as 1 if x is true and 0 otherwise.

Similarly, you can design your own more complex training rule to train topic embedding like TopicVec [13] and Latent Topic Embedding (LTE) [12].

3.3 Word Embeddings Training

LTSG aims to update φ during word embeddings training. Following the similar optimization as Skip-Gram, hierarchical softmax and negative sampling are used for training the word embeddings approximately due to the computationally expensive cost of the full softmax function which is proportional to vocabulary

size W . LTSG uses stochastic gradient descent to optimize the objective function given in Eq. (4).

The hierarchical softmax uses a binary tree (eg. a Huffman tree) representation of the output layer with the W words as its leaves and, for each node, explicitly represents the relative probabilities of its child nodes. There is a unique path from root to each word w and $node(w, i)$ is the i -th node of the path. Let $L(w)$ be the length of this path, then $node(w, 1) = root$ and $node(w, L(w)) = w$. Let $child(u)$ be an arbitrary child of node u , e.g. left child. By applying hierarchical softmax on $\Pr(w_{m,n+j}|T_{w_{m,n}})$ similar to $\Pr(w_{m,n+j}|w_{m,n})$ described in Skip-gram [16], we can compute the log gradient of φ as follows,

$$\frac{\partial \log \Pr(w_{m,n+j}|T_{w_{m,n}})}{\partial \varphi_{k=z_{m,n}, w=w_{m,n}}} = \frac{1}{L(w_{m,n}) - 1} \sum_{i=1}^{L(w_{m,n})-1} \left[1 - h_{i+1}^{w_{m,n+j}} - \sigma(T_{w_{m,n}} \cdot \mathbf{v}_i^{w_{m,n+j}}) \right] \boldsymbol{\tau}_k \cdot \mathbf{v}_i^{w_{m,n+j}}, \quad (10)$$

where $\sigma(x) = 1/(1 + \exp(-x))$. Given a path from root to word $w_{m,n+j}$ constructed by Huffman tree, $\mathbf{v}_i^{w_{m,n+j}}$ is the vector representation of i -th node. And $h_{i+1}^{w_{m,n+j}}$ is the Huffman coding on the path defined as $h_{i+1}^{w_{m,n+j}} = \mathbb{I}(node(w_{m,n+j}, i+1) = child(node(w_{m,n+j}, i))$.

Follow this idea, we can compute the gradients for updating the word w and non-leaf node. From Eq. (10), we can see that φ is updated by using topic embeddings $\boldsymbol{\tau}_k$ directly and word embeddings indirectly via the non-leaf nodes in Huffman tree, which is used for training the word embeddings.

3.4 An Overview of Our LTSG algorithm

In this section we provide an overview of our LTSG algorithm, as shown in Algorithm 1. In line 1 in Algorithm 1, we run the standard LDA with certain iterations and initialize φ based on Eq. (6). From lines 4 to 6, there are the three steps mentioned in Sect. 3. From lines 7 to 13, φ will be updated after training the whole corpus \mathcal{D} rather than per word, which is more suitable for multi-thread training. Function $f(\xi, n_{k,w})$ is a dynamic learning rate, defined by $f(\xi, n_{k,w}) = \xi \cdot \log(n_{k,w})/n_{k,w}$. In line 16, document-topic distribution $\theta_{m,k}$ is computed to model documents.

4 Experiments

In this section, we evaluate our LTSG model in three aspects, i.e., contextual word similarity, text classification, and topic coherence.

We use the dataset 20NewsGroup, which consists of about 20,000 documents from 20 different newsgroups. For the baseline, we use the default settings of parameters unless otherwise specified. Similar to TWE, we set the number of topics $K = 80$ and the dimensionality of both word embeddings and topic embeddings $d = 400$ for all the relative models. In LTSG, we initialize φ with $init_nGS = 2500$. We perform $nItrs = 5$ runs on our framework. We perform $nGS = 200$ Gibbs sampling iterations to update topic assignment with $\alpha = 0.01, \beta = 0.1$.

Algorithm 1. Latent Topical Skip-Gram

Input: corpus \mathcal{D} , # topics K , size of vocabulary W , Dirichlet hyperparameters α, β , # iterations of LDA for initialization $init_nGS$, # iterations of framework $nItrs$, # Gibbs sampling iterations nGS .

Output: $\theta_{m,k}, \varphi_{k,w}, \mathbf{v}_w, \boldsymbol{\tau}_k, m = 1, 2, \dots, M; k = 1, 2, \dots, K; w = 1, 2, \dots, W$

- 1: **Initialization.** Initialize $\varphi_{k,w}$ as in Equation (6) with $init_nGS$ iterations in standard LDA as in Equation (5)
 - 2: $i \leftarrow 0$
 - 3: **while** ($i < nItrs$) **do**
 - 4: **Step 1.** Sample $z_{m,n}$ as in Equation (8) with nGS iterations
 - 5: **Step 2.** Compute each topic embedding $\boldsymbol{\tau}_k$ as in Equation (9)
 - 6: **Step 3.** Train word embeddings with objective function as in Equation (4)
 - 7: Compute the first-order partial derivatives $\mathcal{L}'(\mathcal{D})$
 - 8: Set the learning rate ξ
 - 9: **for** ($k = 1 \rightarrow K$) **do**
 - 10: **for** ($w = 1 \rightarrow W$) **do**
 - 11: $\varphi_{k,w}^{(i+1)} \leftarrow \varphi_{k,w}^{(i)} + f(\xi, n_{k,w}) \frac{\partial \mathcal{L}'(\mathcal{D})}{\partial \varphi_{k,w}}$
 - 12: **end for**
 - 13: **end for**
 - 14: $i \leftarrow i + 1$
 - 15: **end while**
 - 16: Compute each $\theta_{m,k}$ as in Equation (7)
-

4.1 Contextual Word Similarity

To evaluate contextual word similarity, we use Stanford’s Word Contextual Word Similarities (SCWS) dataset introduced by [11], which has been also used for evaluating state-of-art model [14]. There are totally 2,003 word pairs and their contexts, including 1328 noun-noun pairs, 399 verb-verb pairs, 140 verb-noun, 97 adjective-adjective, 30 noun-adjective, 9 verb-adjective pairs. Among all of the pairs, there are 241 same-word pairs which may show different meaning in the giving context. The dataset provide human labeled similarity scores based on the meaning in the context. For comparison, we compute the Spearman correlation similarity scores of different models and human judgments.

Following the TWE model, we use two scores **AvgSimC** and **MaxSimC** to evaluate the multi-prototype model for contextual word similarity. The topic distribution $\Pr(z|w, c)$ will be inferred by using $\Pr(z|w, c) \propto \Pr(w|z) \Pr(z|c)$ with regarding c as a document. Given a pair of words with their contexts, namely (w_i, c_i) and (w_j, c_j) , **AvgSimC** aims to measure the averaged similarity between the two words all over the topics:

$$AvgSimC = \sum_{z, z' \in K} \Pr(z|w_i, c_i) \Pr(z'|w_j, c_j) S(\mathbf{v}_{w_i}^z, \mathbf{v}_{w_j}^{z'}) \quad (11)$$

where \mathbf{v}_w^z is the embedding of word w under its topic z by concatenating word and topic embeddings $\mathbf{v}_w^z = \mathbf{v}_w \oplus \boldsymbol{\tau}_z$. $S(\mathbf{v}_{w_i}^z, \mathbf{v}_{w_j}^{z'})$ is the cosine similarity between $\mathbf{v}_{w_i}^z$ and $\mathbf{v}_{w_j}^{z'}$.

MaxSimC selects the corresponding topical word embedding \mathbf{v}_w^z of the most probable topic z inferred using w in context c as the contextual word embedding, defined as

$$\text{MaxSimc} = S(\mathbf{v}_{w_i}^z, \mathbf{v}_{w_j}^{z'}) \quad (12)$$

where

$$z = \arg \max_z \Pr(z|w_i, c_i), \quad z' = \arg \max_z \Pr(z|w_j, c_j).$$

We consider the two baselines Skip-Gram and TWE. Skip-Gram is a well-known single prototype model and TWE is the state-of-the-art multi-prototype model. We use all the default settings in these two model to train the 20NewsGroup corpus.

Table 2. Spearman correlation $\rho \times 100$ of contextual word similarity on the SCWS dataset.

Model	$\rho \times 100$	
Skip-Gram	51.1	
LTSG-word	53.4	
	AvgSimC	MaxSimC
TWE	52.0	49.2
LTSG	54.2	54.1

From Table 2, we can see that LTSG achieves better performance compared to the two competitive baseline. It shows that topic model can actually help improving polysemous-word model, including word embeddings and topic embeddings. The meaning of a word is certain by giving its specify context so that **MaxSimC** is more relative to real application. Then LTSG model achieves more improvement in **MaxSimC** than **AvgSimC** compared to TWE, which tells that LTSG could perform better in telling a word meaning in specify context.

4.2 Text Classification

In this sub-section, we investigate the effectiveness of LTSG for document modeling using multi-class text classification. The 20NewsGroup corpus has been divided into training set and test set with ratio 60% to 40% for each category. We calculate macro-averaging precision, recall and F1-score to measure the performance of LTSG.

We learn word and topic embeddings on the training set and then model document embeddings for both training set and testing set. Afterwards, we consider document embeddings as document features and train a linear classifier using Liblinear [8]. We use \mathbf{v}_m , $\boldsymbol{\tau}_k$, \mathbf{v}_w to represent document embeddings, topic embeddings, word embeddings, respectively, and model documents on both topic-based and embedding-based methods as shown below.

Table 3. Evaluation results of multi-class text classification.

Model	Accuracy	Precision	Recall	F1-score
BOW	79.7	79.5	79.0	79.2
LDA	72.2	70.8	70.7	70.7
Skip-Gram	75.4	75.1	74.7	74.9
TWE	81.5	81.2	80.6	80.9
LTSG-theta	74.1	73.1	72.7	72.9
LTSG-topic	74.8	74.0	73.3	73.7
LTSG-word	81.4	81.0	80.4	80.7
LTSG	82.7	82.5	81.7	82.1

Table 4. Top words of some topics from LTSG and LDA on 20NewsGroup for $K = 80$.

LTSG	LDA	LTSG	LDA	LTSG	LDA	LTSG	LDA
image	image	jet	printer	stimulation	doctor	anonymous	list
jpeg	files	ink	good	diseases	disease	faq	mail
gif	color	laser	print	disease	coupons	send	information
format	gif	printers	font	toxin	treatment	ftp	internet
files	jpeg	deskjet	graeme	icts	pain	mailing	send
file	file	ssa	laser	newsletter	medical	server	posting
convert	format	printer	type	staffed	day	mail	email
color	bit	noticeable	quality	volume	microorganisms	alt	group
formats	images	canon	printers	health	medicine	archive	news
images	quality	output	deskjet	aids	body	email	onymous
-75.66	-88.76	-91.53	-119.28	-66.91	-100.39	-78.23	-95.47

- **LTSG-theta.** Document-topic distribution θ_m estimated by Eq. (7).
- **LTSG-topic.** $\mathbf{v}_m = \sum_{k=1}^K \theta_{m,k} \cdot \boldsymbol{\tau}_k$.
- **LTSG-word.** $\mathbf{v}_m = (1/N_m) \sum_{n=1}^{N_m} \mathbf{v}_{w_{m,n}}$.
- **LTSG.** $\mathbf{v}_m = (1/N_m) \sum_{n=1}^{N_m} \mathbf{v}_{w_{m,n}}^{z_{m,n}}$, where contextual word is simply constructed by $\mathbf{v}_{w_{m,n}}^{z_{m,n}} = \mathbf{v}_{w_{m,n}} \oplus \boldsymbol{\tau}_{z_{m,n}}$.

Result Analysis. We consider the following baselines, bag-of-word (BOW) model, LDA, Skip-Gram and TWE. The BOW model represents each document as a bag of words and use TFIDF as the weighting measure. For the TFIDF model, we select top 50,000 words as features according to TFIDF score. LDA represents each document as its inferred topic distribution. In Skip-Gram, we build the embedding vector of a document by simply averaging over all word embeddings in the document. The experimental results are shown in Table 3.

From Table 3, we can see that, for topic modeling, LTSG-theta and LTSG-topic perform better than LDA slightly. For word embeddings, LTSG-word

significantly outperforms Skip-Gram. For topic embeddings using for multi-prototype word embeddings, LTSG also outperforms state-of-the-art baseline TWE. This verifies that topic modeling, word embeddings and topic embeddings can indeed impact each other in LTSG, which lead to the best result over all the other baselines.

4.3 Topic Coherence

In this section, we evaluate the topics generated by LTSG on both quantitative and qualitative analysis. Here we follow the same corpus and parameters setting in Sect. 4.2 for LSTG model.

Quantitative Analysis. Although perplexity (held-out likelihood) has been widely used to evaluate topic models, [3] found that perplexity can be hardly to reflect the semantic coherence of individual topics. Topic Coherence metric [17] was found to produce higher correlation with human judgments in assessing topic quality, which has become popular to evaluate topic models [1, 4]. A higher topic coherence score indicates a more coherent topic.

We compute the score of the top 10 words for each topic. We present the score for some of topics in the last line of Table 4. By averaging the score of the total 80 topics, LTSG gets -92.23 compared with -108.72 of LDA. We can conclude that LTSG performs better than LDA in finding higher quality topics.

Qualitative Analysis. Table 4 shows top 10 words of topics from LTSG and LDA model on 20NewsGroup. The words in this two models are ranked based on the probability distribution φ for each topic. As shown, LTSG is able to capture more concrete topics compared with general topics in LDA. For the topic about “image”, LTSG shows about image conversion on different format, while LDA shows the image quality of different format. In topic “printer”, LTSG emphasizes the different technique of printer in detail and LDA generally focus on “good quality” of printing.

5 Related Work

Recently, researches on cooperating topic models and vector representations have made great advances in NLP community. [24] proposed a Markov Random Field regularized LDA model (MRF-LDA) which encourages similar words to share the same topic for learning more coherent topics. [6] proposed Gaussian LDA to use pre-trained word embeddings in Gibbs sampler based on multivariate Gaussian distributions. LFLDA [18] is modeled as a mixture of the conventional categorical distribution and an embedding link function. These works have given the faith that vector representations are capable of helping improving topic models. On the contrary, vector representations, especially topic embeddings, have been promoted for modeling documents or polysemy with great help of topic models.

For examples, [14] used topic model to globally cluster the words into different topics according to their context for learning better multi-prototype word embeddings. [13] proposed generative topic embedding (TopicVec) model that replaces categorical distribution in LDA with embedding link function. However, these models do not show close interactions among topic models, word embeddings and topic embeddings. Besides, these researches lack of investigation on the influence of topic model on word embeddings.

6 Conclusion and Future Work

In this paper, we propose a basic model Latent Topical Skip-Gram (LTSG) which shows that LDA and Skip-Gram can mutually help improve performance on different task. The experimental results show that LTSG achieves the competitive results compared with the state-of-art models.

We consider the following future research directions: (I) We will investigate non-parametric topic models [22] and parallel topic models [15] to set parameters automatically and accelerate training using multi threading for large-scale data. (II) We will construct a package which can be convenient to extend with other topic models and word embeddings models to our framework by using the interfaces. (III) We will deal with unseen words in new documents like Gaussian LDA [6].

Acknowledgments. We thank all reviewers for their valuable comments and feedback that greatly improved our paper. Zhuo thanks the National Key Research and Development Program of China (2016YFB0201900), National Natural Science Foundation of China (U1611262), Guangdong Natural Science Funds for Distinguished Young Scholar (2017A030306028), Pearl River Science and Technology New Star of Guangzhou, and Guangdong Province Key Laboratory of Big Data Analysis and Processing for the support of this research.

References

1. Arora, S., et al.: A practical algorithm for topic modeling with provable guarantees. In: ICML, pp. 280–288 (2013)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *JMLR* **3**, 993–1022 (2003)
3. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: NIPS, pp. 288–296 (2009)
4. Chen, Z., Liu, B.: Topic modeling using topics from many domains, lifelong learning and big data. In: ICML, pp. 703–711 (2014)
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.P.: Natural language processing (almost) from scratch. *JMLR* **12**, 2493–2537 (2011)
6. Das, R., Zaheer, M., Dyer, C.: Gaussian LDA for topic models with word embeddings. In: ACL, pp. 795–804 (2015)
7. Du, J., Jiang, J., Song, D., Liao, L.: Topic modeling with document relative similarities. In: IJCAI 2015, pp. 3469–3475 (2015)

8. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: a library for large linear classification. *JMLR* **9**, 1871–1874 (2008)
9. Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proc. Nat. Acad. Sci.* **101**(Suppl. 1), 5228–5235 (2004)
10. Hofmann, T.: Probabilistic latent semantic indexing. In: *SIGIR 1999*, pp. 50–57 (1999)
11. Huang, E.H., Socher, R., Manning, C.D., Ng, A.Y.: Improving word representations via global context and multiple word prototypes. In: *ACL*, pp. 873–882 (2012)
12. Jiang, D., Shi, L., Lian, R., Wu, H.: Latent topic embedding. In: *COLING*, pp. 2689–2698 (2016)
13. Li, S., Chua, T., Zhu, J., Miao, C.: Generative topic embedding: a continuous representation of documents. In: *ACL* (2016)
14. Liu, Y., Liu, Z., Chua, T., Sun, M.: Topical word embeddings. In: *AAAI*, pp. 2418–2424 (2015)
15. Liu, Z., Zhang, Y., Chang, E.Y., Sun, M.: PLDA+: parallel latent dirichlet allocation with data placement and pipeline processing. *ACM TIST* **2**(3), 26 (2011)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *NIPS*, pp. 3111–3119 (2013)
17. Mimno, D.M., Wallach, H.M., Talley, E.M., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: *EMNLP*, pp. 262–272 (2011)
18. Nguyen, D.Q., Billingsley, R., Du, L., Johnson, M.: Improving topic models with latent feature word representations. *TACL* **3**, 299–313 (2015)
19. Phan, X.H., Nguyen, C., Le, D., Nguyen, M.L., Horiguchi, S., Ha, Q.: A hidden topic-based framework toward building applications with short web documents. *IEEE Trans. Knowl. Data Eng.* **23**(7), 961–976 (2011)
20. Reisinger, J., Mooney, R.J.: Multi-prototype vector-space models of word meaning. In: *NAACL*, pp. 109–117 (2010)
21. Socher, R., Bauer, J., Manning, C.D., Ng, A.Y.: Parsing with compositional vector grammars. In: *ACL*, pp. 455–465 (2013)
22. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. *J. Am. Stat. Assoc.* **101**(476), 1566–1581 (2006)
23. Turian, J.P., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: *ACL 2010*, pp. 384–394 (2010)
24. Xie, P., Yang, D., Xing, E.P.: Incorporating word correlation knowledge into topic modeling. In: *NAACL*, pp. 725–734 (2015)



Improve the Spoofing Resistance of Multimodal Verification with Representation-Based Measures

Zengxi Huang¹, Zhen-Hua Feng², Josef Kittler², and Yiguang Liu³(✉)

¹ School of Computer and Software Engineering,
Xihua University, Chengdu, China
luomull17@hotmail.com

² Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, UK

³ College of Computer Science, Sichuan University, Chengdu, China
lygpapers@aliyun.com

Abstract. Recently, the security of multimodal verification has become a growing concern since many fusion systems have been known to be easily deceived by partial spoof attacks, i.e. only a subset of modalities is spoofed. In this paper, we verify such a vulnerability and propose to use two representation-based measures to close this gap. Firstly, we use the collaborative representation fidelity with non-target subjects to measure the affinity of a query sample to the claimed client. We further consider sparse coding as a competing comparison among the client and the non-target subjects, and hence explore two sparsity-based measures for recognition. Last, we select the representation-based measure, and assemble its score and the affinity score of each modality to train a support vector machine classifier. Our experimental results on a chimeric multimodal database with face and ear traits demonstrate that in both regular verification and partial spoof attacks, the proposed method significantly outperforms the well-known fusion methods with conventional measure.

Keywords: Multimodal verification · Spoof attacks
Representation-based measure · Support vector machine

1 Introduction

A generic biometric system has eight vulnerable points that can be exploited by an intruder to gain unauthorized access [1]. Among them, spoof attacks usually present a counterfeited biometric sample (e.g., a gummy fingerprint, a face image/video/mask) to a system sensor, which do not require knowledge about the system's operational mechanism and internal parameters. Spoof attacks are also known as non-zero effort attacks, presentation attacks, and direct attacks. The concept of non-zero effort attacks is relative to zero effort attempts, where an imposter doesn't fabricate the biometric trait of any specific client and merely presents his/her own biometric trait to the system. In the literature, an imposter is generally regarded as an intruder who performs zero effort attempts. In this paper, for clarity and terminological consistence, a legitimate claim,

zero effort attempt, and non-zero effort attack are termed as genuine, imposter and spoof, respectively, together with their associated executor/sample/score.

Multimodal systems have been considered intrinsically more secure than unimodal systems based on the intuition that an intruder would have to spoof all the biometric traits to successfully impersonate the targeted client [2]. Such a belief has long been established disregarding the possibility that an intruder is falsely accepted by spoofing only a subset of the biometric traits. The vulnerability of multimodal systems to partial spoof attacks has been shown in the worst-case scenario, where the intruder is assumed to be able to replicate a subset of the biometric traits of a genuine client exactly. Under this assumption, Rodrigues [3] showed experimental results on chimeric multimodal databases with face and fingerprint that multimodal systems can be deceived easily by spoofing only a subset of the modalities, if the fusion rule is not designed with any anti-spoofing measure. Wild et al. [4] showed the sensitivity of multimodal systems to partial spoof attacks with real fake biometric databases.

Some efforts to enhance the security of multimodal systems against partial spoof attacks have already been reported. Rodrigues et al. [5] proposed a modification of the classic likelihood ratio (LLR) method that considers the possibility of spoof attacks and the degree of security to individual trait when modelling score distributions. However, these prior probabilities are application dependent and may not be time invariant, hence are quite difficult to quantify. Rodrigues et al. [3] also proposed the idea of using quality measures to protect against spoof attacks. Intuitively, a fake biometric sample is likely to be of inferior quality. However, biometric quality assessment is still an open issue to most biometrics. Besides, fake biometric sample is not necessarily to be inferior with the emerging image/video synthesis, 3D printing, and materials.

Liveness detection is another kind of approach used to improve the spoofing resistance for a given system. Marasco et al. [6] proposed a multimodal system that incorporates a liveness detection algorithm to reject spoofed samples. If a spoof attempt is indicated, the related modality matching score is ignored. Wild et al. [4] combined the recognition score and liveness measure at score level with a 1-median filtering scheme for enhanced tolerance to spoof attacks. Nevertheless, neither one of hardware-based and software-based liveness detection systems have shown acceptable performance and cost against spoof attacks. Physiological and behavioral characteristics are also employed to enhance multimodal verification security in [7].

This paper is enlightened by the fact that in a partial spoof attack, the recognition scores achieved from non-spoofed modalities are generally near the imposter score distribution center, given that they are also zero effort attempts from a unimodal viewpoint. Unlike the quality- and/or liveness-based methods that focus on the spoofed modalities, we propose to take advantage of non-spoofed modalities. To this end, we put forward a representation-based measure to gauge the affinity of a query sample to a claimed client. This is based on the assumption that a biometric sample would result in inferior sparse representation fidelity if it doesn't lie in any subspace spanned by the samples from the same subject [8–10]. Note that, it is unlikely to exhaustively collect the representative samples per subject to construct a class specific overcomplete dictionary. We propose to build the dictionary together with samples from non-target subjects to collaboratively represent a query sample.

This affinity score could be an additional measure to a traditional verification method. However, we further consider sparse coding as a one-to-many comparison among the claimed client and non-target subjects, and hence explore other sparsity-based measures for verification. We evaluate two measures, namely, sparse coding error (SCE) and sparse contribution rate (SCR), on a multimodal database with face and ear. Encouraging performance of SCE-based and SCR-based Sum fusion methods evidently supports the usage of sparsity-based one-to-many comparisons in multimodal verification. However, SCR shows much more inferior performance in spoof attacks. Last, we assemble the proposed affinity score and SCE score of each modality as an input vector to train a support vector machine (SVM) classifier.

To validate the effectiveness of the proposed method, we construct a chimeric multimodal database with face and ear traits. The proposed method is compared with the well-known multimodal methods like LLR, SVM, and Sum fusion that are based on cosine similarity. The experimental results validate that in both no spoof and partial spoof cases, the proposed method significantly outperforms its competitors. For example, the traditional methods get the best equal error rates (EER) of 8.32% and 11.89% in no spoof and spoof cases, while our method achieves 0.27% and 2.12%. Apparently, the proposed method helps to increase the spoofing resistance of multimodal systems.

The remainder of the paper is structured as follows. We discuss the approaches to verification based on one-to-many match, and we review the existing methods using sparse coding in Sect. 2. In Sect. 3, we present the sparsity-based affinity and recognition measures, together with the proposed multimodal verification system. In Sect. 4, we describe our chimeric multimodal database and report the corresponding experimental results. The conclusion is drawn in Sect. 5.

2 Related Work

In a biometric verification system, an individual who desires to be recognized claims an identity and presents biometric samples. Then the system conducts a comparison to determine whether the claim is licit or not. Verification is used for positive recognition, where the aim is to prevent multiple people from using the same identity.

Typically, biometric verification systems conduct a one-to-one match that compares a query image against the gallery template(s), whose identity is being claimed. The comparison produces a similarity score. The system accepts the claim if the score is higher than an operating threshold, otherwise rejects it. The operating threshold is determined in the training phase based on the genuine and imposter score distributions. However, it is unlikely to collect all the representative samples of a client that cover all possible variations, for example, expression, pose, illumination, aging, and occlusion in face. Under such circumstances, it cannot be guaranteed that no imposter score is higher than the predefined operating threshold. The system is at a risk of being cracked by intruders. Therefore, the one-to-one match solely based on a predetermined operating threshold is problematic.

Two decades ago, Verlinde et al. [11] proposed a one-to-many match biometric verification method using a k-NN classifier. To the best of our knowledge, this is one of

the first attempts to consider non-target subjects for verification in the test phase. Nevertheless, the inferior comparison algorithm like k-NN could probably account for the rare use of one-to-many match in verification. Cohort-based score normalization also takes advantage of non-target subjects but serves the traditional one-to-one match verification [12]. In recent years, we have witnessed the great success of sparse coding techniques in biometric recognition [13–15]. The sparse representation-based classification (SRC) conducts one-to-many comparisons in a sparse coding procedure and is naturally applicable to biometric identification. Note that, along with the initial research of SRC-based face identification in [13], a measure called sparse concentration index (SCI) was applied to reject outliers, i.e. the subjects who do not appear in dictionary.

Inspired by the success of SRC identification and sparsity-based outlier verification, SRC-based comparison has been introduced in speaker verification. In [16], GMM mean supervector is used as feature of an utterance. The L_1 -norm value of the representation coefficients associated with the claimed identity is used as genuine score, while the L_1 -norm of the coefficients of each other non-target subject are imposter scores. Based on a similar idea, Li et al. [17] created the dictionary using the total variability i-vectors and evaluated three sparsity-based measures for speaker verification, which achieved better results than a SVM baseline.

3 The Proposed Method

3.1 Affinity Measure

In this section, we present a representation-based measure to gauge the affinity of a query sample to a claimed client, based on the assumption that a biometric sample would result in inferior sparse representation fidelity if it doesn't lie in the subspace spanned by the samples from the same subject [8, 9]. Note that, it is unlikely to exhaustively collect the representative samples per subject to construct a class specific overcomplete dictionary. A feasible way is to use non-target subjects to collaboratively represent the query samples [18].

Therefore, we select a number of non-target subjects together with the claimed client. Their gallery samples/features are used to construct an overcomplete dictionary $\mathbf{A} = [\mathbf{A}_c, \mathbf{A}_b] \in \mathbb{R}^{M \times N}$ ($M \ll N$). The first sub-dictionary $\mathbf{A}_c = [\mathbf{a}_{c,1}, \mathbf{a}_{c,2}, \dots, \mathbf{a}_{c,n}] \in \mathbb{R}^{M \times n}$ is composed of the gallery samples of the claimed client, which is a dynamic part of the dictionary. The other sub-dictionary $\mathbf{A}_b = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \dots, \mathbf{a}_{(N-n)}] \in \mathbb{R}^{M \times (N-n)}$ consists of the samples of non-target subjects. Without any specific instructions, \mathbf{A}_b is fixed for all identity verification processes. Given a query sample \mathbf{y} , if it is from a genuine client and isn't of inferior quality, \mathbf{y} should lie in a subspace spanned by \mathbf{A}_c . In this context, \mathbf{y} can be sparsely represented by $\mathbf{y} = \mathbf{A}\boldsymbol{\alpha}$ with high fidelity (see the genuine distribution in Fig. 1), where $\boldsymbol{\alpha} \in \mathbb{R}^N$ is the coefficient vector. A sparse solution of $\boldsymbol{\alpha}$ can be obtained by the following optimization problem [13]:

$$\hat{\boldsymbol{\alpha}} = \arg \min \|\boldsymbol{\alpha}\|_1 \text{ s. t. } \|\mathbf{y} - \mathbf{A}\boldsymbol{\alpha}\|_2 < \varepsilon, \quad (1)$$

where $\|\cdot\|_1$ denotes the L_1 -norm, and $\varepsilon > 0$ is a positive constant.

In a partial spoof attack, a query sample of non-spoofed modalities is unlikely to lie in any subspace spanned by the dictionary samples given that the non-target subjects are confidential. In this context, only a solution with inferior collaborative representation fidelity (CRF), described in Eq. (2), can be found by optimizing Eq. (1).

$$F(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|_2. \quad (2)$$

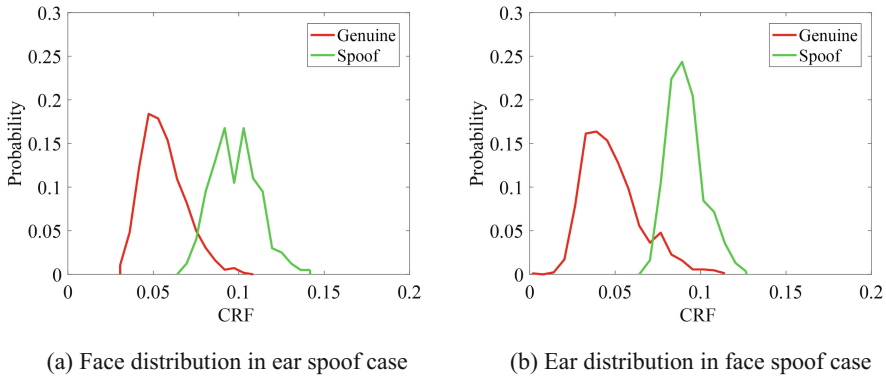


Fig. 1. CRF distributions in partial spoof attacks.

Figure 1 shows the CRF distributions on a chimeric multimodal database using face and ear, detailed in Sect. 4. When the ear of a client is spoofed, the intruder needs to show his/her face or an arbitrary face to complete the biometric data enrollment. Such arbitrary face is unlikely to be from the non-target subjects since the combination of the overcomplete dictionary is confidential. In this context, the non-spoofed face is an outlier that does not lie in the subspace spanned by \mathbf{A} and hence leads to an inferior CRF score, see in Fig. 1(a). When the face is spoofed, we see similar CRF distribution of the non-spoofed ears in Fig. 1(b). From the perspective of the client, CRF score can be used to represent the affinity of the query sample to it.

3.2 Sparsity-Based Recognition Scores

We consider sparse coding as a competing comparison among the client and non-target subjects, and hence explore other two sparsity-based measures, namely, sparse coding error (SCE) and sparse contribution rate (SCR), for multimodal verification.

Since $\hat{\mathbf{x}}$ is achieved in Eq. (1), the SCE value is calculated by

$$E(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}_c \delta_c(\hat{\mathbf{x}})\|_2, \quad (3)$$

where $\delta_c: R^N \rightarrow R^N$ is the characteristic function that selects the coefficients associated with the claimed client.

The well-known SRC and most of its extensions identify a query sample based on comparing the SCEs of all classes in dictionary. Their superior classification

performance validates that SCE is a good candidate to measure the correlation between a query sample and a specific class, as a distance score. Thus, it is reasonable to use SCE for verification.

Wright et al. [13] presented a measure called sparse concentration index (SCI) to reject outliers in face identification. Essentially, the SCI value depends on the class who contributes the most in sparse coding. Given a query sample that isn't an outlier, it generally belongs to the class with the maximal sparse contribution rate (SCR), as defined in Eq. (4). A large value of SCR obtained by a class indicates a greater possibility of the query sample belonging to this class. Therefore, SCR could possibly be used as a similarity score for verification.

$$R(\hat{\alpha}) = \|\delta_c(\hat{\alpha})\|_1 / \|\hat{\alpha}\|_1. \tag{4}$$

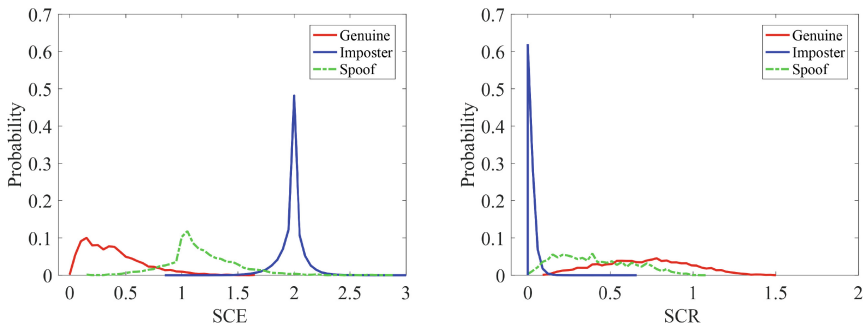


Fig. 2. The distributions of SCE and SCR with Sum fusion on our multimodal dataset.

Figure 2 plots the distributions of SCE and SCR scores obtained on the proposed chimeric multimodal database of face and ear. For convenience to illustrate the effectiveness of SCE and SCR in multimodal verification, we use the Sum rule to fuse face and ear scores. As for SCE, the distribution centers of the genuine and imposter scores are far away from each other with little overlap. Although there is no a clear distribution center peak of the genuine SCR, the overlap is not evident as well. More experimental evidence supporting SCE and SCR is shown in Sect. 4. In addition, Fig. 2 also demonstrates that most spoof scores are located between the distribution centers of genuine and imposter scores. This implies that the multimodal fusion methods based on SCE or SCR are vulnerable to spoof attacks.

Some variants of SCE and SCR have been used in speaker verification and shown to achieve comparable performance with the traditional one-to-one verification. However, in our face and ear unimodal experiments, a genuine client might lose his/her chance to obtain an eligible SCE or SCR score in the competing comparison, owing to the variations in query samples. If it happens, the genuine score will be extremely low. It means that many licitly claimed clients could not pass the verification system by tuning a client specific operating threshold. Instead, more user cooperation will be necessary, which would degrade the user experience. Therefore, for high accuracy and

user convenience of identity verification, sparsity-based one-to-many comparisons would be rather preferable in multimodal scenarios rather than in unimodal applications.

3.3 Multimodal Verification

The CRF score that measures the affinity of a query sample to its claimed client can be utilized to enhance the system’s resistance to partial spoof attacks in a serial or parallel fusion mode. In a serial fusion mode, multimodal systems firstly examine the CRF scores of each modality to determine whether they are spoofed or not, and then conduct multimodal verification.

However, as shown in Fig. 1, the overlap of the genuine and the spoof CRF score distribution is still rather obvious. A hard CRF threshold would lead to high false acceptance rate (FAR), while a loose one may compromise the multimodal system. Note that, there is a high possibility that the non-spoofed modalities get inferior recognition scores along with inferior CRF scores from the same sparse coding. The CRF score and sparsity-based recognition score are complementary. Hence, it is worthwhile to combine them in a parallel way to achieve better performance.

Two sparsity-based recognition scores, i.e., SCE and SCR, are introduced in Sect. 3.2. Both the Sum fusion methods based on them get promising verification performance in zero effort attempts, as shown by the distributions in Fig. 2. These results support the use of the sparsity-based one-to-many comparison in multimodal systems. On the other hand, SCR is much more inferior to SCE in spoof attacks. The detailed experimental results will be given in Sect. 4.

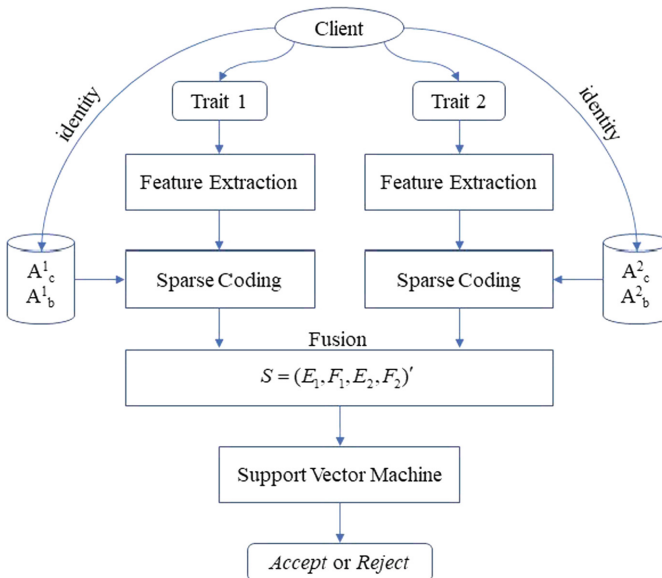


Fig. 3. An overview of the multimodal system architecture.

Last, we select the SCE and CRF scores of each modality to form a score vector for a verification claim. Suppose there are K modalities, e_k and f_k are the SCE and CRF scores of the k^{th} modality. The final score vector can be denoted by $S = (e_1, f_1, e_2, f_2, \dots, e_K, f_K)'$. In the training phase, we use genuine, imposter, and spoof score samples to train a SVM classifier with RBF kernel. For simplicity but without the loss of generality, an overview of system architecture with two modalities ($K = 2$) is shown in Fig. 3 to illustrate the proposed method.

The chimeric multimodal database introduced in Sect. 4 contains 79 subjects with 7 gallery samples each. All these samples are used to form an overcomplete dictionary with 553 atoms. We don't have abundance data to discuss how to optimally select the non-target subjects in this paper. Note that, we ignore the issue of score normalization, given that the scores of face and ear are compatible in our experiments.

4 Experiments and Discussion

4.1 Databases

The proposed method is general for verification using multiple biometric traits. In this paper, we construct a chimeric multimodal database with publicly available face and ear databases. All the 79 subjects in USTB III ear database [19] are randomly paired with the first 79 subjects of AR face database [20]. For each subject, the 7 face images without occlusion of Session 1 are used as gallery samples, while the same type of 7 images of Session 2 are used as probe samples. The USTB III is a multi-view ear database with 20 images per subject. We use the same gallery and probe partition rule in [8, 9], where 7 ear gallery images and 13 ear probe images are selected for each subject. In our experiments, the 2 probe images per subject with extreme pose variation are discarded. For each subject on the multimodal database, in the gallery set, 7 face images are uniquely paired with the 7 ear images to form $79 \times 7 = 553$ multimodal samples. In the probe set, each face image is paired with all the ear images to form $79 \times 7 \times 11 = 6083$ multimodal samples.

To simulate the worst-case partial spoof attacks, in a face spoof case, we replace the ear part of a multimodal sample with the image of USTB II ear database (77 subjects, 4 images per subject) [19]. In an ear spoof case, we replace the face part with the image of Georgia Tech face database (GT, 50 subjects, 8 images per subject) [21]. Finally, we get 77 subjects, 28 face spoof multimodal samples per subject, and 50 subjects, 88 ear spoof multimodal samples per subject.

In the experiments, we use the features of gallery samples of all 79 subjects to construct the overcomplete dictionary. The SCE, SCR, and CRF scores are derived from the comparison between one-sample and one-set. The numbers of genuine, imposter and spoof score samples are 6083, 474474 (6083×78), and 6556, respectively. As for the competing methods using cosine similarity, we empirically select the best match score from each comparison, hence their score sample numbers are the same.

4.2 Settings

The 2D-DCT method is applied for feature extraction of face and ear images, since it is fast, general, and without specific training. The DCT coefficients are scanned in a zigzag manner starting from the top-left corner of the entire transformed image to form a feature vector with 200 dimensions.

The proposed multimodal method uses SVM with RBF kernel ($\sigma = 0.25$). It is compared with the Sum fusion methods of SCE and SCR, denoted by SUM(sce) and SUM(scr), respectively. The competing multimodal methods include the well-known LLR [22], SVM [23], and Sum fusion methods, which use cosine similarity and are respectively denoted by LLR(cos), SVM(cos), and SUM(cos). SVM(cos) also uses RBF kernel ($\sigma = 1$).

Without specific instructions, half of the genuine, imposter and spoof scores are randomly selected for training, and the remainder are for testing. To alleviate the imbalance of training samples, SVM-based classifiers use 1/10 imposters to train. The LLR(cos) uses half of all kinds of samples to fit Gaussian mixture models for score distribution estimation. We run all experiments 5 times, the results presented here are based on the average from these 5 runs.

5 Results

The metrics like false acceptance rate (FAR), false rejection rate (FRR), equal error rate (EER), and the receiver operating characteristic (ROC) curves are generally used to evaluate methods in regular verification. The spoof FAR (SFAR) is specifically used to note the FAR in spoof attacks.

In the first part of the experiments, we train all the learning-based classifiers without considering the spoof samples, namely Regular training. Figure 4 plots the ROC curves of all competing methods in regular verification. The methods with sparsity-based measures are observed to be significantly better than the methods with traditional measure. Among the former methods, SUM(scr) is obviously inferior to SUM(sce) and the proposed method. The ROC curves and the EERs summarized in Table 1 do not show evident advantage of our method when compared with SUM(sce).

Table 1. Performance in terms of EER (%).

Training	Testing	SUM(cos)	SVM(cos)	LLR(cos)	SUM(sce)	SUM(scr)	Ours
Regular training	Regular	11.83	6.632	6.85	0.20	0.39	0.18
	Spoof attacks	12.44	22.05	21.04	8.73	28.26	4.13
Spoof training	Regular	11.83	8.79	8.32	0.20	0.39	0.27
	Spoof attacks	12.44	11.89	12	8.73	28.26	2.12

Figure 5(a) demonstrates that all these methods without spoof training are vulnerable to partial spoof attacks. Both the EERs of LLR(cos) and SVM(cos) increase by about 15%, and even that of SUM(scr) soars to 28.26%. On the other hand, our method achieves a 4.13% EER, which is less than half of the second best.

In the second part of the experiments, all the learning-based classifiers are trained with genuine, imposter and spoof samples, namely spoof training. We can see from Table 1 that, compared with the former experiments of spoof attacks, both LLR(cos) and SVM(cos) get about 10% improvements, while the EER of ours reduces by half, down to 2.12%. The overwhelming advantage of our method can be seen vividly with the ROC curves plotted in Fig. 5(b). It is quite promising provided that the experiments here are in the worst-case spoof conditions where the fake score distribution of the spoofed modalities is identical to that of genuine.

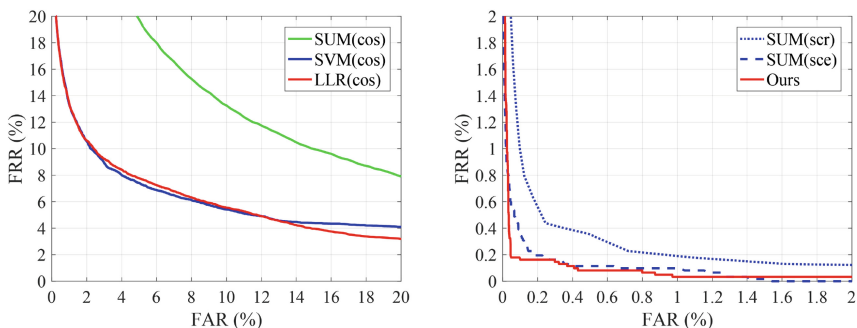


Fig. 4. Performance in regular verification.

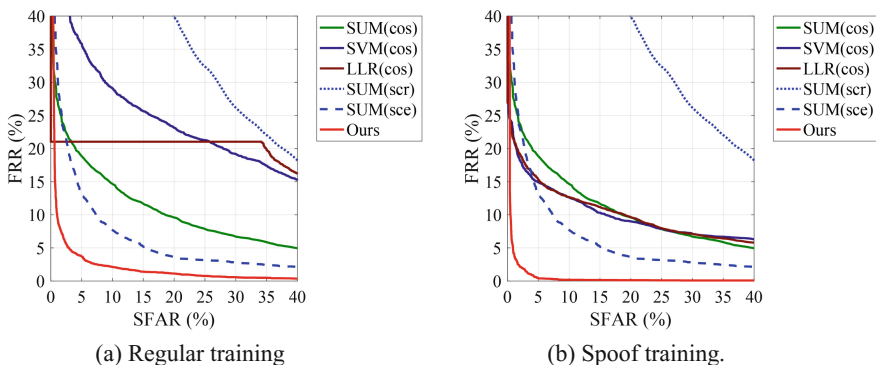


Fig. 5. Performance in partial spoof attacks.

Although LLR(cos) and SVM(cos) also exhibit obvious improvements, they encounter obvious accuracy decline in regular verification, see Table 1. These results show again that the spoof training may bring about unacceptable performance

degradation in regular identity verification [2]. As for the proposed method, the EER increases from 0.18% to 0.27%, which is still very low. Above all, the proposed method is able to achieve very low EER in both regular verification and partial spoof attacks.

6 Conclusion

In this paper, aiming to improve the multimodal system's resistance to partial spoof attacks, we proposed the use of collaborative representation fidelity with non-target subjects to measure the affinity of a query sample to a claimed client. We further considered sparse coding as a competing comparison among the claimed client and non-target subjects, and hence explored two sparsity-based measures associated with individual subjects for recognition. The encouraging performance evidently supports the use of sparsity-based one-to-many comparisons in multimodal systems. However, based on their performance in spoof attacks, only the representation-based one is selected as recognition score. Last, two types of representation-based scores for each modality are assembled to train a SVM classifier.

The proposed method was compared with well-known multimodal methods like LLR, SVM, and Sum fusion methods, using the cosine similarity measure, on a chimeric multimodal database of face and ear traits. The experimental results demonstrate that in both regular verification and partial spoof attacks, the proposed method overwhelmingly outperforms its competitors. The proposed method is a general model for combining multiple biometric traits. In the future work, we plan to evaluate more biometric traits like palmprint, iris, and with real spoofed data. We believe the method can be further enhanced by using more robust feature extraction method like CNN-based, and advanced multimodal joint sparse coding techniques [24].

Acknowledgements. This work was partly supported by the National Natural Science Foundation of China (61602390, 61860206007, 61532009, 61571313, 61605054), the EPSRC Programme Grant (FACER2VM) EP/N007743/1, EPSRC/dstl/MURI project EP/R018456/1, Chinese Ministry of Education (Z2015101), Department of Science and Technology of Sichuan Province (2017RZ0009, 2017FZ0029, and 18GJHZ0138), and by funding under 2016CDLZ-G02-SCU from Sichuan University and Lu-Zhou city.

References

1. Ratha, N.K., Connell, J.H., Bolle, R.M.: An analysis of minutiae matching strength. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 223–228. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-45344-X_32
2. Biggio, B., Fumera, G., Marcialis, G.L., Roli, F.: Statistical meta-analysis of presentation attacks for secure multibiometric systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(3), 561–575 (2017)
3. Rodrigues, R.N., Ling, L.L., Govindaraju, V.: Robustness of multimodal biometric fusion methods against spoof attacks. *J. Vis. Lang. Comput.* **20**(3), 169–179 (2009)

4. Wild, P., Radu, P., Chen, L., et al.: Robust multimodal face and fingerprint fusion in the presence of spoofing attacks. *Pattern Recogn.* **50**, 17–25 (2016)
5. Rodrigues, R.N., Kamat, N., Govindaraju, V.: Evaluation of biometric spoofing in a multimodal system. In: *IEEE International Conference on Biometrics: Theory Applications & Systems*, pp. 1–5 (2010)
6. Marasco, E., Johnson, P., Sansone, C., Schuckers, S.: Increase the security of multibiometric systems by incorporating a spoofing detection algorithm in the fusion mechanism. In: Sansone, C., Kittler, J., Roli, F. (eds.) *MCS 2011. LNCS*, vol. 6713, pp. 309–318. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21557-5_33
7. Bhardwaj, I., Londhe, N.D., Kopparapu, S.K.: A spoof resistant multibiometric system based on the physiological and behavioral characteristics of fingerprint. *Pattern Recogn.* **62**, 214–224 (2017)
8. Huang, Z., Liu, Y., Li, C., et al.: A robust face and ear based multimodal biometric system using sparse representation. *Pattern Recogn.* **46**(8), 2156–2168 (2013)
9. Huang, Z., Liu, Y., Li, C., et al.: An adaptive bimodal recognition framework using sparse coding for face and ear. *Pattern Recogn. Lett.* **53**, 69–76 (2015)
10. Song, X., Feng, Z.H., Hu, G., Kittler, J., Wu, X.J.: Dictionary integration using 3D morphable face models for pose-invariant collaborative-representation-based classification. *IEEE Trans. Inf. Forensics Secur.* **13**(11), 2734–2745 (2018)
11. Verlinde, P., Cholet, G.: Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application. In: *AVBPA*, pp. 188–193 (1999)
12. Merati, A., Poh, N., Kittler, J.: User-specific cohort selection and score normalization for biometric systems. *IEEE Trans. Inf. Forensics Secur.* **7**(4), 1270–1277 (2012)
13. Wright, J., Yang, A.Y., Ganesh, A., et al.: Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(2), 210–227 (2009)
14. Cheng, H., Liu, Z., Yang, L., Chen, X.: Sparse representation and learning in visual recognition: Theory and applications. *Signal Process.* **93**(6), 1408–1425 (2013)
15. Shao, C., Song, X., Feng, Z.H., Wu, X.J., Zheng, Y.: Dynamic dictionary optimization for sparse-representation-based face classification using local difference images. *Inf. Sci.* **393**, 1–14 (2017)
16. Kua, J., Ambikairajah, E., Epps, J., Togneri, R.: Speaker verification using sparse representation classification. In: *IEEE ICASSP*, pp. 4548–4551 Prague, Czech Republic, (2011)
17. Li, M., Zhang, X., Yan, Y., Narayanan, S.: Speaker verification using sparse representations on total variability i-vectors. In: *12th Annual Conference of the International Speech Communication Association*, Florence, Italy, pp. 2729–2732 (2011)
18. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: which helps face recognition? In: *ICCV*, Barcelona, Spain, pp. 471–478 (2011)
19. University of Science & Technology Beijing (USTB). <http://www1.ustb.edu.cn/resb/>. Accessed Jan 2016
20. Martinez, A.M., Benavente, R.: The AR Face Database. *CVC Technical Report 24* (1998)
21. Georgia Tech Face Database. http://www.anefian.com/research/face_reco.htm. Accessed June 2016
22. Figueiredo, T., Jain, A.K.: Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(3), 381–396 (2002)
23. Liu, Y., You, Z., Cao, L.: A novel and quick SVM-based multi-class classifier. *Pattern Recogn.* **39**(11), 2258–2264 (2006)
24. Yuan, X.T., Liu, X., Yan, S.: Visual classification with multitask joint sparse representation. *IEEE Trans. Image Process.* **21**(10), 4349–4360 (2012)

Machine Learning



Function-Guided Energy-Precision Optimization with Precision-Rate-Complexity Bivariate Models

Hao Liu¹, Rong Huang^{1(✉)}, and Zhihai He²

¹ College of Information Science and Technology,
Donghua University, Shanghai 201620, China
{liuhao, rong.huang}@dhu.edu.cn

² Department of Electrical and Computer Engineering,
University of Missouri, Columbia, MO 65211, USA

Abstract. In an intelligent wireless vision sensor network, an intra encoder is used for the energy-precision optimization with two control parameters: sampling ratio and quantization parameter, which have a direct impact on the coding bit rate, encoder complexity, wireless transmission energy, as well as the server-end object classification precision. Through extensive experiments, we construct the precision-rate-complexity bivariate models to understand the behaviors of the intra encoder and the deep convolutional neural networks, and then characterize the inherent relationship between bit rate, encoding complexity, classification precision and these two control parameters. With these models, we study the problem of optimization control of the wireless vision sensor node so that the node-end energy can be minimized subject to the server-end object classification precision. Our experimental results demonstrate that the proposed control method is able to effectively adjust the energy consumption of the sensor node while achieving the target classification performance.

Keywords: Intra encoder · Energy-precision optimization · Bivariate models
Deep convolutional neural networks

1 Introduction

In an intelligent wireless vision sensor networks (iWVSN), the vision analysis task is performed on the compressed images. Therefore, the reconstruction quality of the compressed image, as well as the encoder design and configuration, will have direct impact on the subsequent vision analysis performance. The latest standardization efforts in compression coding have led to the specification of high efficiency video coding (HEVC) [1]. Studies have been performed to analyze and model the complexity behavior of the HEVC encoder. In [2], the encoding complexity is incorporated into the rate-distortion analysis to reduce the encoder's energy consumption, where the macroblock-level computational complexity of the H.264 encoder is modeled for each prediction mode. Authors in [3] proposed a rate-power allocation scheme for wireless video chat applications, where the transmission parameters are adaptively adjusted based on a power-rate-distortion model.

Recently, researchers have recognized the importance of joint design of image compression and vision analysis. For traffic surveillance, an unequal error protection scheme was developed in [4] to increase the vehicle tracking accuracy by allocating more resources to the image region of interest. By classifying macroblocks into different groups in video frames, a rate control method was also proposed for preserving the important local image features [5]. For moving object surveillance, a dynamic rate control scheme was developed in [6] to achieve higher image quality for the regions of interest. For lossy image compression of plant phenotyping, a λ -domain HEVC rate-distortion model was implemented to reduce the object segmentation errors at different bit rates [7].

In this work, we choose the deep convolutional neural networks (DCNN) for object classification of target images at the server end. Deep neural networks are able to construct complex representations and automatically learn a compositional relationship between inputs and outputs, mapping input images to output labels [8]. Once a DCNN is trained using the back-propagation learning procedure, the classification or test is a purely feedforward process [9]. During the past several years, a significant amount of works have been done to push the performance limits of DCNN in vision analysis. However, the joint design of image compression, wireless transmission, and DCNN-based object classification has not been studied.

Within the context of iWVSN with DCNN-based target classification, this work has identified two important system control parameters, image sampling ratio (S) and quantization parameter (Q) of the HEVC intra encoder, play a critical role in determining the encoder complexity, coding bit rate, energy consumption in encoding and wireless transmission, reconstructed image quality, and object classification precision. Following an operational approach with extensive experiments, we establish models to characterize the behaviors of coding bit rate, encoding energy, wireless transmission energy, and DCNN classification precision with respect to two control parameters. Based on these models, we then develop optimal resource allocation schemes to minimize the sensor-node energy consumption while achieving the object classification precision.

2 Energy-Precision Control Framework

As discussed in the above, the task objective of the iWVSN is to identify targets. The target images are collected, encoded, transmitted and analyzed for automated classification. As illustrated in Fig. 1, each iWVSN sensor node encodes the target image using the HEVC intra encoder. The compressed bit stream is transmitted over a wireless channel, and then forwarded to the cloud server through Internet. At the server side, the bit stream is decoded to reconstruct the image. The DCNN is then applied to classify this reconstructed image to determine the target class. The iWVSN system is controlled by two important parameters: (1) the sampling ratio S and (2) the quantization parameter Q . Specifically, before encoding, we perform down-sampling on the target image X with a sampling ratio of S . As we know, the sampling ratio S has a direct impact on the following: (1) the encoding complexity which translates into encoder power consumption, (2) the coding bit rate which translates into power consumption in wireless transmission, and (3) the complexity and precision of the DCNN classifier.

The quantization parameter Q has a direct impact on (1) the coding bit rate, (2) the quality of reconstructed images, and (3) the precision of target classification.

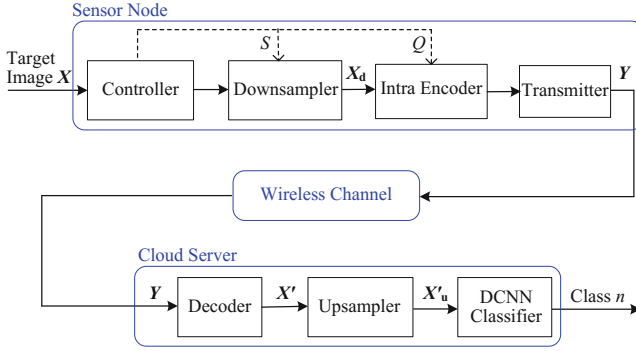


Fig. 1. The module diagram of energy-precision control framework.

HEVC image compression and wireless transmission are two major tasks for each node, consuming most of its energy. With S and Q as the control parameters, $P(S, Q)$ denotes the classification precision in percentage (%), and $R(S, Q)$ denotes the coding bit rate per image in Kbps, and $C(S, Q)$ denotes the average complexity per image in millisecond (ms). The node-end energy consumption includes two additive components: the encoding energy E_c for compressing images, and the transmission energy E_t for sending bit data to a cloud server. The encoding energy E_c is related to the computational complexity $C(S, Q)$ of the encoder, which depends on the two control parameters: S and Q . In other words, we have

$$E_c = \Phi[C(S, Q)] \quad (1)$$

where $\Phi[\cdot]$ is a task-specific mapping the computational complexity or processor cycles into energy consumption. The transmission energy E_t is related to the bit rate $R(S, Q)$ of the compressed image data stream which also depends on (S, Q) . Therefore, we write

$$E_t = \Theta[R(S, Q)] \quad (2)$$

where $\Theta[\cdot]$ is also a task-specific mapping which depends on the wireless transmission scheme. In this work, we consider the concise mapping mechanism for $\Phi[\cdot]$ and $\Theta[\cdot]$. In iWVSN, the node-end processor power is stable and the wireless transmission is delay-tolerant. The encoding energy E_c exhibits a linear relation with the computational complexity $C(S, Q)$, and the wireless transmission energy E_t also exhibits a linear relation with the coding bit rate $R(S, Q)$ [10]. In this way, the total amount of energy consumption by the sensor node is given as follows:

$$E(S, Q) = E_c + E_t = p_c \cdot C(S, Q) + e_t \cdot R(S, Q) \quad (3)$$

where the encoding power p_c is a constant in J/ms, and the wireless transmission power e_t is another constant in J/Kbps. At the server end, the HEVC decoder decodes the received bit stream and reconstructs the image. The reconstructed image is then used as input to the DCNN module for target classification. Note that overall objective of the iWVSN is to determine the target classes. Therefore, we propose to use the classification precision $P(S, Q)$ as the performance metric, which depends on the size and quality of the input image.

One major motivation of this work is from the following observation: the vision sensor nodes may have spent too much computational and energy resources in encoding and transmitting the image samples whose quality is much higher than that needed for accurate target classification. In other words, from the target classification perspective, the sensor nodes may have wasted a lot of energy. This leads to the optimal resource allocation and control problem under DCNN precision constraints:

$$\min E(S, Q) \quad \text{s.t. } P(S, Q) \geq P_{\min} \quad (4)$$

In this work, we aim to minimize the energy consumption of the iWVSN node while achieving the required precision P_{\min} for target classification. To successfully solve the above control problem, we need to establish those precision-rate-complexity models: $P(S, Q)$, $R(S, Q)$ and $C(S, Q)$, which will be presented in the following section.

3 Precision-Rate-Complexity Modeling

Through extensive experiments, we will establish models to characterize the behaviors of rate, complexity, and precision with respect to the two control parameters: S and Q .

3.1 Datasets and Experimental Setup

In this paper, we consider the application scenario of remote wildlife monitoring and protection. A network of vision sensors are deployed to monitor wildlife and human presence in the monitoring region. Triggered by animal motion, the sensor node will capture an image and transmit it to the cloud server for object classification: animal, human, or no-object. For example, if a human is detected in the wildlife protection zone, an alarm will be generated. To test the DCNN classification module, we have assembled a dataset of 1001 images of size 640×480 , with about $1/3$ images for each class. The basic unit of HEVC is a coded tree block (CTB) whose minimum size is 16×16 pixels. Let (W, H) and (W_d, H_d) be the (width, height) of the original image X and its down-sampled image X_d , respectively. With a given sampling ratio S , the (width, height) of the down-sampled image X_d can be denoted as follows:

$$(W_d, H_d) = ([W/\sqrt{S}], [H/\sqrt{S}]) \quad (5)$$

where $[k]$ denotes a multiple of 16 that is closest to k ; the width and height of a down-sampled image uniformly increase or decrease. For each target image, we will use the

HEVC intra encoder to compress the image with different sampling ratios S and quantization parameters Q . The candidate values of S are $\Omega_s = \{1, 2, \dots, 50\}$ and the candidate values of Q are $\Omega_q = \{0, 1, 2, \dots, 51\}$. In total, we have 50×52 different (S, Q) configurations. In this paper, we assume that the compressed bit stream is correctly received at the server side for successful image decoding and reconstruction. The DCNN is then applied to classify the reconstructed image into one of three classes: Human, Animal, and Background. The DCNN model is previously trained with a large set of labeled images, which are uncompressed and have the original resolution of 640×480 .

3.2 Precision-Rate-Complexity Analysis

Note that S and Q are two independently control parameters. We propose to firstly analyze the precision-rate-complexity behaviors with respect to each individual parameter. Once we have understood and established these 1-Dimensional models, we then proceed to establish the joint model with these two control parameters. Figure 2(a) shows the actual $P(S, Q)$ curves at different S and different Q . We can see that for small values of Q , for example, from 0 to 30, the compressed image quality is high, and the precision does not change much. When Q is larger than the threshold (e.g., 30), the precision drop exponentially. This implies that the image quality does not affect the DCNN classification performance if it is above a certain threshold. This example suggests that the sensor node will waste the bits and energy resources if the image quality is already above the threshold since an even higher image quality level does not help the DCNN classification. We can see that the $P(S)$ curves follows a decreasing near-exponential behavior. For actual coding bit rate, Fig. 2(b) plots the actual $R(S, Q)$ curves at different S and different Q , whose average bit rate is 1725 Kbps. These curves show an exponentially decreasing relationship with the increasing S or Q . For a given encoder, its computational complexity is directly related to its encoding time. Figure 2 (c) plots the actual $C(S, Q)$ curves at different S and different Q , whose average complexity is 258 ms. We can see that the quantization parameter Q does not affect the complexity much. Certainly, the complexity will decrease for smaller input images or larger sampling ratios.

3.3 Precision-Rate-Complexity Bivariate Models

A fundamental goal of the precision-rate-complexity modeling is to solve the node-end energy minimization problem under server-end classification precision constraints. By heuristically feeding actual data into the constrained minimization task in (4), the actual distribution of all optimal control parameters can be obtained by exhaustively testing all possible (S, Q) configurations. With all cases, Fig. 3 shows the distribution of actual optimal Q values at different precisions, where a dot denotes an optimal Q value at its precision. It can be seen that all optimal Q values are limited to a range from $Q = 24$ and $Q = 51$. When the smaller Q values vary from 0 to 23, the resulting precision (bit rate, complexity) have no influence on the optimal solution of the energy-precision optimization task, which motivates us neglect some (S, Q) configurations so as to produce more accurate precision-rate-complexity models.

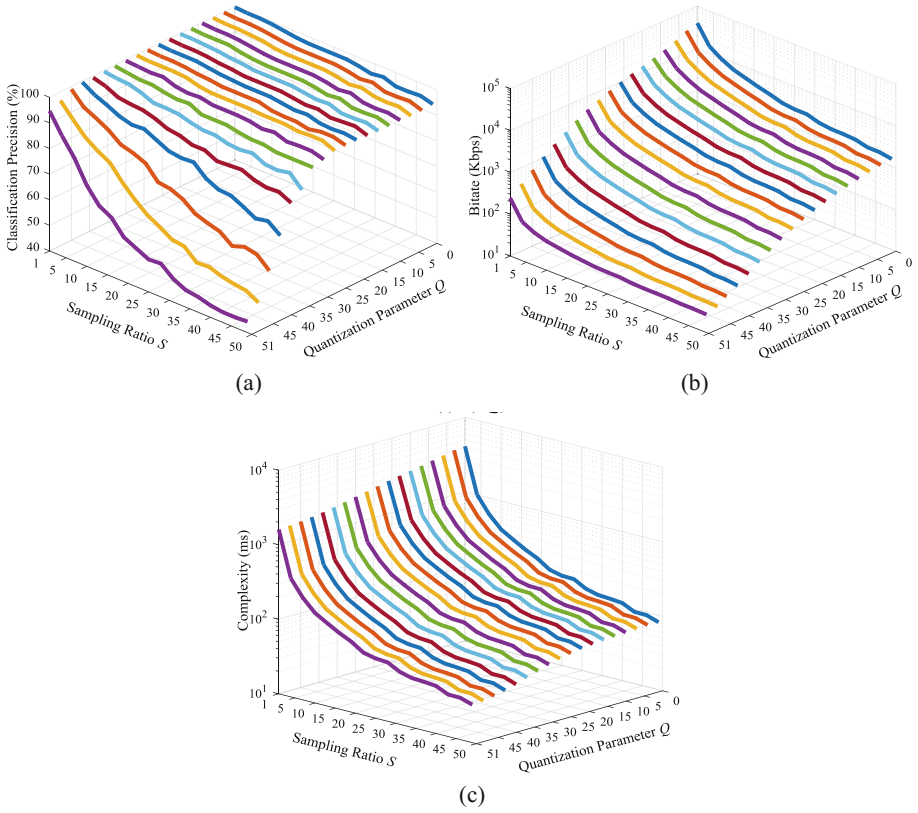


Fig. 2. The actual behaviors of classification precision, coding bit-rate and complexity: (a) $P(S, Q)$ curves, (b) $R(S, Q)$ curves, (c) $C(S, Q)$ curves.

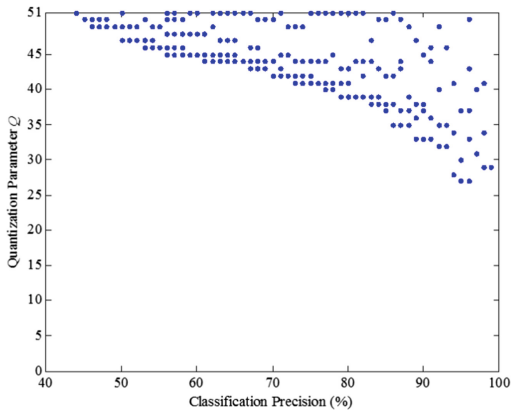


Fig. 3. The distribution of actual optimal Q values.

Based on the experiments, our curve fitting goal may only considers those larger Q values in the range of [24, 51] and all 50 possible values for S . Thus, we have 50×28 possible (S, Q) configurations that needs to be fitted. It can be seen that the curves of actual precision and complexity also exhibits a certain linear behavior, and a first-order polynomial may approximate such a behavior. We relax the maximum value constraint in a smaller fitting space. By comparing various exponential forms and their parameters, the precision-rate-complexity bivariate models can be constructed as follows:

$$P(S, Q) = \beta_{p1} - \beta_{p2} \cdot e^{\beta_{p3} \cdot Q + \beta_{p4} \cdot S} - \beta_{p5} \cdot Q - \beta_{p6} \cdot S \quad (6)$$

$$R(S, Q) = \beta_{r1} \cdot e^{\beta_{r2} \cdot Q + \beta_{r3} \cdot S} + \beta_{r4} \cdot e^{\beta_{r5} \cdot Q + \beta_{r6} \cdot S} \quad (7)$$

$$C(S, Q) = \beta_{c1} \cdot e^{\beta_{c2} \cdot Q + \beta_{c3} \cdot S} + \beta_{c4} \cdot e^{\beta_{c5} \cdot Q + \beta_{c6} \cdot S} + \beta_{c7} \cdot Q + \beta_{c8} \cdot S + \beta_{c9} \quad (8)$$

By continuous approximation, Table 1 reports the optimal parameter values of the precision-rate-complexity bivariate models. With better fitting results, the bivariate models can be used to search the appropriate S and Q for the energy-precision optimization task.

Table 1. The parameters values of precision-rate-complexity bivariate models.

Parameter	Value	Parameter	Value	Parameter	Value
β_{p1}	102.6	β_{r1}	132020	β_{c1}	5850.2
β_{p2}	0.0863	β_{r2}	-0.1059	β_{c2}	-0.01508
β_{p3}	0.1123	β_{r3}	-0.5508	β_{c3}	-0.792
β_{p4}	0.01999	β_{r4}	41189	β_{c4}	1336.5
β_{p5}	-0.04603	β_{r5}	-0.1159	β_{c5}	-0.02882
β_{p6}	0.1242	β_{r6}	-0.04011	β_{c6}	-0.07452
				β_{c7}	-0.5673
				β_{c8}	0.7176
				β_{c9}	23.73

4 Resource Allocation and Energy Minimization

In the above section, we have established models to predict the encoder computational complexity $C(S, Q)$, coding bit rate $R(S, Q)$, and the DCNN precision $P(S, Q)$. Based on these models, we are ready to study the resource allocation problem, answering the following important question: what is the minimum energy consumption that the iWVSN node needs to spend in order to achieve the desired DCNN object classification precision at the server end? As discussed in the above section, the iWVSN resource allocation problem can be formulated by:

$$\min E(S, Q) = p_c \cdot C(S, Q) + e_t \cdot R(S, Q) \quad \text{s.t.} \quad P(S, Q) \geq P_T \quad (9)$$

In the above section, we have obtained analytical models for the encoder complexity $C(S, Q)$, the encoding bit rate $R(S, Q)$, and the DCNN classification precision $P(S, Q)$. We resort to a numerical solution. Specifically, with the precision-rate-complexity bivariate models, we are able to compute the values of $P(S, Q)$, $R(S, Q)$, and $C(S, Q)$ for a dense grid of points (S, Q) . We then find the set of grid points which satisfy the precision constraint. Finally, within this set, we find the optimal (S, Q) which has the minimum energy $E(S, Q)$. Figure 4 shows the optimal sampling ratio S^* and encoder quantization parameter Q^* for a given target classification precision P_T . Each dot represents an optimal look-up-table value of S^* or Q^* for a given target precision P_T . The jig-saw effect is caused by the fact that the quantization parameter Q has to be an integer and the input image size has to be a multiple of 16. For easy implementation in actual system control, we propose to approximate optimal sampling ratio $S^*(P_T)$ using a piece-wise linear function as shown in Fig. 4(a) in solid lines, and approximate the optimal encoder quantization parameter $Q^*(P_T)$ using an exponential function as shown in Fig. 4(b):

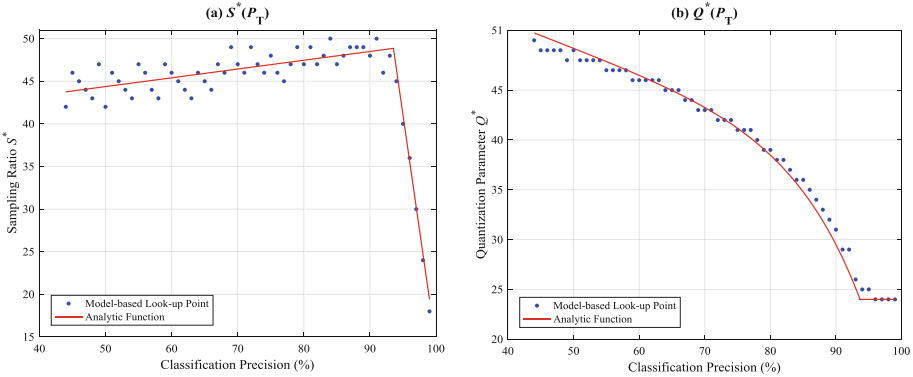


Fig. 4. The look-up-table solution and analytic solution for energy-precision optimization: (a) the $S^*(P_T)$ function; (b) the $Q^*(P_T)$ function.

$$S^*(P_T) = \begin{cases} \text{Round}(\omega_1 \cdot P_T + \omega_2), & P_T < P_0 \\ \text{Round}(\omega_3 \cdot P_T + \omega_4), & P_T \geq P_0 \end{cases} \quad (10)$$

$$Q^*(P_T) = \text{Round}(\tau_1 \cdot e^{\tau_2 \cdot P_T} + \tau_3 \cdot e^{\tau_4 \cdot P_T}) \quad (11)$$

where the values of $S^*(P_T)$ belong to $\{1, 2, \dots, 49, 50\}$, and the values of $Q^*(P_T)$ belong to $\{24, 25, \dots, 50, 51\}$. The model parameters are listed in Table 2.

Table 2. The coefficients of analytic functions.

Coefficient	Value	Coefficient	Value
ω_1	0.103	τ_1	-0.002
ω_2	39.22	τ_2	0.096
ω_3	-5.4	τ_3	62.01
ω_4	553.3	τ_4	-0.0046
P_0	93.6		

5 Experimental Results

In this section, we conduct experiments to evaluate the proposed method. Our test dataset consists of 1001 uncompressed camera-trap images. The original input image size is 640×480 in RGB color format. The DCNN classifier is constructed and trained by using CAFFE which has 5 convolutional layers followed by 3 fully connected layers [11]. The DCNN classifier has been well trained and tested on original target images, where the target images are categorized into three object classes, namely: *Human*, *Animal*, and *Background*. The image sampling ratio S and quantization parameter Q jointly affect the complexity, bit rate, and object classification precision. The candidate values of S are set to be $\{1, 2, 3, \dots, 49, 50\}$, and the candidate values of Q are set to be $\{0, 1, 2, \dots, 50, 51\}$. For image compression, we adopt the HM-16.7 main profile HEVC intra coding [12]. During simulation, to translate the computational complexity into computational energy, we set the thermal design power (TDP) of the microprocessor to be $p_c = 0.14$ J/ms. The transmission power e_t is set to 2.6×10^{-3} Kbps [10].

Figures 5, 6 and 7 show the estimation results by the precision-rate-complexity bivariate models obtained from the above section. Specifically, Fig. 5(a) shows the estimation results for the $P(Q)$ curve at different S . Figure 5(b) shows the estimation results for the $P(S)$ curve for different Q . We can see that the model is able to accurately capture the behavior of actual classification precision. For the estimation performance, we have R - square = 0.9548, and RMSE = 3.057%. Figure 6(a) shows the estimation results for the $R(Q)$ curve at different S . Figure 6(b) shows the estimation results for the $R(S)$ curve for different Q . We can see that this rate model is very accurate with R - square = 0.991. Figure 7(a) shows the estimation results for the $C(Q)$ curves at different S . Figure 7(b) shows the estimation results for the $C(S)$ distributions for different Q . We can see that the complexity model is very accurate with R - square = 0.997.

Figure 8(a) shows the minimum energy consumption (in lines with circles) of the iWVSN node to achieve the target DCNN classification precision at the server end using the precision-rate-complexity bivariate model and resource allocation. For comparison, we also include the actual optimal value of minimum energy consumption (in lines with crosses) which are obtained from brute-force search based on experiments with all possible combinations of control parameters (S, Q) . We can see that our analysis and optimization approaches the actual optimal values. Figure 8(b) shows the operating bit rate and complexity of the iWVSN node. We can see that, if we allow a very small percentage of performance drop, for example, from dropping the precision from 97% to 96%, we can save the total energy at the iWVSN node by up to 2 times, which is very significant.

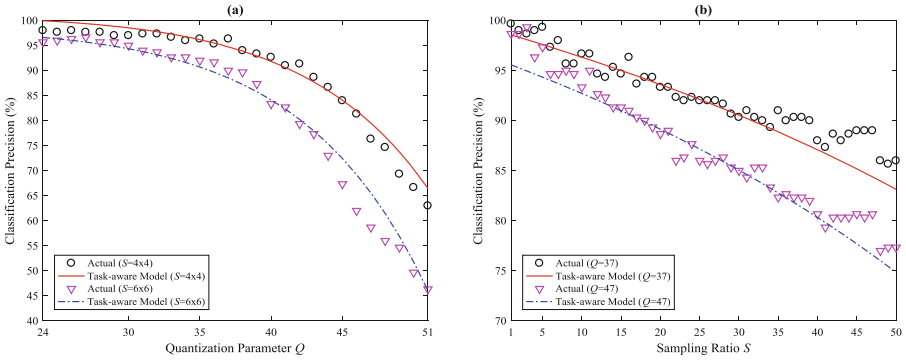


Fig. 5. The fitting results for precision model: (a) $P(Q)$ at different S ; (b) $P(S)$ at different Q .

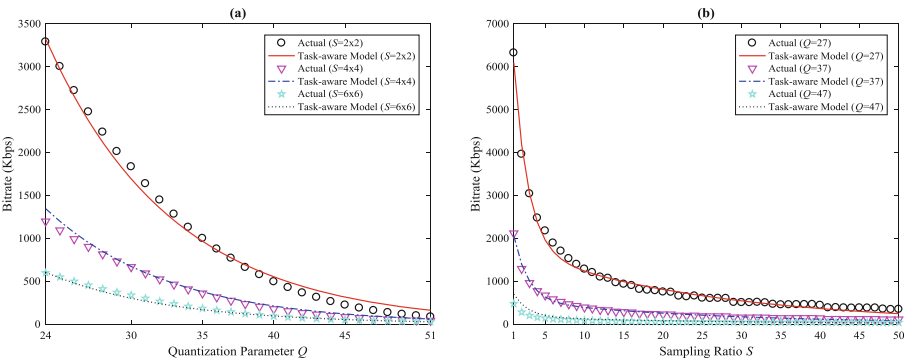


Fig. 6. The fitting results for rate model: (a) $R(Q)$ at different S ; (b) $R(S)$ at different Q .

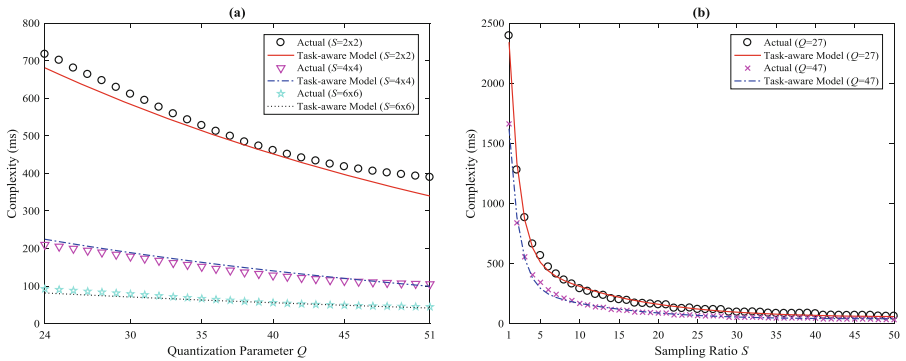


Fig. 7. The fitting results for complexity model: (a) $C(Q)$ at different S ; (b) $C(S)$ at different Q .

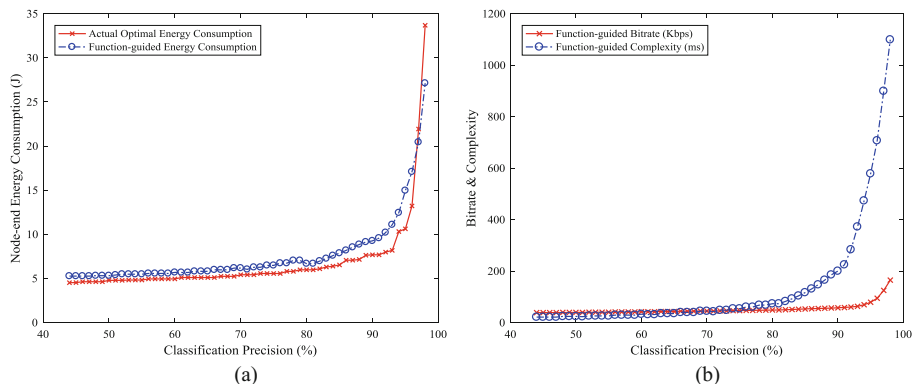


Fig. 8. The actual optimal energy consumption vs. function-guided energy consumption.

6 Conclusion

In this paper, we have studied the resource modeling, allocation, and optimization problem for an intelligent wireless vision sensor network which collects image samples of the targets, encodes and transmits the data to a cloud server for object classification using DCNN. We developed a new framework for energy-precision analysis and optimization. Specifically, we use the HEVC intra encoder for image compression configured with two control parameters: the image sampling ratio and quantization parameter. Through extensive experiments, we construct the precision-rate-complexity bivariate models to understand the behaviors of the HEVC intra encoder and the DCNN, and characterize the inherent relationship between bit rate, encoding complexity, classification precision and these two control parameters. Based on these models, we study the problem of optimization control of the wireless vision sensor node so that the node-end energy can be minimized subject to the server-end object classification precision. Our experimental results demonstrate that the proposed control method is able to effectively adjust the energy consumption of the sensor node while achieving the target classification performance.

Acknowledgments. This work is supported by the Natural Science Foundation of Shanghai (18ZR1400300).

References

1. Pastuszak, G., Abramowski, A.: Algorithm and architecture design of the H.265/HEVC intra encoder. *IEEE Trans. Circuits Syst. Video Technol.* **26**(1), 210–222 (2016)
2. Li, X., Wien, M., Ohm, J.R.: Rate-complexity-distortion optimization for hybrid video coding. *IEEE Trans. Circuits Syst. Video Technol.* **21**(7), 957–970 (2011)
3. Chuah, S.P., Tan, Y.P., Chen, Z.: Rate and power allocation for joint coding and transmission in wireless video chat applications. *IEEE Trans. Multimed.* **17**(5), 687–699 (2015)

4. Chen, Z., Tsaftaris, S.A., Soyak, E., Katsaggelos, A.K.: Application-aware approach to compression and transmission of H.264 encoded video for automated and centralized transportation surveillance. *IEEE Trans. Intell. Transp. Syst.* **14**(4), 2002–2007 (2013)
5. Chao, J., Huitl, R., Steinbach, E., Schroeder, D.: A novel rate control framework for SIFT/SURF feature preservation in H.264/AVC video compression. *IEEE Trans. Circuits Syst. Video Technol.* **25**(6), 958–972 (2015)
6. Ko, J.H., Mudassar, B.A., Mukhopadhyay, S.: An energy-efficient wireless video sensor node for moving object surveillance. *IEEE Trans. Multi-Scale Comput. Syst.* **1**(1), 7–18 (2015)
7. Minervini, M., Tsaftaris, S.A.: Classification-aware distortion metric for HEVC intra coding. In: *Proceedings of IEEE Visual Communications and Image Processing*, Singapore, pp. 1–4 (2015)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, pp. 1097–1105 (2012)
9. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* (2014)
10. Redondi, A., Baroffio, L., Bianchi, L., Cesana, M., Tagliasacchi, M.: Compress-then-analyze versus analyze-then-compress: what is best in visual sensor networks? *IEEE Trans. Mob. Comput.* **15**(12), 3000–3013 (2016)
11. <http://caffe.berkeleyvision.org/installation.html>
12. HEVC Software Repository — HM-16.7 Reference Model. https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.7



Point Cloud Noise and Outlier Removal with Locally Adaptive Scale

Zhenxing Mi^{1,2} and Wenbing Tao^{1,2}(✉)

¹ Shenzhen Huazhong University of Science and Technology Research Institute, Shenzhen 518057, China

² National Key Laboratory of Science and Technology on Multi-spectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China

{m201772503,wenbingtao}@hust.edu.cn

Abstract. This paper introduced a simple and effective algorithm to remove the noise and outliers in point sets generated by multi-view stereo methods. Our main idea is to discard the points that are geometrically or photometrically inconsistent with its neighbors in 3D space using the input images and corresponding depth maps. We attach a scale value to each point reflecting the influence to the adjacent area of the point and define a geometric consistency function and a photometric consistency function for the point. We employ a very efficient method to find the neighbors of a point using projection. The consistency functions are related to the normal and scale of the neighbors of points. Our algorithm is locally adaptive, feature preserving and easy to implement for massive parallelism. It performs robustly with a variety of noise and outliers in our experiments.

Keywords: Multi-view stereo · Noise filtering · Scale · Local adaptive

1 Introduction

The state of the art in multi-view stereo methods has seen great development in robustness and accuracy these years. However, point sets produced by multi-view stereo methods are usually redundant and inevitably with a lot of noise and outliers due to imperfection of acquisition hardware and algorithms, as is shown in Fig. 1(b). Modern MVS algorithms use different output scene representations, such as depth maps, a point cloud, or a mesh. Depth map scene representation is one of the most popular choices due to the flexibility and scalability [7] but suffers more noise. This poses a great challenge to surface reconstruction.

We can impose strong regularization in MVS methods to reduce outliers, but this will destroy sharp features and may be time consuming. Some denoising methods directly operate on unorganized point cloud and using k nearest neighbors to optimize the position and normal of a reference point [13]. Depth map, however, often provides us with additional information such as connectivity and

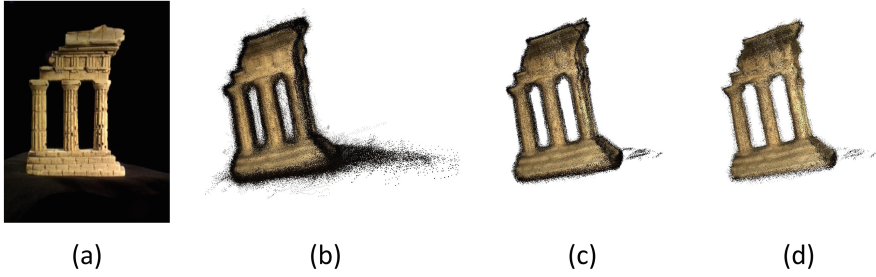


Fig. 1. We use the multi-view stereo methods MVE [8] to reconstruct a dense 3D point cloud (b) for the Middlebury Temple dataset [11] (a). The output point cloud is very noisy. We denoise the depth maps only use geometric consistency (c). A lot of noise and outliers are removed but there are still some black points from the background retained on the border of the temple. We use geometric consistency and photometric consistency together in (d) and get better result.

scale [3]. Therefore, in our method, we computed a scale value for each point using the input depth maps in image space. The scale value provides valuable information about the surface area each point was acquired from, as discussed by Fuhrmann et al. [3]. With scale information, we can handle datasets containing non-uniform noise and sample resolution.

In our method, we do not discretize the 3D space, avoiding large memory and time usage. We project a reference point to other depth maps and find its neighbors in the image space. The neighbors obtained from image space are not necessarily but most likely to be neighbors in the 3D space. Then we project them back to the 3D space to evaluate the geometric and photometric consistency between the reference point and its neighbors. Our locally adaptive geometric consistency function and photometric consistency are related to the scale of the reference point and its neighbors. The functions are defined compactly supported, namely, the neighbors used for evaluating the functions must be near the reference point in spatial space. Because of the redundancy of the depth maps, we do not change the position, normal and color of the points but just remove the points that are not consistent with its neighbors. For the sake of efficiency, we employ view selection strategy to identify nearby views using the feature points reconstructed in the previous SFM phase [6, 8]. This enables our methods the ability to operate on extremely large photo collections.

Our contributions are:

- An approach using *scale* information to evaluate the geometric and photometric consistency, which is local adaptive feature preserving and more accurate.
- Finding neighbors of reference points in image space by depth map triangulation and projection, which is very efficiency.

In the remainder of this paper, we first review related work (Sect. 2). Then introduce our denoise approach (Sect. 3), perform experiments on a variety of data sets (Sect. 4) and conclude our work (Sect. 5).

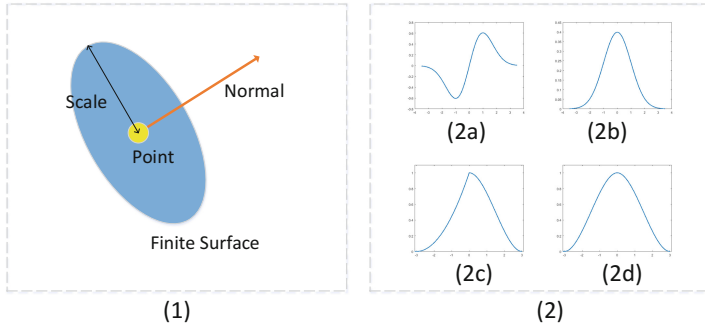


Fig. 2. (1) A point with a scale value represents a finite surface in the spatial space. (2) The shape of the functions $f_x(x)$ (2a), $f_y(y)$ and $f_z(z)$ (2b), $w_x(x)$ (2c) and $w_{yz}(r)$ (2d).

2 Related Work

Here we describe some closely related work in point set denoising, focusing on how they handle point sets generated by images with varying resolution and viewing parameters, what parameters they use and to what extent they are time and memory consuming.

Most multi-view stereo methods integrate a depth map fusion strategy into the depth estimation stage or after the whole reconstruction. They usually enforce visibility and consistency across views. Wu et al. [18] firstly use an indicator function based on visibility cues in [16] to remove outliers. Then they enforce visibility consistency across views. Such method is not sophisticated thus there remains a lot of noise and outliers. Schönberger et al. [10] define a directed graph of consistent pixels with their photometric and geometric consistency support set, then find and fuse the clusters of consistent pixels in this graph. The fused point cloud are of high quality and have little outliers. However, finding clusters is very time consuming and not easy to parallelize. In addition, they use the photometric and geometric consistency terms computed in the MVS procedure of their reconstruction method, which are only available in their approach.

The above methods proposed as part of multi-view stereo methods usually use parameters that are unique in their depth reconstruction and thus their use is restricted. There are also some methods independent of the MVS. Sun et al. [13] directly denoise point clouds using the L_0 norm to preserving sharp features. Wolff et al. [17] take depth maps as input and implicitly uses a surface represented by the input depth maps to check geometric consistency and photometric consistency between each per-view point and other input views. Our method are relevant to their method, projecting the points to the image space of other depth maps. However, we take a completely different, local adaptive strategy to examine consistency using the finite surface represented by points with scale value.

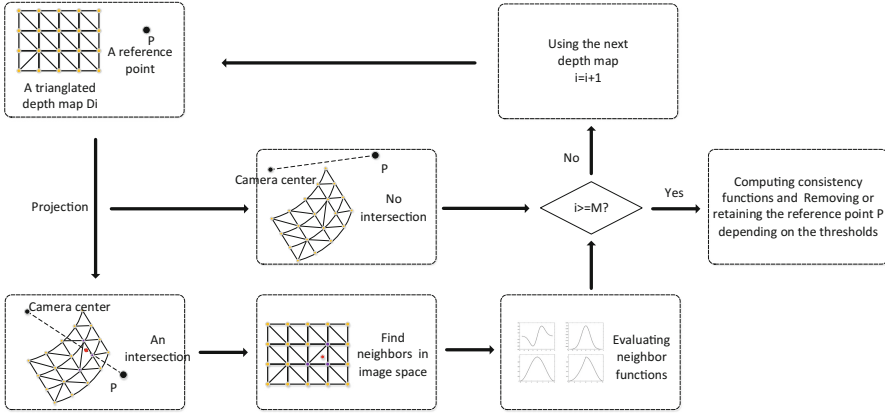


Fig. 3. Our point denoising pipeline: we examine a reference point \mathbf{p} against other depth maps. A depth map D_i is triangulated in the image space. Then we project the reference point to the depth map and get which triangle it falls into. If no such triangle exist, we do not compute any function and examine \mathbf{p} against next depth map. If it falls into an triangle, we regard the three vertexes as the neighbor of \mathbf{p} and use them to evaluate our functions. Our functions are related to the scale of the points. After examining the reference point against all the depth maps, we compare the functions with threshold and decide if the point will be removed.

The quality of the reconstructed surface strongly depends on the quality of the input point set which is inevitably with noise and outliers. Therefore, many surface reconstruction methods explicitly use some strategy to handle the noise and outliers. Poisson surface reconstruction [9] estimate local sampling density and scale the contribution of each point accordingly. However, sampling density is not necessarily related to the sample resolution, and an increased sampling density may simply be caused by data redundancy as discussed in [4]. Fuhrmann et al. [3] construct a discrete, multi-scale signed distance field capable of representing surfaces at multiple levels of detail and produce output surfaces that are adaptive to the scale of the input data. Our methods apply the same depth map triangulation step and compute the scale of every points. Fuhrmann et al. [4] attach the scale value to each sample point and use the weighted average of locally estimated functions to define the implicit surface compactly around the input data. The method is virtually parameter-free for mixed-scale datasets and does not require any global operations. Our method draws inspiration from this method and uses scale value computed from the triangulated depth maps to handle the noise outliers.

3 Denoising and Outlier Removal

In this section, we describe the evaluation of geometric and photometric consistency between a reference point \mathbf{p} and its neighbors in spatial space. We assume

that M input depth maps are given and points in them are equipped with a position, a normal and a color.

3.1 Definition of *Scale*

We define a scale value for each point related to the depth map it comes from. As illustrated in Fig. 3, we first find the adjacent points for a point in the input depth map in image space, and then computed a scale value for each point by averaging the spatial distances between the point and its adjacent points. As discussed by Fuhrmann et al. [3], the scale value provides valuable information about the surface area each point was acquired from. The points in depth maps are not ideal points. Instead, they represent a surface at a particular scale depending on viewing distance, focal length and image resolution [3] as illustrated in Fig. 2. With scale information, we can define local adaptive functions for geometric consistency and photometric consistency to handle datasets containing non-uniform noise and sample resolution.

3.2 Neighbors in Image Space and LCS

To determine the geometric and photometric consistency, every reference point \mathbf{p} has to be examined against its neighbors in the spatial space. Depth maps can provide us with additional information such as connectivity. As illustrated in Fig. 3, We triangulate the depth maps in image space using the method proposed by [3]. Then we project the reference point \mathbf{p} to other depth maps and get the triangles it falls into. The three vertices of the triangle are regarded as the neighbors of the reference point. After the whole projection, we get a set of neighbors $N_{\mathbf{p}} = \{\mathbf{p}_i | i = 1, \dots, M\}$ for \mathbf{p} . Each of them are equipped with a position $\mathbf{p}_i \in \mathbb{R}^3$, a normal $\mathbf{n}_i \in \mathbb{R}^3$, $\|\mathbf{n}_i\| = 1$, and a scale value $s_i \in \mathbb{R}$. Generally, such neighbors are most likely near the reference point in spatial space. Since our functions are compactly supported, we can ensure that the neighbor points used to evaluate geometric and photometric consistency are actually near the reference point. When examining \mathbf{p} against \mathbf{p}_i , we use the local coordinate of \mathbf{p} in the local coordinate system (LCS) of \mathbf{p}_i . The local coordinate is $\mathbf{x}_i = R_i \cdot (\mathbf{p} - \mathbf{p}_i)$ with a rotation matrix $R_i = R(\mathbf{n}_i)$ such that \mathbf{p}_i is located in the origin and the normal \mathbf{n}_i coincides with the positive x-axis [4]. The LCS is only up to the position and normal of \mathbf{p}_i so the functions should be invariant to the choice of the LCS orthogonal to the normal.

3.3 Geometric Consistency

Given a reference point \mathbf{p} , and a set of neighbors $N_{\mathbf{p}} = \{\mathbf{p}_i | i = 1, \dots, M\}$, we define a signed geometric consistency function $F(\mathbf{p})$ as a weighted sum of basis functions, as proposed in the surface reconstruction method [4]:

$$F(\mathbf{p}) = \frac{\sum_i w d_i(\mathbf{x}_i) w n_i(\mathbf{p}_i) f_i(\mathbf{x}_i)}{\sum_i w d_i(\mathbf{x}_i) w n_i(\mathbf{p}_i)}$$

$$W(\mathbf{p}) = \sum_i wd_i(\mathbf{x}_i)wn_i(\mathbf{p}_i) \tag{1}$$

where \mathbf{x}_i is the local coordinate of \mathbf{p} in local coordinate system of (LCS) \mathbf{p}_i . The basis function $f_i(\mathbf{x}_i)$ is a signed function which is positive in front of the surface and negative otherwise (similar to a signed distance function). The function $f_i(\mathbf{x}_i)$ and weight $wd_i(\mathbf{x}_i)$, $wn_i(\mathbf{p}_i)$ are parameterized by the i th neighbor’s position \mathbf{p}_i , normal \mathbf{n}_i and scale s_i . Similar to [4], for each neighbor \mathbf{p}_i , we define a basis function that is unit-integral and stretched depending on the scale of the neighbor.

With $\mathbf{x}_i = (x, y, z)$, we use a function $f_x(x)$ that is like the derivative of the Gaussian in the x-coordinate. The standard deviation of $f_x(x)$ is set to the scale of the neighbor, that is $\sigma = s_i$. It is positive when $x > 0$ and negative when $x < 0$. Normalized Gaussians $f_y(y)$, $f_z(z)$ are used orthogonal to the normal in y-coordinate and z-coordinate.

$$f_x(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, f_y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}}, f_z(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}} \tag{2}$$

We define the basis function of the i th neighbor as:

$$f_i(\mathbf{x}_i) = f_x(x)f_y(y)f_z(z) = \frac{x}{\sigma^4 2\pi} \cdot e^{-\frac{1}{2\sigma^2}(x^2+y^2+z^2)} \tag{3}$$

The function meets the condition that it must be unit-integral as discussed before:

$$\int \int \int |f_i(\mathbf{x}_i)| d\mathbf{x}_i = \int |f_x(x)| dx \int f_y(y) dy \int f_z(z) dz = 1 \tag{4}$$

In the following, we define a weighting function $wd_i(\mathbf{x}_i)$ related to the distance between the neighbor \mathbf{p}_i . It is designed to ensure that the neighbor used to evaluate $F(\mathbf{p})$ are actually near the reference point \mathbf{p} . As illustrated in the Fig. 2, $f_i(\mathbf{x}_i)$ is almost zero beyond 3σ , and thus $wd_i(\mathbf{x}_i)$ is define as 0 beyond 3σ to ensure the compact support. As discussed by Curless and Levoy [1] and Vruble et al. [14]: if a point has been observed, the existence of a surface between the observer and the point is not possible. Therefore, if $x < 0$, the existence of a reference point behind the neighbor cause conflict. Therefore, we want to reduce the weight quickly. The weighting function $wd_i(\mathbf{x}_i)$ is non-symmetric in x-direction and rotation invariant in y- and z-direction:

$$wd_i(\mathbf{x}_i) = w_x(x) \cdot w_{(yz)}(\sqrt{y^2 + z^2}) \tag{5}$$

$$w_x(x) = \begin{cases} \frac{1}{9} \frac{x^2}{\sigma^2} + \frac{2}{3} \frac{x}{\sigma} + 1 & x \in [-3\sigma, 0) \\ \frac{2}{27} \frac{x^3}{\sigma^3} - \frac{1}{3} \frac{x^2}{\sigma^2} + 1 & x \in (0, 3\sigma] \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$w_{yz}(r) = \begin{cases} \frac{2}{27} \frac{r^3}{\sigma^3} - \frac{1}{3} \frac{r^2}{\sigma^2} + 1 & r < 3\sigma \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$r = \sqrt{y^2 + z^2} \quad (8)$$

Additionally, to better preserve the sharp features in the point set and avoid over smoothing, we define a weighting function $wn_i(\mathbf{p}_i)$ related to the similarity between the normals of the points.

$$wn_i(\mathbf{p}_i) = \begin{cases} \frac{\mathbf{n}_p^T \mathbf{n}_i}{\|\mathbf{n}_p\| \cdot \|\mathbf{n}_i\|} & \mathbf{n}_p^T \mathbf{n}_i > 0 \\ 0 & \mathbf{n}_p^T \mathbf{n}_i \leq 0 \end{cases} \quad (9)$$

We define $wn_i(\mathbf{p}_i)$ as 0 if $\mathbf{n}_p^T \mathbf{n}_i \leq 0$ to eliminate the influence of neighbors that have a much different normal direction with the reference point, which can improve the robustness.

Since $F(\mathbf{p})$ is compactly supported, some extremely isolated outliers with little neighbors will have small $F(\mathbf{p})$. They cannot be filtered if we only make use of $F(\mathbf{p})$. We observe that if a reference point is an outlier with little neighbors, its $W(\mathbf{p})$, the sum of the weighting function, will be very small. In practice, points with a weight below a certain value are also removed, which can filter out extremely isolated outliers.

3.4 Photometric Consistency

In practice, our algorithm can filter out common noise and outliers with geometric consistency function. However, as illustrated by Fig. 1(b) (c), the noisy points near the border of object are hard to remove. Our observation is that such points usually have a blurred color that is quite different from its neighbors. So we define a function $E(\mathbf{p})$ to evaluate the photometric consistency between the reference point \mathbf{p} , with a color $\mathbf{c}(\mathbf{p})$, and its neighbors $N_p = \{\mathbf{p}_i | i = 1, \dots, M\}$, whose colors are $\mathbf{c}(\mathbf{p}_i)$. $E(\mathbf{p})$ is defined as

$$E(\mathbf{p}) = \frac{\|\mathbf{c}(\mathbf{p}) - \mathbf{c}'(\mathbf{p})\|}{\|\mathbf{c}(\mathbf{p})\|} \quad (10)$$

where $\mathbf{c}'(\mathbf{p})$ is the temporary color of \mathbf{p} computed by the color of its neighbors. Inspired by the anisotropic and feature-preserving nature of bilateral filtering [2], we compute $\mathbf{c}'(\mathbf{p})$ as

$$\mathbf{c}'(\mathbf{p}) = K(\mathbf{p}) \sum_i W_c(\mathbf{p}_i) W_s(\mathbf{p}_i) \mathbf{c}(\mathbf{p}_i) \quad (11)$$

where $W_c(\mathbf{p}_i)$ is the spatial weighting term, $W_s(\mathbf{p}_i)$ is the signal weighting term and $K(\mathbf{p}) = \frac{1}{\sum_i W_c(\mathbf{p}_i) W_s(\mathbf{p}_i)}$ is the normalization factor. $W_c(\mathbf{p}_i)$ is a spatial Gaussian that decreases the influence of distant neighbors:

$$W_c(\mathbf{p}_i) = \exp(-\|\mathbf{p} - \mathbf{p}_i\|^2 / 2\sigma^2) \quad (12)$$

where $\sigma = s_p$, which is the scale value of the reference point \mathbf{p} . We do not define $W_s(\mathbf{p}_i)$ as Gaussian but just use the normalized dot product of the normals between \mathbf{p} and \mathbf{p}_i for efficiency.

$$W_s(\mathbf{p}_i) = \begin{cases} \frac{\mathbf{n}_p^T \mathbf{n}_i}{\|\mathbf{n}_p\| \cdot \|\mathbf{n}_i\|} \mathbf{n}_p^T \mathbf{n}_i > 0 \\ 0 & \mathbf{n}_p^T \mathbf{n}_i \leq 0 \end{cases} \quad (13)$$

The influence of neighbors that have a much different normal direction with the reference point, i.e. $\mathbf{n}_p^T \mathbf{n}_i \leq 0$, are eliminated.

3.5 Depth Map Selection for Scalability

Our algorithm proposed above does not perform costly optimizations and thus is very efficient and easy to parallel. However, assuming we have N input depth maps with a resolution of K , the time complexity of our algorithm is $O(KN^2)$. It increases quadratically with the number of depth maps N . In practice, we do not consider depth maps whose viewing direction \mathbf{v}_i differs too much from the viewing direction \mathbf{v} under which \mathbf{p} was observed, i.e. $\mathbf{v}_i^T \mathbf{v} < 0$. However, the time complexity still increase quickly when operating extremely large data sets. In order to make our algorithm more scalable, we introduce a view selection method as an option when operating on large data sets. We use SFM points to select nearby depth maps for a reference depth map. The number of shared SFM points between the reference depth map and other depth maps is a good indicator whether the reference point is visible in other depth maps. We calculate the number of shared feature points, sort them from large to small and only examine the points in the reference depth map against the first C depth maps. Now the time complexity is $O(KCN)$, increasing linearly with the number of depth maps N . Since the reference point is not likely visible by the depth maps with few shared SFM points, our algorithm still yields good results with view selection in our experiments.

3.6 Point Filtering Strategy

After evaluating $F(\mathbf{p})$, $W(\mathbf{p})$ and $E(\mathbf{p})$ for a reference point \mathbf{p} , we use them to decide whether the point \mathbf{p} will be retained. We *retain* a point if it satisfies all of the following three conditions:

$$-T_p < F(\mathbf{p}) < T_p, \quad W(\mathbf{p}) > \alpha, \quad E(\mathbf{p}) < \varepsilon \quad (14)$$

Since $F(\mathbf{p})$ is an locally adaptive function, we define a locally adaptive threshold $T_p = \beta F(x = s_p, \sqrt{y^2 + z^2} = s_p, \sigma = s_p)$ for $F(\mathbf{p})$. Actually, $F(x = s_p, \sqrt{y^2 + z^2} = s_p, \sigma = s_p)$ is the function value of a virtual point whose local coordinates are relate to the scale of reference point. This definition can ensure the adaptivity of filtering. β is a constant decided by users to control the degree of filtering. It performs well in feature preserving in our experiments. The threshold of $W(\mathbf{p})$ is a constant α to filter out the extremely isolated outliers. It is related to the number of input depth maps and typically we set it to 25 when there are hundreds of input depth maps. The threshold of $E(\mathbf{p})$ is a constant ε . We typically set it to 0.1, that is, if the difference between the real color and the temporary color is above 10%, we filter the point out. It performs well in eliminating the color blur in the point sets.

4 Results

In this section, we perform evaluation of our algorithm on different types of datasets. In Sect. 4.1 we compare our filtering results with the method proposed by Wolff et al. [17] on several datasets released by Yücer et al. [15]. We use (Screened) Poisson Surface Reconstruction (PSR) [9] for surface reconstruction. In Sect. 4.2 we analyze the performance of our strategy for filtering using the Fountain data set of Strecha et al. [12]. In Sect. 4.3 we check the validity of the photometric consistency function on the Temple Full dataset from the Middlebury benchmark [11].

4.1 Comparison Against the Method of Wolff et al.

Figure 4 shows the results of comparison of our method and the method proposed by Wolff et al. [17] on the datasets released by Yücer et al. [15]. Wolff et al. [17] also takes depth maps as input and use these datasets for the evaluation of their method. We use two of state-of-the-art multi-view stereo methods, the colmap of Schönberger et al. [10] and the MVE of Fuhrmann et al. [5] for the dense multi-view depth reconstruction. While Fuhrmann et al. (MVE) [5] do not integrate a fusion step into the MVS reconstruction, colmap of Schönberger et al. [10] fuse their resulting depth maps into a point cloud. In our experiment, we disable the fusion step in colmap [10] and use its raw depth maps for filtering. We also show the result of the fusion result of colmap [10] for comparison.

We use about 200 input images for the reconstructions of each dataset. For MVE we used the level-2 depth maps (4*downsampling) the same as the experiments of Wolff et al. [17]. We also limit the max image size in colmap to the same resolution as the experiment of MVE for comparison. We run PSR for each point cloud in our experiment after the filtering. As shown in Fig. 4, the outliers of the results of MVE and colmap are very dense so that it is not easy to filter them out. However, our method employ both the $F(\mathbf{p})$ and $W(\mathbf{p})$ in Geometric consistency and thus more robust to such outliers. Comparing to the results of Wolff et al. [17], we get more clean and dense point cloud and little outliers with our method. In all the experiments, the run time of our method and Wolff et al. are almost the same. With the use of scale value, our method are not only perform well in removing outliers but also preserve more sharp features in the point cloud. Since the method of Wolff et al. are actually global, the results of it often retains some outliers while destroying the sharp features.

4.2 Analysis of Filtering Strategy

In this section, we analyze the filtering strategy of our methods using the datasets released by Yücer et al. [15] and the Fountain data set of Strecha et al. [12]. In our experiments, we use the locally adaptive threshold for $T_{\mathbf{p}}$. As is shown in Fig. 4, the result of locally adaptive threshold is more clean nearby the surface of the objects. That is, $F(\mathbf{p})$ with a locally adaptive threshold performs better in feature preserving with the scale information. We also use different constant

	MVE			Colmap			
	Unfiltered	Wolff et al.	Ours	Unfiltered	Wolff et al.	Ours	Fusion
Points							
Mesh							
Points							
Mesh							
Points							
Mesh							
Points							
Mesh							

Fig. 4. We use the MVE [5] and colmap [10] to generate the depth maps. After filtering, we use (PSR) [9] to reconstruct a surface for the point cloud. We compare our output point clouds and surfaces with those of Wolff et al. [17]. We also show the result of the fusion method of colmap as a comparison.

threshold of α for $W(\mathbf{p})$. As illustrated by Fig. 5, as the increase of α , the number of outliers in the point cloud decreases quickly because $W(\mathbf{p})$ play an important role in extreme outliers removing.

4.3 Performance of Photometric Consistency

Figure 1 shows the importance of photometric consistency function. The Temple Full dataset from the Middlebury benchmark [11] contains 312 images. Their background are black, so as shown in Fig. 1, the resulting point cloud using MVE contains a mass of black points near the border of the object. These black points are retained when we only apply the photometric consistency. When we integrate the photometric consistency in filtering, most of the black points are removed and the colors of the surface of the object are more uniform.

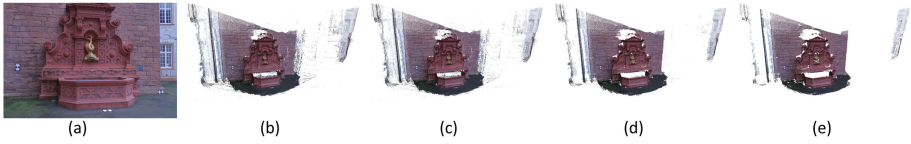


Fig. 5. The sum of weight, $W(\mathbf{p})$ performs an important role in outliers removing. The α for $W(\mathbf{p})$ in (b), (c), (d), (e) are 0, 2, 4, 6. It is clear that as the increase of α , the number of outliers decreases quickly.

5 Conclusions

We propose a very efficient point cloud denoiser which is locally adaptive. We are mainly inspired by the surface reconstruction method [4]. Since scale and efficiency are common topics in 3D reconstruction, we hope that other people can be inspired by our work and solve some other problems.

Acknowledgment. We would like to thank the reviewers for their time and the valuable comments. This work is supported by the National Natural Science Foundation of China (Grant 61772213) and in part by Grants JCYJ20170818165917438 and 2017010201010121.

References

1. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 303–312. ACM (1996)
2. Fleishman, S., Drori, I., Cohen-Or, D.: Bilateral mesh denoising. In: ACM SIGGRAPH, pp. 950–953 (2003)
3. Fuhrmann, S., Goesele, M.: Fusion of depth maps with multiple scales. In: SIGGRAPH Asia Conference, p. 148 (2011)
4. Fuhrmann, S., Goesele, M.: Floating scale surface reconstruction. *ACM Trans. Graph.* **33**(4), 1–11 (2014)
5. Fuhrmann, S., Langguth, F., Goesele, M.: MVE-A multi-view reconstruction environment. In: GCH, pp. 11–18 (2014)
6. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1434–1441. IEEE (2010)
7. Furukawa, Y., Hernández, C., et al.: Multi-view stereo: a tutorial. *Found. Trends® Comput. Graph. Vis.* **9**(1–2), 1–148 (2015)
8. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE (2007)
9. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. *ACM Trans. Graph.* **32**(3), 29 (2013)
10. Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 501–518. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_31

11. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 519–528 (2006)
12. Strecha, C., Hansen, W.V., Gool, L.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)
13. Sun, Y., Schaefer, S., Wang, W.: Denoising point sets via l0 minimization. *Comput. Aided Geom. Des.* **35**, 2–15 (2015)
14. Vrabel, A., Bellon, O.R., Silva, L.: A 3D reconstruction pipeline for digital preservation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2687–2694. IEEE (2009)
15. Yücer, K., Sorkine-Hornung, A., Wang, O., Sorkine-Hornung, O.: Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction. *ACM Trans. Graph.* **35**(3), 22 (2016)
16. Wei, J., Resch, B., Lensch, H.P.: Multi-view depth map estimation with cross-view consistency. In: BMVC (2014)
17. Wolff, K., et al.: Point cloud noise and outlier removal for image-based 3D reconstruction. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 118–127. IEEE (2016)
18. Wu, P., Liu, Y., Ye, M., Li, J., Du, S.: Fast and adaptive 3D reconstruction with extensively high completeness. *IEEE Trans. Multimed.* **19**(2), 266–278 (2017)



Robust Multi-view Subspace Learning Through Structured Low-Rank Matrix Recovery

Jiamiao Xu², Xinge You^{1,2(✉)}, Qi Zheng², Fangzhao Wang², and Peng Zhang²

¹ Research Institute of Huazhong University of Science and Technology in Shenzhen, Shenzhen, China

² Huazhong University of Science and Technology, Wuhan 430074, China
{jiamiao.xu,youxg}@hust.edu.cn

Abstract. Multi-view data exists widely in our daily life. A popular approach to deal with multi-view data is the multi-view subspace learning (MvSL), which projects multi-view data into a common latent subspace to learn more powerful representation. Low-rank representation (LRR) in recent years has been adopted to design MvSL methods. Despite promising results obtained on real applications, existing methods are incapable of handling the scenario when large view divergence exists among multi-view data. To tackle this problem, we propose a novel framework based on structured low-rank matrix recovery. Specifically, we get rid of the framework of graph embedding and introduce class-label matrix to flexibly design a supervised low-rank model, which successfully learns a discriminative common subspace and discovers the invariant features shared by multi-view data. Experiments conducted on CMU PIE show that the proposed method achieves the state-of-the-art performance. Performance comparison under different random noise disturbance is also given to illustrate the robustness of our model.

Keywords: Subspace learning · Multi-view learning
Low-rank representation

1 Introduction

In our daily life, people or objects can be captured at different viewpoints or by different sensors. Consequently, one object has multiple representations, this is also known as multi-view data. Multi-view data is generally heterogeneous [4, 13] (i.e., intra-class samples from another views may have lower similarity than inter-class samples from the same view), which brings a large challenge to recognition or classification tasks. For this reason, numerous work focusing on multi-view subspace learning (MvSL) appears.

Early work on MvSL aims to learn multiple mapping functions, one for each view, to respectively project multi-view data into a common latent subspace,

in which the view divergence can be decreased and the similarity of heterogeneous samples can be measured. Among these approaches, the most well-known unsupervised method is Canonical Correlation Analysis (CCA) [8]. However, CCA can only be applied to two-view scenarios. Multi-view Canonical Correlation Analysis (MCCA) [20] was later proposed to generalize CCA to multi-view situations. Moreover, some state-of-the-art methods (e.g., Generalized Multi-view Analysis (GMA) [21], Multi-view discriminant analysis (MvDA) [10] and Multi-view Hybrid Embedding (MvHE) [26]) also have been proposed. Different from MCCA, these methods take into consideration discriminant information, thus improving the representation power of subspace. Despite significant results obtained by them, they fail to work during the testing phase, when the view-related information of test samples is not provided [5].

Low-rank multi-view subspace learning (LRMSL) circumvents this drawback by learning a common mapping function for all views, with the help of low-rank representation (LRR). Compared with aforementioned methods, this type of approaches do not need view-related information in testing process. Based on how the prior knowledge (i.e., view-related information and class-label information) is involved in the training phase, LRMSL approaches can be divided into three categories: unsupervised methods, weakly-supervised methods and supervised methods. Unsupervised methods (e.g., Latent Low-rank Representation (LatLRR) [17]) make no use of these two kinds of information, weakly-supervised methods (e.g., Low-rank Common Subspace (LRCS) [4]) only take into consideration view-related information, whereas supervised methods take full advantage of class-label information (e.g., Supervised Regularization based Robust Subspace (SRRS) [12] and Robust Multi-view Subspace Learning (RMSL) [5]).

LRMSL approaches did make a great progress for multi-view data, but there still exist some problems. The success of low-rank representation bases on the assumption that samples from a same class have higher similarity, but the assumption is invalid for multi-view data. Hence, unsupervised and weakly-supervised methods are incapable of effectively discovering the invariant features shared by multi-view data. Although supervised methods provide a feasible solution, existing methods (e.g., SRRS and RMSL) do not achieve significant improvement. One possible reason is that some graph embedding (e.g., Locally Linear Embedding (LLE) [19] and Locality Preserving Projections (LPP) [7]) can not be applied to multi-view data. This is because these methods require manifolds are locally linear. Unfortunately, this condition is also not met for multi-view data [22, 25].

To overcome the problems discussed above, we get rid of the framework of graph embedding and introduce class-label matrix to flexibly design a supervised low-rank model. In the process, a discriminative subspace and the shared information of multi-view data are discovered. Experimental results on face recognition demonstrate the superiority of our method.

The remainder of this paper is organized as follows. Section 2 introduces related work and Sect. 3 presents the proposed method. Optimization is given

in Sect. 4. Experimental results are provided in Sect. 5. Finally, Sect. 6 concludes this paper.

2 Related Work

In this section, related work is presented to make interested readers more familiar with the low-rank multi-view subspace learning (LRMSL).

Low-rank Representation (LRR) is a popular approach that has been widely applied in many computer vision and machine learning tasks. In [3], Robust Principle Component Analysis (Robust PCA) was proposed to recover a low-rank component and a sparse component from given data, which assumes that data is homogeneous. To handle data sampled from multiple spaces, Liu *et al.* [15,16] proposed LRR methods which learn a lowest-rank representation at a given dictionary. Besides discovering the global class structure, it also eliminates the influence of noises. Similar to dictionary learning approaches [1,18], the dictionary used in LRR is also expected to be overcomplete. However, this condition is not always easily met. Thus, LatLRR [17] was proposed to construct the dictionary with both observed data and hidden data. In the area of LRR, methods all aim to find an optimal (i.e. structured) representation matrix \mathbf{Z} with respect to data \mathbf{X} [15,16]. Specifically, assume that we have a dataset $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_c]$ and a dictionary \mathbf{A} , then the optimal representation \mathbf{Z} is expected to be block-diagonal as follows:

$$\mathbf{Z}^* = \begin{pmatrix} \mathbf{Z}_1^* & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2^* & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_c^* \end{pmatrix}, \quad (1)$$

where c is number of classes.

Low-rank Multi-view Subspace Learning (LRMSL) uses low-rank representation technology to learn a robust subspace, in which the intrinsic structure of data is preserved. In [4], LRCS was proposed to capture the shared structure from multiple views. SRRS [12] used fisher criterion to learn a discriminant subspace. Considering there are two kinds of structure embedded in multi-view data (i.e. class structure and view structure), Ding *et al.* [5] proposed RMSL to learn two kinds of low-rank structure simultaneously.

3 Robust Low-Rank Multi-view Subspace Learning

3.1 Problem Formulation

Suppose we have a multi-view dataset $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n]$, where n is the number of views. $\mathbf{X}_k = [\mathbf{X}_{k_1}, \mathbf{X}_{k_2}, \dots, \mathbf{X}_{k_c}]$ denotes the k -th view data, where c is the number of classes and \mathbf{X}_{k_i} represent all samples of the i -th class under the k -th view. Low-rank multi-view subspace learning (LRMSL) aims to find

a component mapping function $\mathbf{P} \in \mathbb{R}^{d \times p}$ to project multi-view data from d -dimensional space into a p -dimensional subspace ($p \leq d$), in which projected samples $\mathbf{P}^T \mathbf{X}$ can be represented as a linear combination of the bases of dictionary \mathbf{A} , and the representation matrix exhibits low-rank characteristic. Its objective can be formulated as:

$$\begin{aligned} \min_{\mathbf{Z}, \mathbf{E}, \mathbf{P}} \quad & \|\mathbf{Z}\|_* + \lambda_1 \|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{P}^T \mathbf{X} = \mathbf{AZ} + \mathbf{E}, \quad \mathbf{P}^T \mathbf{P} = \mathbf{I}, \end{aligned} \tag{2}$$

where \mathbf{E} in Eq. (2) is introduced to remove random noise, the orthogonal constraint on \mathbf{P} is used to obtain an orthogonal subspace and $\lambda_1 > 0$ can be determined by cross validation.

Equation (2) is a basic framework of LRMSL algorithms. To learn a discriminant subspace, we develop a novel supervised model below.

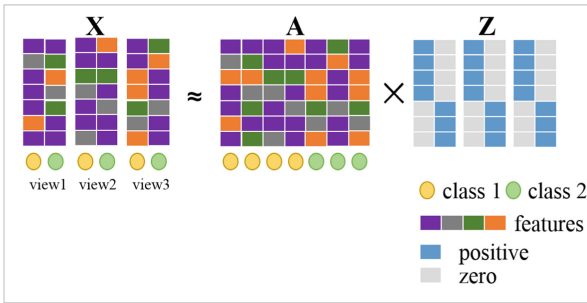


Fig. 1. Illustration of structured low-rank matrix recovery for multi-view data.

3.2 Structured Low-Rank Matrix Recovery

Suppose $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_c]$ denotes the dictionary, where \mathbf{A}_i are the bases of the i -th class. According to the discussion in Sect. 2, structured low-rank matrix \mathbf{Z} of multi-view projected samples $\mathbf{P}^T \mathbf{X}$ can be defined as follows:

$$\mathbf{Z}^* \triangleq (\mathbf{Z}_1^*, \mathbf{Z}_2^*, \dots, \mathbf{Z}_n^*), \tag{3}$$

where \mathbf{Z}_k^* is the structured representation matrix of $\mathbf{P}^T \mathbf{X}_k$, which can be represented as

$$\mathbf{Z}_k^* = \begin{pmatrix} \mathbf{Z}_{k_1}^* & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{k_2}^* & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{k_c}^* \end{pmatrix}. \tag{4}$$

Obviously, low-rank matrix \mathbf{Z} is a structured matrix when each sample from the i -th class can be represented as a linear combination of the dictionary bases from the i -th class. The illustration of the structured low-rank matrix recovery for multi-view data is shown in Fig. 1. As can be seen, intra-class representations are united and inter-class representations are deviated from each other.

To this end, we use class-label matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m]$ to design a supervised model, where m is the number of samples. Assume that $\mathbf{y}_k \in \mathbb{R}^{c \times 1}$ is from the j -th class, it can be defined as

$$\mathbf{y}_k = \begin{bmatrix} \underbrace{0, \dots, 0}_{j-1}, 1, \underbrace{0, \dots, 0}_{C-j} \end{bmatrix}^T. \quad (5)$$

The objective of the proposed supervised algorithm can be formulated as

$$\begin{aligned} & \min_{\mathbf{Z}, \mathbf{E}, \mathbf{P}} \|\mathbf{Z}\|_* + \lambda_1 \|\mathbf{E}\|_1 \\ & \text{s.t. } \mathbf{P}^T \mathbf{X} = \mathbf{A} \mathbf{Z} + \mathbf{E}, \quad \mathbf{P}^T \mathbf{P} = \mathbf{I}, \quad \mathbf{Y} \mathbf{Z} = \mathbf{Y}_s, \quad \mathbf{e}^T \mathbf{Z} = \mathbf{e}^T, \quad \mathbf{Z} \geq 0, \end{aligned} \quad (6)$$

where $\mathbf{Y} \in \mathbb{R}^{c \times m_1}$ and $\mathbf{Y}_s \in \mathbb{R}^{c \times m_2}$ are the class-label matrices of the dictionary \mathbf{A} and the dataset \mathbf{X} respectively, and \mathbf{e} is a column vector with all elements equal to one. $\mathbf{e}^T \mathbf{Z} = \mathbf{e}^T$ in Eq. (6) is used to normalize the representation coefficients (i.e., the sum of each column in \mathbf{Z} is equal to one), $\mathbf{Z} \geq 0$ is used to guarantee that each element in \mathbf{Z} is non-negative. Based on the normalization and non-negative constraints, $\mathbf{Y} \mathbf{Z} = \mathbf{Y}_s$ can guarantee that the \mathbf{Z} we learned is a structured matrix.

The dictionary \mathbf{A} is generally represented by training samples in previous algorithms, thus we replace \mathbf{A} with $\mathbf{P}^T \mathbf{X}$ and we have $\mathbf{Y} = \mathbf{Y}_s$. Moreover, to improve the generalization performance, we introduce an error term \mathbf{E}_L . Then, the objective function (6) can be reformulated as:

$$\begin{aligned} & \min_{\mathbf{Z}, \mathbf{E}, \mathbf{E}_L, \mathbf{P}} \|\mathbf{Z}\|_* + \lambda_1 \|\mathbf{E}\|_1 + \lambda_2 \|\mathbf{E}_L\|_F^2 \\ & \text{s.t. } \mathbf{P}^T \mathbf{X} = \mathbf{P}^T \mathbf{X} \mathbf{Z} + \mathbf{E}, \quad \mathbf{P}^T \mathbf{P} = \mathbf{I}, \quad \mathbf{Y}_s \mathbf{Z} = \mathbf{Y}_s + \mathbf{E}_L, \quad \mathbf{e}^T \mathbf{Z} = \mathbf{e}^T, \quad \mathbf{Z} \geq 0, \end{aligned} \quad (7)$$

where λ_2 controls the contribution of \mathbf{E}_L .

4 Optimization

Through introducing relax variable \mathbf{J} , problem (7) can be translated into

$$\begin{aligned} & \min_{\mathbf{J}, \mathbf{Z}, \mathbf{E}, \mathbf{E}_L, \mathbf{P}} \|\mathbf{J}\|_* + \lambda_1 \|\mathbf{E}\|_1 + \lambda_2 \|\mathbf{E}_L\|_F^2 \\ & \text{s.t. } \mathbf{P}^T \mathbf{X} = \mathbf{P}^T \mathbf{X} \mathbf{Z} + \mathbf{E}, \quad \mathbf{P}^T \mathbf{P} = \mathbf{I}, \quad \mathbf{J} = \mathbf{Z} \\ & \quad \mathbf{Y}_s \mathbf{Z} = \mathbf{Y}_s + \mathbf{E}_L, \quad \mathbf{e}^T \mathbf{Z} = \mathbf{e}^T, \quad \mathbf{Z} \geq 0, \end{aligned} \quad (8)$$

where the augmented Lagrangian function is formulated as

$$\begin{aligned} & \|J\|_* + \lambda_1 \|E\|_1 + \lambda_2 \|E_L\|_F^2 + tr \left(Y_1^T (P^T X - P^T X Z - E) \right) + tr \left(Y_2^T (Z - J) \right) \\ & + tr \left(Y_3^T (Y_s Z - Y_s - E_L) \right) + tr \left(Y_4^T (e^T Z - e^T) \right) \\ & + \frac{\mu}{2} \left(\|P^T X - P^T X Z - E\|_F^2 + \|Z - J\|_F^2 \right) + \frac{\mu}{2} \left(\|Y_s Z - Y_s - E_L\|_F^2 + \|e^T Z - e^T\|_F^2 \right), \end{aligned} \tag{9}$$

where Y_1, Y_2, Y_3 and Y_4 are Lagrange multipliers and μ is a positive penalty parameter. There are five parameters in problem (9) to be optimized, and it is difficult to optimize them simultaneously. For this reason, we employ the alternating direction method of multipliers (ADMMs) [6] to alternately optimize J, Z, E, E_L and P one by one through fixing the other variables. For example, during the $t+1$ iteration of optimization, when we optimize J , variables Z, E, E_L and P are regarded as constants, i.e. inherit results of the t th iteration. In detail, we define $J_t, Z_t, E_t, E_{L,t}, P_t, Y_{1,t}, Y_{2,t}, Y_{3,t}$ and $Y_{4,t}$ as variables in the t th iteration, and then we optimize variables in the $t + 1$ iteration as follows.

Updating J :

$$J_{t+1} = \arg \min_J \frac{1}{\mu_t} \|J\|_* + \frac{1}{2} \left\| J - \left(Z_t + \frac{Y_{2,t}}{\mu_t} \right) \right\|_F^2. \tag{10}$$

Updating E :

$$E_{t+1} = \arg \min_E \frac{\lambda_1}{\mu_t} \|E\|_1 + \frac{1}{2} \left\| E - \left(P_t^T X - P_t^T X Z_t + \frac{Y_{1,t}}{\mu_t} \right) \right\|_F^2, \tag{11}$$

The two problems above can be optimized by the iterative thresholding approach [14].

Updating E_L :

$$E_{L,t+1} = (2\lambda_2 + \mu_t)^{-1} (Y_{3,t} + \mu_t Y_s Z_t - \mu_t Y_s). \tag{12}$$

Updating P :

$$P_{t+1} = \left((X - X Z_t) (X - X Z_t)^T \right)^{-1} ((X - X Z_t) (E_t^T - Y_{1,t}^T / \mu_t)). \tag{13}$$

Updating Z :

$$Z = Z_1^{-1} Z_2, \tag{14}$$

where Z_1 and Z_2 are represented as follows:

$$\begin{aligned} Z_1 &= X^T P_t P_t^T X + I + Y_s^T Y_s + e e^T, \\ Z_2 &= X^T P_t (P_t^T X - E_t) + J_t + Y_s^T (Y_s + E_{L,t}) + e e^T \\ &\quad + (X^T P_t Y_{1,t} - Y_{2,t} - Y_s^T Y_{3,t} - e Y_{4,t}) / \mu_t. \end{aligned}$$

Algorithm 1. Solving Problem (7) by ADMM

Input: X, Y_s, λ_1 and λ_2 ;

Initialization: $J = Z = E = E_L = P = 0$,

$$Y_1 = Y_2 = Y_3 = Y_4 = 0,$$

$$\mu_{max} = 10^6, \quad \mu = 10^{-3},$$

$$\rho = 1.03, \quad \epsilon = 10^{-6};$$

While not converged **do**

 Step 1. Update J by solving problem (10);

 Step 2. Update E by solving problem (11);

 Step 3. Update E_L by (12);

 Step 4. Update P by (13), and then $P \leftarrow \text{orthogonal}(P)$ [9];

 Step 5. Update Z by (14), and then $Z = \max(0, Z)$ [27];

 Step 6. Update multipliers and parameter μ by (15);

Step 7. Check the convergence conditions:

$$\|P^T X - P^T X Z - E\|_\infty < \epsilon,$$

$$\|Z - J\|_\infty < \epsilon,$$

$$\|Y_s Z - Y_s - E_L\|_\infty < \epsilon,$$

$$\|e^T Z - e^T\|_\infty < \epsilon;$$

End while
Output: P, Z, E, E_L .

Afterwards, we update multipliers $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ and \mathbf{Y}_4 in the following way

$$\begin{aligned} \mathbf{Y}_{1,t+1} &= \mathbf{Y}_1, t + \mu_t (\mathbf{P}_{t+1}^T \mathbf{X} - \mathbf{P}_{t+1}^T \mathbf{X} \mathbf{Z}_{t+1} - \mathbf{E}_{t+1}), \\ \mathbf{Y}_{2,t+1} &= \mathbf{Y}_2, t + \mu_t (\mathbf{Z}_{t+1} - \mathbf{J}_{t+1}), \\ \mathbf{Y}_{3,t+1} &= \mathbf{Y}_3, t + \mu_t (\mathbf{Y}_s \mathbf{Z}_{t+1} - \mathbf{Y}_s - \mathbf{E}_{L,t+1}), \\ \mathbf{Y}_{4,t+1} &= \mathbf{Y}_4, t + \mu_t (e^T \mathbf{Z}_{t+1} - e^T), \\ \mu_{t+1} &= \min(\rho \mu_t, \mu_{max}), \end{aligned} \tag{15}$$

where $\rho > 1$ and μ_{max} is a constant. We iteratively update variables and the penalty parameter until the algorithm satisfies the convergence conditions or reaches the maximum iterations. The detailed iteration process is summarized in Algorithm 1.

5 Experiments

In this section, we first specify the evaluation protocol of MvSL algorithms. Following this, one public dataset is introduced and experimental setting is presented. In order to evaluate the performance of the proposed method, three baselines (i.e., PCA [24], LDA [2], LPP [7]) and three state-of-the-art low-rank multi-view subspace learning (LRMSL) algorithms (i.e., LRCS [4], SRRS [12] and RMSL [5]) are selected for comparison.

5.1 Evaluation Protocol

Evaluation protocol of single-view subspace learning (SvSL) methods can not precisely evaluate the performance of multi-view learning algorithms. To this end, similar to [11], we adopt a more convincing evaluation protocol as follows:

$$acc_{v_1}^{v_2} = \frac{\sum(x : x \in X_{probe}^{v_2} \wedge \bar{y} = y)}{\sum(x : x \in X_{probe}^{v_2})}, \quad mACC = \left(\sum_{v_1=1}^n \sum_{v_2=1}^n acc_{v_1}^{v_2} \right) / n^2, \quad (16)$$

where n is the number of views, $acc_{v_1}^{v_2}$ denotes the accuracy when gallery and probe sets are from view v_1 and view v_2 respectively. y and \bar{y} are the true label and the predicted label of data x respectively. In experiments, we average results of all pairwise views as the mean accuracy (mACC).

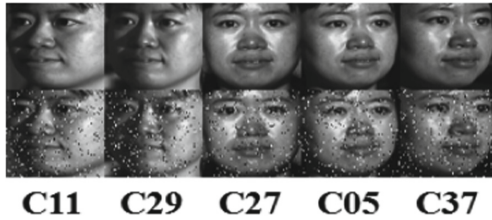


Fig. 2. Exemplar subjects from the CMU PIE dataset. C11, C29, C27, C05 and C37 poses are selected to construct multi-view data. The top row shows clean images and the bottom row shows images with 10% random noise.

5.2 Dataset and Experimental Setting

The CMU Pose, Illumination, and Expression (PIE) Database. (CMU PIE) [23] contains 41,368 images of 68 people with 13 different poses, 43 diverse illumination conditions and 4 various expressions. Five poses (i.e., C11, C29, C27, C05 and C37) are selected to construct multi-view data (see Fig. 2 for exemplar subjects). In experiments, each person at a given pose has 4 images, and images are cropped and resized to 64×64 . To make results more convincing, experiments on CMU PIE are repeated ten times by randomly dividing data into training set, validation set and test set, and we report average result as the final accuracy. Hyper-parameters of all approaches are determined by validation set.

5.3 The Superiority of the Proposed Method

The CMU PIE is used to evaluate face recognition across poses. Similar to [4, 5], experiments are conducted in 5 cases, namely case 1: {C27, C29}, case 2: {C27, C11}, case 3: {C05, C27, C29}, case 4: {C37, C27, C11} and case 5:

Table 1. The average recognition accuracy (%) in 5 cases of CMU PIE in terms of mean accuracy (mACC). **Bold** denotes the best performance.

	Case 1	Case 2	Case 3	Case 4	Case 5
PCA [24]	76.6 ± 7.3	66.0 ± 5.7	62.1 ± 5.7	55.1 ± 4.8	55.5 ± 4.1
LDA [2]	64.5 ± 6.6	58.6 ± 4.1	41.0 ± 10.0	45.0 ± 1.8	46.7 ± 3.6
LPP [7]	72.8 ± 7.5	65.5 ± 4.8	62.3 ± 4.6	53.5 ± 3.7	54.0 ± 4.1
LRCS [4]	74.0 ± 6.7	66.1 ± 4.4	68.9 ± 5.5	56.3 ± 3.1	58.1 ± 3.8
SRRS [12]	74.3 ± 6.5	66.8 ± 4.3	69.0 ± 5.2	56.2 ± 3.4	59.4 ± 3.3
RMSL [5]	75.7 ± 7.7	68.0 ± 4.7	70.7 ± 4.4	57.9 ± 3.2	62.0 ± 3.0
Proposed	83.0 ± 5.9	76.7 ± 7.0	78.5 ± 5.7	67.0 ± 4.5	70.9 ± 5.0

{C37, C05, C27, C29, C11}. In our experiments, 40 people are used as training set, 14 people serve as validation set and the rest comprise the test set.

In the first experiment, we evaluate our performance with three baselines and three state-of-the-art methods. The experimental results are summarized in Table 1. As can be seen, SvSL based methods rank the lowest due to the neglect of the view divergence. Benefited from the consideration of discriminant information, SRRS and RMSL perform better than LRCS. As expected, our method achieves a remarkable improvement compared with RMSL, which we argue can be attributed to the more effectively exploiting discriminant information.

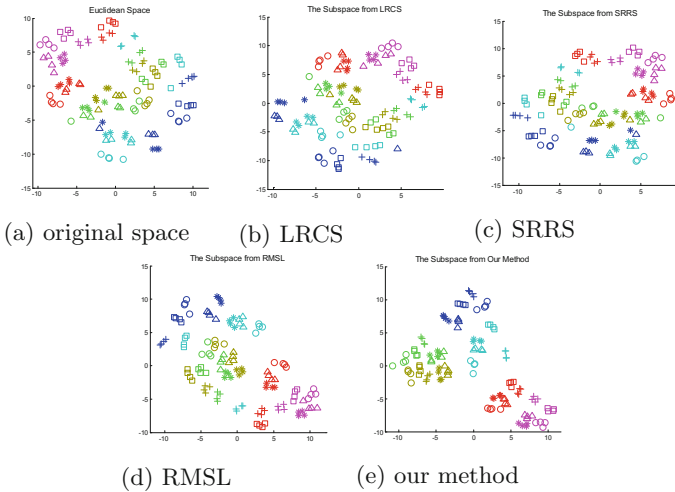


Fig. 3. Illustration of 2D embedding of Euclidean space, the subspace generated by LRCS, SRRS, RMSL and the proposed method in case 5 of CMU PIE dataset. Different colors denote different classes, and different views are denoted by different markers.

Table 2. The average recognition results (%) in case 5 of CMU PIE. **Bold** denotes the best performance.

Gallery	Probe	PCA [24]	LDA [2]	LPP [7]	LRCS [4]	SRRS [12]	RMSL [5]	Proposed
C37	C37	98.8	99.8	99.1	95.2	98.4	100.0	99.8
	C05	70.9	53.6	67.3	71.6	69.3	69.5	81.4
	C27	38.8	23.4	35.5	29.1	30.5	37.7	52.9
	C29	27.1	18.0	26.3	29.8	28.6	30.4	45.0
	C11	25.2	21.3	23.4	28.0	29.8	32.5	46.3
C05	C37	73.0	57.3	70.5	80.5	83.2	84.1	89.3
	C05	96.1	100.0	95.9	85.7	87.9	99.8	99.6
	C27	62.0	40.7	60.5	60.0	63.0	67.3	75.9
	C29	32.9	26.4	29.3	47.7	47.3	49.6	57.9
	C11	31.6	28.2	28.4	43.2	45.9	45.7	58.2
C27	C37	44.3	28.9	41.4	49.3	50.2	48.0	65.5
	C05	58.8	42.0	56.3	70.4	72.3	74.3	78.6
	C27	84.1	99.1	90.0	81.4	85.0	89.1	92.5
	C29	54.1	33.9	51.3	62.3	64.1	66.6	70.0
	C11	48.0	28.9	43.8	49.1	50.1	52.1	65.2
C29	C37	27.9	22.3	25.4	37.5	38.2	37.3	50.9
	C05	34.5	24.6	30.2	42.5	43.8	46.4	59.5
	C27	46.6	33.9	52.1	45.5	49.1	48.8	67.0
	C29	97.0	99.8	97.9	89.6	91.6	98.9	99.5
	C11	75.4	60.5	75.7	83.6	85.9	85.5	91.3
C11	C37	25.0	19.3	17.3	33.9	34.8	37.0	45.7
	C05	29.3	25.4	20.7	29.6	27.7	31.8	45.5
	C27	38.0	21.1	38.8	34.1	32.5	38.0	49.1
	C29	73.6	58.6	76.8	82.1	84.3	82.9	88.9
	C11	94.6	99.8	96.3	89.3	91.8	96.6	97.1
Average		55.5	46.7	54.0	58.1	59.4	62.0	70.9
Standard deviation		4.1	3.6	4.1	3.8	3.3	3.0	5.0

Table 3. The average recognition accuracy (%) in case 5 of CMU PIE with random noise in terms of mean accuracy (mACC). **Bold** denotes the best performance. the values in parentheses denote the relative performance loss (%) with respect to the random noise scenario. “NR” denotes noise ratio.

NR	LRCS [4]	SRRS [12]	RMSL [5]	Proposed
0%	58.1 (0.0)	59.4 (0.0)	62.0 (0.0)	70.9 (0.0)
5%	56.4 (2.9)	57.2 (3.7)	60.0 (3.2)	70.0 (1.3)
10%	56.5 (2.8)	56.6 (4.7)	55.6 (10.3)	70.0 (1.3)
15%	54.9(5.5)	56.8 (4.4)	56.2 (9.4)	68.2 (3.8)
20%	53.4(8.1)	55.6 (6.4)	48.8 (21.3)	66.7 (5.9)

To better evaluate performance of the proposed method, detailed results in case 5 of CMU PIE are shown in Fig. 3 and Table 2. As can be seen in Fig. 3, all

low-rank subspace learning approaches can remove the view divergence to some extent. However, LRCS, SRRS and RMSL approaches fail to distinguish the yellow class from the green one correctly, whereas these two classes are separated obviously in the subspace generated by our method. As a whole, the embeddings shown in Fig. 3 corroborate the results summarized in Table 1. Moreover, as can be seen in Table 2, one should note that our method does not achieve the best performance when the gallery and the probe data come from the same view. The reason for this phenomenon is that the constraint with respect to intra-view and intra-class samples is only based on low-rank representation. Compared with traditional graph embedded, this is a weak constraint.

At last, we evaluate the robustness of the proposed methods. we add random noise to original images by randomly replacing 5%, 10%, 15% and 20% pixels (see Fig. 2 for exemplar subjects) and report the results in case 5 in Table 3. As can be seen, LRCS, SRRS and RMSL are more sensitive to random noise than our method. Take the 20% random noise scenario as an example, our method only suffers from a relative 5.9% performance drop from its original 70.9% accuracy, whereas the accuracy of RMSL decreases to 48.8% with a relative performance drop nearly 21.3%.

6 Conclusion

In this paper, we proposed an novel framework based on structured low-rank matrix recovery to learn a discriminant subspace for multi-view data. Experiments conducted on CMU PIE show that the proposed method successfully discovers the discriminant information shared by multi-view data, thus improving the performance of subsequent recognition or classification tasks. Moreover, experimental results in the scenario of random noise disturbance indicate that our method is more robust to random noise. In the future, we are interested in develop a nonlinear version of our method to handle more challenge scenarios.

Acknowledgment. This work was supported partially by National Key Technology Research and Development Program of the Ministry of Science and Technology of China (No. 2015BAK36B00), in part by the Key Science and Technology of Shenzhen (No. CXZZ20150814155434903), in part by the Key Program for International S&T Cooperation Projects of China (No. 2016YFE0121200), in part by the Key Science and Technology Innovation Program of Hubei (No. 2017AAA017), in part by the National Natural Science Foundation of China (No. 61571205), in part by the National Natural Science Foundation of China (No. 61772220).

References

1. Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P., Tandon, R.: Learning sparsely used overcomplete dictionaries. In: Conference on Learning Theory, pp. 123–137 (2014)
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 711–720 (1997)

3. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? *J. ACM (JACM)* **58**(3), 11 (2011)
4. Ding, Z., Fu, Y.: Low-rank common subspace for multi-view learning. In: 2014 IEEE International Conference on Data Mining (ICDM), pp. 110–119. IEEE (2014)
5. Ding, Z., Fu, Y.: Robust multi-view subspace learning through dual low-rank decompositions. In: AAAI, pp. 1181–1187 (2016)
6. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* **2**(1), 17–40 (1976)
7. He, X., Niyogi, P.: Locality preserving projections. In: Advances in Neural Information Processing Systems, pp. 153–160 (2004)
8. Hotelling, H.: Relations between two sets of variates. *Biometrika* **28**(3/4), 321–377 (1936)
9. Jhuo, I.H., Liu, D., Lee, D., Chang, S.F.: Robust visual domain adaptation with low-rank reconstruction. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2168–2175. IEEE (2012)
10. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.: Multi-view discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(1), 188–194 (2016)
11. Li, J., Wu, Y., Zhao, J., Lu, K.: Low-rank discriminant embedding for multiview learning. *IEEE Trans. Cybern.* **47**, 3516–3529 (2016)
12. Li, S., Fu, Y.: Robust subspace discovery through supervised low-rank constraints. In: Proceedings of the 2014 SIAM International Conference on Data Mining, pp. 163–171. SIAM (2014)
13. Lian, W., Rai, P., Salazar, E., Carin, L.: Integrating features and similarities: flexible models for heterogeneous multiview data. In: AAAI, pp. 2757–2763 (2015)
14. Lin, Z., Chen, M., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint [arXiv:1009.5055](https://arxiv.org/abs/1009.5055) (2010)
15. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 171–184 (2013)
16. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: Proceedings of the 27th International Conference on Machine Learning (ICML2010), pp. 663–670 (2010)
17. Liu, G., Yan, S.: Latent low-rank representation for subspace segmentation and feature extraction. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1615–1622. IEEE (2011)
18. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 689–696. ACM (2009)
19. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)
20. Rupnik, J., Shawe-Taylor, J.: Multi-view canonical correlation analysis. In: Conference on Data Mining and Data Warehouses (SiKDD 2010), pp. 1–4 (2010)
21. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2160–2167. IEEE (2012)
22. Silva, V.D., Tenenbaum, J.B.: Global versus local methods in nonlinear dimensionality reduction. In: Advances in Neural Information Processing Systems, pp. 721–728 (2003)

23. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: 2002 Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 53–58. IEEE (2002)
24. Turk, M., Pentland, A.: Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**(1), 71–86 (1991)
25. Van Der Maaten, L., Postma, E., Van den Herik, J.: Dimensionality reduction: a comparative. *J. Mach. Learn. Res.* **10**, 66–71 (2009)
26. Xu, J., Yu, S., You, X., Leng, M., Jing, X.Y., Chen, C.: Multi-view hybrid embedding: a divide-and-conquer approach. arXiv preprint [arXiv:1804.07237](https://arxiv.org/abs/1804.07237) (2018)
27. Yin, M., Gao, J., Lin, Z., Shi, Q., Guo, Y.: Dual graph regularized latent low-rank representation for subspace clustering. *IEEE Trans. Image Process.* **24**(12), 4918–4933 (2015)



An Online Learning Approach for Robust Motion Tracking in Liver Ultrasound Sequence

Chunxu Shen¹, Huabei Shi¹, Tao Sun¹, Yibin Huang³,
and Jian Wu^{1,2(✉)}

¹ Tsinghua University, Beijing 100084, China
{scxl6, shbl6, suntl6}@mails.tsinghua.edu.cn,
wuj@sz.tsinghua.edu.cn

² Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

³ Shenzhen Traditional Chinese Medicine Hospital, Shenzhen 518034, China
huangyb2004@126.com

Abstract. Suffering from respiratory motion and drift, radiotherapy requires real-time and accuracy motion tracking to minimize damage to critical structures and optimize dosage delivery to target. In this paper, we propose a robust tracker to minimize tracking error and enhance the quality of radiotherapy based on two-dimensional ultrasound sequences. We firstly develop a scale adaptive kernel correlation filter to compensate deformation. Then the filter with an improved update rule is utilized to predict target position. Moreover, displacement and appearance constrains are elaborately devised to restrict unreasonable positions. Finally, a weighted displacement is calculated to further improve the robustness. Proposed method has been evaluated on 53 targets, yielding 1.13 ± 1.07 mm mean and 2.31 mm 95%ile tracking error. Extensive experiments are performed between proposed and state-of-the-art algorithms, and results show our algorithm is more competitive. Favorable agreement between automatically and manually tracked displacements proves proposed algorithm has potential for target motion tracking in abdominal radiotherapy.

Keywords: Target tracking · Kernel correlation filter · Scale adaptation
Displacement and appearance constrain · Radiotherapy

1 Introduction

Motion in the abdomen is worth accounting for during radiotherapy image guided intervention [1] and focus ultrasound surgery [2]. The motion induced in abdominal organs is mainly due to breathing motion, drift and surgical instruments. Therefore, motion tracking of abdominal target is crucial to minimize the damage to surrounding crucial structure and optimize dosage delivery to target.

Respiratory gating is one of the most conventional approach to deal with abdomen motion, whereas it potentially increases treatment time [3]. Motion modeling like implanting fiducial markers to target region [4] is an alternative method, but it is usually at the expense of healthy tissue. Tracking base on medical image e.g. magnetic

resonance (MR), ultrasound (US) generally becomes a superior to localize abdomen target. De Senneville [5] generates an atlas of motion fields based on magnitude data of temperature-sensitive MR acquisitions. They suppose that motion of target region is periodic and can be estimated in the next moment, so it just recovers deformation caused by periodic component. 4D MR [6] is also introduced to respiratory motion reconstruction, but low signal-to-noise ratio and additional high cost must be considered in clinical practice. US is an appealing choice for abdominal target tracking, by contrast, as it has high temporal resolution and sub-millimeter spatial resolution along the beam direction.

Recently several literatures focus on tracking hepatic landmark and reconstructing liver motion of free breathing. Block matching [7], optical flow [8], particle filter [9], image registration and mechanical simulation [10] are widely investigated. Meanwhile temporal regularization [7] and distance metric [10] are also introduced to reject false tracking results. While some results have achieved a great process, many limitations remain to be discussed like tradeoff between real-time and accuracy, as well as robustness for acoustic shadowing and large deformation due to out-of-plane motion.

Our tracking approach is motivated by kernel correlation filter (KCF) [11], which achieves a fast and high performance on Visual Tracker Benchmark [12]. KCF provides an effective solution for translation, but its performance would degrade because of the scale and deformation of targets. Li et al. [13] suggests an effective scale adaptive scheme. Without discussing update strategy adequately, however, better tracking results cannot be remerged in US sequence. Besides, we integrate intensity feature, namely speckle patten, to proposed tracking frame as it includes much information about anatomical structure. In fact, if all the speckle patterns are stable, target motion can be easily reconstructed. Unluckily, speckle patterns are not identical because of out-of-plane motion and acoustic shadowing [14]. Moreover, similarity metrics is another important ingredient in proposed method. While mutual information (MI) has been suggested to be the most suitable metric for US to US match, high computation limits its usage in real-time target tracking. In this work, normalized cross-correlation (NCC) is chosen as it is easy to implement and effective to perform block matching.

In this work, we propose a real-time, robust tracking algorithm to compensate target motion in abdominal radiotherapy. Our contributions mainly focus on four aspects: first, we propose a scale adaptation strategy to alleviate deformation and scale change. Second, an improved update rule for proximate periodic motion is applied to reducing accumulation error in long-term tracking. Third, we integrate displacement and appearance constrains to proposed method in order to restrict unreasonable target prediction. And fourth, we suggest to use weighted displacement to determine target displacement.

2 Method

2.1 The KCF Tracker

In KCF tracker, Henriques et al. [11] suppose that the cyclic shifts version of base sample is approximate the dense samples over the base sample. Take one-dimension data $\mathbf{x} = [x_1, x_2, \dots, x_n]$ for example, a cyclic shift of \mathbf{x} is defined as

$\mathbf{P}\mathbf{x} = [x_n, x_1, x_2, \dots, x_{n-1}]$. Therefore, all the cyclic shift samples, $\{\mathbf{P}^u\mathbf{x}|u = 0, \dots, n - 1\}$, can be concatenated to form sample matrix \mathbf{X} , which also called circulant matrix as the matrix is purely generated by the cyclic shifts of \mathbf{x} . This matrix has a helpful property that all the circulant matrices can be formulated as follows:

$$\mathbf{X} = \mathbf{F}^H \text{diag}(\mathbf{F}\mathbf{x})\mathbf{F} \tag{1}$$

Where, \mathbf{F} is the Discrete Fourier Transformation (DFT) matrix. \mathbf{F}^H is the Hermitian transpose of \mathbf{F} . Benefit from the decomposition of circulant matrix, it can be used to the solution of linear regression. Moreover, the objective function of linear ridge regression can be written as:

$$\min_{\mathbf{w}} \sum_i^n (f(\mathbf{x}_i) - y_i)^2 + \lambda \|\mathbf{w}\| \tag{2}$$

Where, f is linear combination of basis samples, $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. The ridge regression has a close-form solution, $\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. The solution can be rewritten with Eq. 1, $\hat{\mathbf{w}}^* = \frac{\hat{\mathbf{x}}^* \odot \hat{\mathbf{y}}}{\hat{\mathbf{x}}^* \odot \hat{\mathbf{x}} + \lambda}$. Where, $\hat{\mathbf{x}} = \mathbf{F}\mathbf{x}$ donates the DFT of \mathbf{x} ; $\hat{\mathbf{x}}^*$ is the complex-conjugate of $\hat{\mathbf{x}}$; \odot denotes element-wise multiplication. So during the process of extracting patches explicitly and solving a general regression problem, this step can save much computational cost. In order to construct a more powerful classifier in case of non-linear regression, Henriques et al. [11] adopt a kernel tracker, $f(\mathbf{z}) = \mathbf{w}^T\mathbf{z} = \sum_{i=1}^n \alpha_i \mathcal{K}(\mathbf{z}, \mathbf{x}_i)$. Then dual space confident α can be learned as follows:

$$\hat{\alpha}^* = \frac{\hat{\mathbf{y}}}{\hat{\mathbf{k}}^{\mathbf{xx}} + \lambda} \tag{3}$$

$\mathbf{k}^{\mathbf{xx}}$ is defined as kernel correlation. Similar to the linear classifier, the dual coefficients are learned in Fourier domain. \mathbf{y} is a regression target vector in Fourier domain and has the same size with \mathbf{x} ; λ is regularization weight in ridge regression. Note that the search window, which is the size of \mathbf{x} , has 2.5 times the size of the target in the implementation of KCF. In case of Gaussian kernel function, the kernel correlation can be denoted as:

$$\mathbf{k}^{\mathbf{xx}'} = \exp \left(-\frac{1}{\sigma^2} \left(\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 \right) - 2\mathbf{F}^{-1}(\hat{\mathbf{x}} \odot \hat{\mathbf{x}}'^*) \right) \tag{4}$$

Where \mathbf{F}^{-1} denotes inverse Fourier transform.

In detection step, the regression function Eq. 5 is applied to predict the position of target where the maximum regression value locates.

$$\hat{\mathbf{f}}(\mathbf{z}) = \left(\hat{\mathbf{k}}^{\tilde{\mathbf{x}}\mathbf{z}} \right)^* \odot \hat{\alpha} \tag{5}$$

Where $\tilde{\mathbf{x}}$ denotes basic data template to be learned in the model; \mathbf{z} is the candidate patch, which has the same size and location with \mathbf{x} in next frame. When we transform

$\hat{f}(z)$ back into the spatial domain, the translation with respect to the maximum response is considered as the displacement of the tracked target.

2.2 Scale Adaptive KCF

Deformations and scale variations of targets is potential to increase the tracking error and reduce robustness, even fail. However, these negative factors are common in abdominal targets. In our clinical practice, there are two situations leading to target deformation. First, with the contraction and relaxation of the diaphragm in free breathing situation [15], the hepatic targets would suffer from deformation. Second, because of free breathing and drift, the appearance of cross section between ultrasound beam and targets would change. In this part, we propose a scale adaptive strategy to compensate these deformations and scale variations.

Suppose that the size of search window sets as $\mathbf{s}_T = (s_x, s_y)$, we define a scaling pool $\boldsymbol{\eta} = \{\eta_1, \eta_2, \dots, \eta_m\}$ to expand search range to different scale space, which can be donated as $\tilde{\mathbf{s}}_T = \{(\eta_i s_x, \eta_j s_y) | \eta_i, \eta_j \in \boldsymbol{\eta}\}$. Because the dot-product requires the search window with the fixed size in kernel correlation filter, we resize $\tilde{\mathbf{s}}_T$ into the fixed size of \mathbf{s}_T using bilinear-interpolation. Note that our proposed scale adaptive method is different from Li's work [13], which adopts $\check{\mathbf{s}}_T = \{\eta_i \mathbf{s}_T | \eta_i \in \boldsymbol{\eta}\}$. Therefore, the response $\mathbf{R}(\eta_i, \eta_j)$ in difference scale space can be calculated.

$$\mathbf{R}(\eta_i, \eta_j) = \mathbf{F}^{-1} \hat{f}(z(\eta_i, \eta_j)) \tag{6}$$

Where $z(\eta_i, \eta_j)$ is the sample patch resampled by scaling pool and the size of $z(\eta_i, \eta_j)$ is $(\eta_i s_x, \eta_j s_y)$, which is subsequently resized to the fixed size of \mathbf{s}_T .

2.3 Improved Update Rule for Approximate Periodic Motion

According to Eq. 5, there are two sets of coefficient should be update. One is dual space coefficient $\boldsymbol{\alpha}$, another is basic template $\tilde{\mathbf{x}}$. Original update rule is realized by combining new filter with old one linearly as Eq. 7 illustrates.

$$\begin{cases} \tilde{\mathbf{x}}_{t+1} = \mu \tilde{\mathbf{x}}_{t+1} + (1 - \mu) \tilde{\mathbf{x}}_t \\ \hat{\boldsymbol{\alpha}}_{t+1} = \mu \hat{\boldsymbol{\alpha}}_{t+1} + (1 - \mu) \hat{\boldsymbol{\alpha}}_t \end{cases} \tag{7}$$

Where μ is the linear interpolation factor.

While the update rule above achieves impressive success for nature video tracking, it is so sensitive that cannot support for long-term tracking in our work. An explanation is that Eq. 7 pays more attention to learn new characteristics from a new image. Once ultrasound images suffer from noise severely, like acoustic shadowing and speckle decorrelation, the performance of online classifier could degrade largely. With prior knowledge that motion of liver is approximate periodic in free breathing, the target in first frame would also appear in subsequent sequence. Therefore, an improved update rule for long-term tracking of approximate periodic motion is proposed as Eq. 8 shows:

$$\begin{cases} \tilde{\mathbf{x}}_{t+1} = \beta \tilde{\mathbf{x}}_1 + (1 - \beta - \mu) \tilde{\mathbf{x}}_{t+1} + \mu \tilde{\mathbf{x}}_t \\ \hat{\boldsymbol{\alpha}}_{t+1} = \beta \hat{\boldsymbol{\alpha}}_1 + (1 - \beta - \mu) \hat{\boldsymbol{\alpha}}_{t+1} + \mu \hat{\boldsymbol{\alpha}}_t \end{cases} \quad (8)$$

Where β is recurrence factor.

2.4 Restricting Unreasonable Target Prediction

Though NCC has been a popular similarity measure in speckling tracking, it still suffers from acoustic shadowing, speckle decorrelation and other artifacts. Here, in order to alleviate these adverse effect, we provide displacement and appearance constrains to restrict unreasonable target prediction.

Displacement Constrain. In clinical ultrasound image guided abdominal radiotherapy, we notice that the target displacement in two consecutive frames is very small (<3 mm, acquisition frequency is 13–23 Hz). So a displacement cost function is employed to restrict unreasonable prediction. Suppose that $\mathbf{D} = (\Delta \mathbf{x}(\eta_i, \eta_j), \Delta \mathbf{y}(\eta_i, \eta_j))$ is the displacement prediction and $\mathbf{R}(\mathbf{d}_{ij} | \mathbf{d}_{ij} \in \mathbf{D})$ is corresponding response map, therefore, the response with displacement constrain can be expressed by:

$$\mathbf{R}_{\text{dis}}(\eta_i, \eta_j) = \mathbf{R}(\eta_i, \eta_j) \odot \exp\left(-\frac{\Delta \mathbf{x}^2 + \Delta \mathbf{y}^2}{\sigma_{\text{dis}}}\right) \quad (9)$$

Where σ_{dis} is the bandwidth of displacement constrain.

Appearance Constrain. For alleviating the unreasonable matching from NCC, we also employ a set of confidence response to determine target displacement instead of selecting the displacement that the best response locates. Supposing the threshold of confidence response is θ_{app} , the appearance constrain can be expressed as Eq. 10 shows.

$$\mathbf{R}_{\text{dis}}^{\text{app}}(\eta_i, \eta_j) = \begin{cases} \mathbf{R}_{\text{dis}}(\eta_i, \eta_j), & \text{if } \mathbf{R}_{\text{dis}} \geq \max\{\mathbf{R}_{\text{dis}}\} \cdot \theta_{\text{app}} \\ 0, & \text{others} \end{cases} \quad (10)$$

With constrains of displacement and appearance, the best scale space can be determined by maximize the average response $\overline{\mathbf{R}_{\text{dis}}^{\text{app}}(\eta_i, \eta_j)}$ in Eq. 10:

$$\arg \max \overline{\mathbf{R}_{\text{dis}}^{\text{app}}(\eta_i, \eta_j)} \quad (11)$$

2.5 Weighted Displacement

Motivated by Carletti’s work [9], a weighted displacement is calculated to enhance the robustness of proposed tracking algorithm. The displacements used to calculate

weighted displacement are from Eq. 10, namely $r_{ij} \in \mathbf{R}_{\text{dis}}^{\text{app}}(\eta_i, \eta_j)$. Finally the target displacement can be determined in adjacent frames.

$$\bar{\mathbf{d}} = \frac{\sum_{i=1}^M \sum_{j=1}^N r_{ij} \mathbf{d}_{ij}}{\sum_{i=1}^M \sum_{j=1}^N r_{ij}} \quad (12)$$

Note that $\bar{\mathbf{d}}$ is the displacement in best scale space, we get the real displacement $\bar{\mathbf{d}}_r$ by performing scale inverse transformation with scale parameters from Eq. 11. Therefore, by combining the target position in last frame \mathbf{p}_{old} and displacement $\bar{\mathbf{d}}_r$, new target position \mathbf{p}_{new} in current frame can be determined.

$$\mathbf{p}_{\text{new}} = \mathbf{p}_{\text{old}} + \bar{\mathbf{d}}_r \quad (13)$$

Finally, the overall algorithm is summarized into Algorithm 1

Algorithm 1. Overall of proposed tracking algorithm

Require:

- The template of tracked target, $\tilde{\mathbf{x}}$;
- The dual space coefficient, α ;
- The newly search window, \mathbf{y} ;
- The last target position, \mathbf{p}_{old} ;

Ensure:

- The updated template of tracked target, $\tilde{\mathbf{x}}$;
 - The updated dual space coefficient, α ;
 - The new target position, \mathbf{p}_{new} ;
- 1: **for** every $(\eta_i s_x, \eta_j s_y)$ in $\tilde{\mathbf{s}}_{\mathbf{T}}$ **do**
 - 2: Sample the new search window $\mathbf{z}(\eta_i, \eta_j)$ based on size $(\eta_i s_x, \eta_j s_y)$ and resize it to $\mathbf{s}_{\mathbf{T}}$ with bilinear interpolation
 - 3: Calculate the corresponding response $\mathbf{R}(\eta_i, \eta_j)$ in different scale space with Eqn. 5
 - 4: Apply displacement and appearance constrains with Eqn.9 and Eqn.10
 - 5: Get the best scale space and corresponding scale parameters
 - 6: Calculate the weighted displacement $\bar{\mathbf{d}}$ in best scale space with Eqn.11
 - 7: **end for**
 - 8: Get the real displacement $\bar{\mathbf{d}}_r$ with scale parameters
 - 9: Get the new target position \mathbf{p}_{new} with Eqn.13.
 - 10: Get the new $\tilde{\mathbf{x}}$ and α base on the new position \mathbf{p}_{new} with Eqn.3.
 - 11: Update $\tilde{\mathbf{x}}$ and α with Equation 8.
 - 12: **return** update $\tilde{\mathbf{x}}$ and α ;
-

3 Experiments and Results

3.1 Dataset and Parameter Settings

Datasets and Resource. Our 2D liver ultrasound sequences are provided by MICCAI 2015 Challenge on Liver Ultrasound Tracking (CLUST) [16] training database, and it consists of five different datasets CIL, ETH, ICR, MED1 and MED2. Each dataset is acquired by different scanner with different image resolution (0.30–0.55 mm) and acquisition frequency (13–23 Hz). Besides, our code is implemented using MATLAB R2017b on an Intel Core i7-4910MQ CPU @ 2.90 GHz.

Parameter Settings. The parameters in our algorithm come from two parts. One is from the original KCF tracker and we adopt the default parameters as [11] recommends. The learning rate λ in Eqs. 2 and 3 sets to 10^{-4} ; the σ used in Gaussian function Eq. 4 sets to 0.2; the linear interpolation factor μ in Eq. 8 sets to 0.1; and the size of search window is 2.5 times to the size of target. Another part is from our contributions, which is used to ensure proposed tracker more accuracy and robust. We adopt scaling pool with the suggestion from our experienced radiologist $\eta = \{0.85, 0.90, 0.95, 1.00, 1.05, 1.10, 1.15\}$. And the recurrence factor β in Eq. 8, bandwidth of displacement σ_{dis} in Eq. 9 and the threshold of confidence response θ_{app} in Eq. 10 set to 0.15, 10 and 0.95 respectively. Parameters are same for all following experiments.

Note that proposed method needs image patches as initialization. Therefore, we generate a rectangular region manually with the guidance of experienced radiologist in the first frame. During online tracking process, the center of rectangular region is recorded and then used to evaluate tracking performance.

3.2 Tracking Results

We employ Euclidean distance suggested by Organizers of CLUST [16] to evaluate the tracking performance. In our experiments, we compute errors between each manual annotation and the output of proposed algorithm, and then mean, standard deviation (SD), 95%ile and maximum errors are counted. Additionally, processing speed is estimated by counting frames that are tracked per second (FPS).

Performance Evaluation on CLUST. Firstly, we evaluate the performance of proposed tracking algorithm using the five datasets of CLUST database. The number of objects means the total objects being tracked in corresponding dataset. The following Table 1 shows the tracking error distribution of each dataset and the total 2D ultrasound sequences respectively.

Comparison Proposed with Baseline Algorithm. Then a performance comparison experiment is performed between proposed and baseline algorithm, and the results are shown in Fig. 1.

Compared with baseline algorithm, proposed method achieves state-of-the-art results with mean decreasing by 78.8% (from 5.33 mm to 1.13 mm), 95%ile error

decreasing by 77.1% (from 10.08 mm to 2.31 mm) and maximum error decreasing by 82.8% (from 66.10 mm to 11.37 mm) respectively.

Table 1. A summary for performance evaluation on CLUST. All tracking errors are in millimeters and processing speed is presented by frames per second.

Dataset	No. objects	Mean	SD	95%ile	Maximum	FPS
CIL	3	0.99	1.16	2.02	3.61	20.33
ETH	16	0.89	0.60	1.73	4.18	23.57
ICR	12	1.00	0.54	2.31	6.23	23.14
MED1	19	1.39	1.62	2.74	11.37	22.21
MED2	3	1.38	2.04	3.01	7.88	31.00
Total	53	1.13	1.07	2.31	11.37	23.22

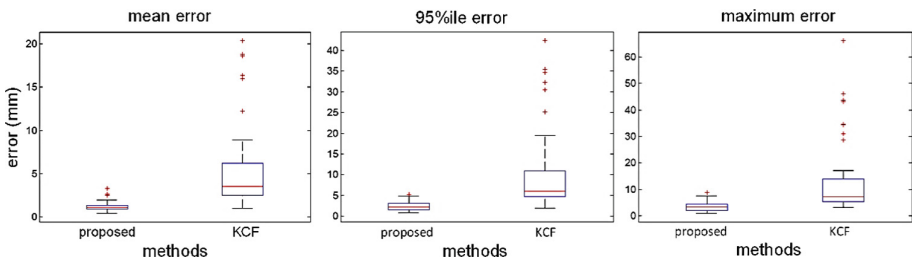


Fig. 1. Tracking errors comparison between proposed and baseline algorithm on CLUST. Left is mean tracking error; middle is 95%ile error; right is maximum error.

Comparison Proposed with State-of-the-art Algorithms. Extensive comparison experiments are performed among our tracker and some state-of-the-art trackers. The following Table 2 gives a summary of tracking error distribution. It is worth mentioning that we compare these algorithms whose tracking performance is also evaluated on CLUST training database. Compared with TMG [17], RMTwS [17] and Hybrid [18], proposed algorithm achieves a competitive accuracy with maximum tracking error decreasing by 40.4%–47.8%, which means it would provide a more effective guidance for clinical operation. Experimental results also indicate our tracker is more real time than the existing state-of-the-art trackers.

Table 2. Comparison of published results with our tracking results. All tracking errors are in millimeters and processing speed is presented by frames per second.

Algorithms	Mean	SD	95%ile	Maximum	FPS
Proposed	1.13	1.07	2.31	11.37	17–34
TMG [17]	1.17	0.89	2.61	21.78	8–23
RMTwS [17]	1.12	0.81	2.19	21.78	3–16
Hybrid [18]	0.80	0.80	1.85	19.08	8–32

3.3 Experimental Analysis

In this section, we first perform an ablation analysis to understand the benefit of scale adaptive strategy. Then a detailed parameters analysis are performed to find out the effectiveness of improved update rule (Eq. 8) and appearance/displacement (Eqs. 9 and 10) constraints.

Ablation Study About Scale Adaptive Strategy. Deformation is common in liver ultrasound sequence. In this part, we perform a comparison experiment between non-rigid (with Eq. 6) and rigid (without Eq. 6) tracking. Results are shown in Fig. 2.

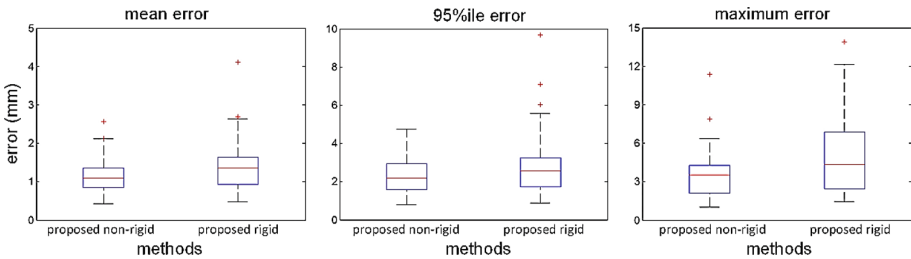


Fig. 2. Tracking error distributions (mm) for proposed non-rigid and rigid tracking method. Left is mean error; middle is 95%ile error, right is maximum error.

Compared with rigid tracking, non-rigid tracking achieves a better performance with mean decreasing by 20.4% (from 1.42 mm to 1.13 mm), 95%ile error decreasing by 19.5% (from 2.87 mm to 2.31 mm) and maximum error decreasing by 18.1% (from 13.88 mm to 11.37 mm) respectively. That means non-rigid deformation should be considered seriously in precise radiotherapy.

Figure 3 shows an instance to compare the results from non-rigid and rigid tracking. The target position calculated by rigid tracking yields larger deviations, by contrast, the positions from proposed method are more accurate and robust.

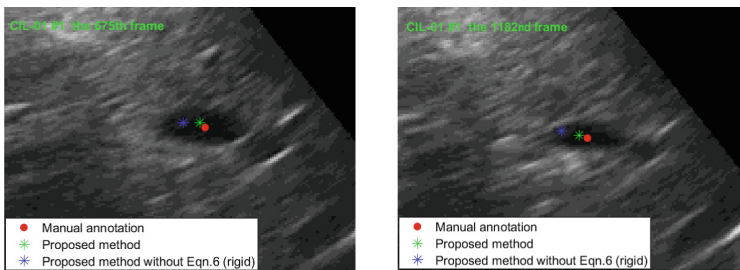


Fig. 3. An example for showing deviation between non-rigid and rigid tracking. Images are both from CIL-01 #1 in CLUST. Left is the 675th frame and right is the 1182nd frame.

Parameters Analysis. There are four parameters, $[\boldsymbol{\eta}, \sigma_{\text{dis}}, \theta_{\text{app}}, \beta]$, needing more discussion. Among them, scaling pool $\boldsymbol{\eta}$ can be designed when the deformation of target is estimated. And we also can determine σ_{dis} by magnitude of target motion and frequency of image acquisition. However, θ_{app} and β are assigned empirically. In this part, we investigate the effect when we change the threshold of confidence response and recurrence factor. Without loss of generality, we choose $\theta_{\text{app}} \in [0.90, 0.95, 1.00]$ and $\beta \in [0.10, 0.15, 0.20]$ to perform parameters analysis on CLUST training database. Here, mean and 95%ile tracking errors, as regardful indicators for our project, are chosen to evaluate the results of parameters analysis. Results are shown in Fig. 4 and Table 3.

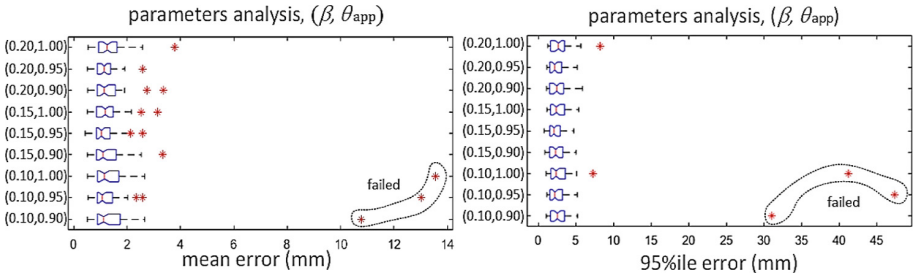


Fig. 4. The results of parameters analysis on CLUST. Left is results of mean error and right is results of 95%ile error with parameters $(\beta, \theta_{\text{app}})$ changing.

Table 3. Statistic results of mean and 95%ile errors (mm) with $(\beta, \theta_{\text{app}})$ changing.

$(\beta, \theta_{\text{app}})$	(0.10,0.90)	(0.10,0.95)	(0.10,1.00)	(0.15,0.90)	(0.15,0.95)
mean	1.45	1.43	1.53	1.24	1.13
95%ile	3.23	3.36	3.54	2.54	2.31
$(\beta, \theta_{\text{app}})$	(0.15,1.00)	(0.20,0.90)	(0.20,0.95)	(0.20,1.00)	
mean	1.26	1.24	1.18	1.32	
95%ile	2.65	2.56	2.43	2.78	

Therefore, recurrence factor is a crucial parameter in proposed algorithm. A smaller β has a terrible effect on long-term tracking (like $\beta = 0.10$, see Fig. 4). But a larger one would also enlarge tracking error by unduly limiting learning ability for proposed method. Besides, a smaller or larger θ_{app} are not a wise chose, which would potentially introduce more unreasonable position or be not adaptive for artifacts well respectively. Therefore, (0.15, 0.95) is a better combination for accuracy and robust tracking in our project.

4 Conclusion and Discussion

In this paper, we present a 2D real-time tracking approach, which consists four steps namely (1) initial target regions selection, (2) tracking with scale adaptive kernel correlation filter, (3) displacement and appearance constrains, and (4) weighted displacement. The initial target regions are generated by our experienced radiologist. Then we train an online classifier to predict targets position. Because deformation of targets can lead to error accumulation in learning phase, we employ adaptive scale strategy to mitigate this adverse effect. Considering US images suffer from acoustic shadowing and speckle decorrelation, NCC is more susceptible to bias. We employ displacement and appearance constrains to constrict unreasonable position prediction by carefully investigating the motion extents of landmarks in liver under free breathing. Furthermore, with prior knowledge that target motion in liver is approximately periodic under free breathing, we revise the update rule by introducing a recurrence factor to improve robustness in long-term tracking. Finally, inspired by success of particle filter in noise circumstance, we obtain new target positions by calculating weighted displacement.

However, we just adopt single feature to realize target tracking. Accuracy and robustness for proposed method may continue to improve by combining other image features like texture and shape, which is a major research direction for future work. Also, similarity metrics is a core ingredient for target tracking. While a large of similarity metrics have been proposed in computer vision community, there are no clear rules about how to select the most suitable one but to try them in different condition.

There are several avenues of future work that would potentially improve proposed method. Integrating texture feature into our tracking method would be helpful to improve accuracy. And adaptive recurrence factor strategy will be investigated to improve robustness for long-time tracking.

In conclusion, we propose an online learning approach for robust and real-time motion tracking in liver ultrasound sequences and evaluate it on five different datasets. Favorable agreement between automatically and manually tracked displacements, along with real-time processing speed prove that proposed algorithm has potential for target motion tracking in abdominal radiotherapy.

Acknowledgement. This work is supported in part by Knowledge Innovation Program of Basic Research Projects of Shenzhen under Grant JCYJ20160428182053361, in part by Guangdong Science and Technology Plan under Grant 2017B020210003 and in part by National Natural Science Foundation of China under Grant 81771940, 81427803.

References

1. Riley, C., Yang, Y., Li, T., Zhang, Y., Heron, D.E., Huq, M.S.: Dosimetric evaluation of the interplay effect in respiratory-gated RapidArc radiation therapy. *Med. Phys.* **41**, 011715 (2014)
2. Jenne, J.W., Preusser, T., Günther, M.: High-intensity focused ultrasound: principles, therapy guidance, simulations and applications. *Zeitschrift Für Medizinische Physik* **22**, 311–322 (2012)

3. Okada, A., et al.: A case of hepatocellular carcinoma treated by MR-guided focused ultrasound ablation with respiratory gating. *Magn. Reson. Med. Sci. Mrms Off. J. Jpn. Soc. Magn. Reson. Med.* **5**, 167 (2006)
4. Kothary, N., Dieterich, S., Louie, J.D., Chang, D.T., Hofmann, L.V., Sze, D.Y.: Percutaneous implantation of fiducial markers for imaging-guided radiation therapy. *AJR Am. J. Roentgenol.* **192**, 1090–1096 (2009)
5. de Senneville, B.D., Mougnot, C., Moonen, C.T.: Real-time adaptive methods for treatment of mobile organs by MRI-controlled high-intensity focused ultrasound. *Magn. Reson. Med.* **57**, 319–330 (2007)
6. Rank, C.M., et al.: 4D respiratory motion-compensated image reconstruction of free-breathing radial MR data with very high undersampling. *Magn. Reson. Med.* **77**, 1170 (2016)
7. De Luca, V., Tschannen, M., Székely, G., Tanner, C.: A learning-based approach for fast and robust vessel tracking in long ultrasound sequences. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) *MICCAI 2013. LNCS*, vol. 8149, pp. 518–525. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_65
8. Chuang, B., Hsu, J.H., Kuo, L.C., Jou, I., Su, F.C., Sun, Y.N.: Tendon-motion tracking in an ultrasound image sequence using optical-flow-based block matching. *Biomed. Eng. Online* **16**, 47 (2017)
9. Carletti, M., Dall’Alba, D., Cristani, M., Fiorini, P.: A robust particle filtering approach with spatially-dependent template selection for medical ultrasound tracking applications. In: 11th International Conference on Computer Vision Theory and Applications, pp. 522–531. SCITE Press, Rome (2016)
10. Royer, L., Krupa, A., Dardenne, G., Le, B.A., Marchand, E., Marchal, M.: Real-time target tracking of soft tissues in 3D ultrasound images based on robust visual information and mechanical simulation. *Med. Image Anal.* **35**, 582–598 (2017)
11. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 583–596 (2015)
12. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411–2418. IEEE press, Portland (2013)
13. Li, Y., Zhu, J.: A scale adaptive kernel correlation filter tracker with feature integration. In: Agapito, L., Bronstein, Michael M., Rother, C. (eds.) *ECCV 2014. LNCS*, vol. 8926, pp. 254–265. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16181-5_18
14. Liang, T., Yung, L., Yu, W.: On feature motion decorrelation in ultrasound speckle tracking. *IEEE Trans. Med. Imaging* **32**, 435–448 (2016)
15. Lei, P., Moeslein, F., Wood, B.J., Shekhar, R.: Real-time tracking of liver motion and deformation using a flexible needle. *Int. J. Comput. Assist. Radiol. Surg.* **6**, 435–446 (2011)
16. Luca, V.D., et al.: The 2014 liver ultrasound tracking benchmark. *Phys. Med. Biol.* **60**, 5571–5599 (2015)
17. Ozkan, E., Tanner, C., Kastelic, M., Mattausch, O., Makhinya, M., Goksel, O.: Robust motion tracking in liver from 2D ultrasound images using supporters. *Int. J. Comput. Assist. Radiol. Surg.* **12**, 941–950 (2017)
18. Williamson, T., Cheung, W., Roberts, S.K., Chauhan, S.: Ultrasound-based liver tracking utilizing a hybrid template/optical flow approach. *Int. J. Comput. Assist. Radiol. Surg.* **13**, 1–11 (2018)



Set-to-Set Distance Metric Learning on SPD Manifolds

Zhi Gao, Yuwei Wu^(✉), and Yunde Jia

Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology (BIT),
Beijing 100081, People's Republic of China
{gaozhi_2017, wuyuwei, jiayunde}@bit.edu.cn

Abstract. The Symmetric Positive Definite (SPD) matrix on the Riemannian manifold has become a prevalent representation in many computer vision tasks. However, learning a proper distance metric between two SPD matrices is still a challenging problem. Existing metric learning methods of SPD matrices only regard an SPD matrix as a global representation and thus ignore different roles of intrinsic properties in the SPD matrix. In this paper, we propose a novel SPD matrix metric learning method of discovering SPD matrix intrinsic properties and measuring the distance considering different roles of intrinsic properties. In particular, the intrinsic properties of an SPD matrix are discovered by projecting the SPD matrix to multiple low-dimensional SPD manifolds, and the obtained low-dimensional SPD matrices constitute a set. Accordingly, the metric between two original SPD matrices is transformed into a set-to-set metric on multiple low-dimensional SPD manifolds. Based on the learnable alpha-beta divergence, the set-to-set metric is computed by summarizing multiple alpha-beta divergences assigned on low-dimensional SPD manifolds, which models different roles of intrinsic properties. The experimental results on four visual tasks demonstrate that our method achieves the state-of-the-art performance.

Keywords: SPD manifold · Metric learning · Set-to-set metric
Multiple manifolds

1 Introduction

The Symmetric Positive Definite (SPD) matrix has become a prevalent representation in many visual tasks, such as face recognition [12], action recognition [30], and object detection [25]. It utilizes the second-order or higher-order statistics information to capture the desirable feature distribution. There are several works try to model a more discriminative SPD matrix [16, 27, 28] from local features. Meanwhile, calculating the distance metric in the SPD manifold is a crucial problem coming along with the SPD matrix representation. Due to the no-Euclidean structure of SPD manifolds, the Euclidean metric can't be applied

directly on it. In this paper, we focus on a robust metric learning method on SPD manifolds.

Many efforts have been devoted to the SPD matrix metric, such as the Affine Invariant Metric (AIM) [19], Log-Euclidean Metric (LEM) [2], Bregman divergence [14], Stein divergence [21], and alpha-beta divergence [3, 4, 22]. Given a concrete metric, metric learning aims at learning proper metric parameters that keep similar pairs close and separate dissimilar pairs. Most of the existing metric learning methods on the SPD manifold learn a discriminative metric on the tangent Euclidean space [11, 23, 31].

However, how to learn a proper SPD matrix metric is still a challenging problem. The SPD matrix is aggregated from local features, and contains different essential intrinsic properties. Existing SPD matrix metric learning methods [11, 23, 31] just regard an SPD matrix as a global representation and exploit a direct metric on the complex manifold, ignoring the different roles of intrinsic properties in the SPD matrix. It is unsuitable to treat intrinsic properties equally when they have different roles, *e.g.*, different distribution or significance. Therefore, we argue that an SPD matrix metric modeling different roles of intrinsic properties will achieve a better performance.

In this paper, a novel metric learning method on SPD manifolds is proposed to solve the issues mentioned above. Firstly we discover intrinsic properties of an SPD matrix, and then calculate the SPD matrix metric considering different roles of them. In particular, our method aims to jointly learn multiple low-dimensional projections and a set-to-set metric. As the property discovery can be seen as the feature extraction, we apply multiple low-dimensional manifold projections on the SPD matrix to discover discriminative intrinsic properties. Thus, the distance metric between two original SPD matrices is transformed into the distance metric between the two sets which contain several corresponding projected low-dimensional SPD matrices. The alpha-beta divergences is a learnable SPD matrix metric, so it is applied in our set-to-set metric to be adaptive to the intrinsic property. We assign multiple alpha-beta divergences on different low-dimensional manifolds as the sub-metrics and summarize these sub-metrics discriminatively as the SPD matrix metric. Through this, the different roles of intrinsic properties are involved in the SPD matrix metric. Evaluated by experiments, the proposed learnable metric is extremely helpful to capture meaningful nearest neighbors of different original SPD matrices.

In summary, our contributions are three-fold.

- (1) We propose a robust SPD matrix metric learning method of discovering discriminative intrinsic properties and modeling their different roles in metric computation.
- (2) We formulate the metric learning as the two-component joint optimization problem, *i.e.*, multiple low-dimensional manifold projections and a set-to-set metric are learned jointly.
- (3) We introduce the manifold optimization method which can learn metric parameters to guarantee the robustness of the proposed metric.

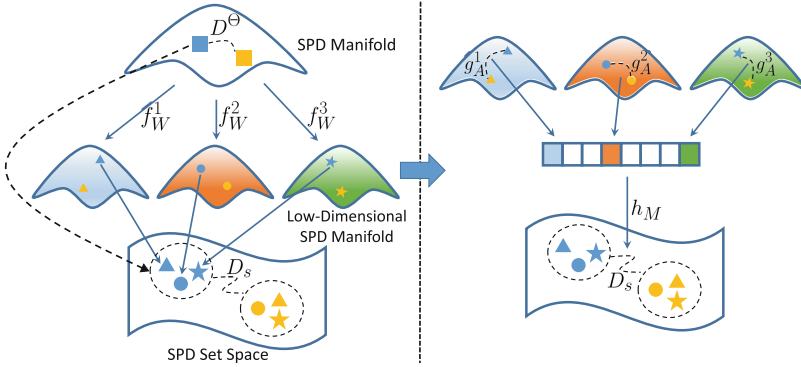


Fig. 1. The flowchart of our SPD matrix metric learning method. Left: multiple projections f_W^1 , f_W^2 , and f_W^3 used to discover intrinsic properties; Right: the computation of the set-to-set distance D_s which considers different roles of intrinsic properties.

2 The Proposed Method

Throughout this paper, scalars are denoted by the lower-case letters; the vectors are represented by the bold lower-case letters; the matrices are denoted by the upper-case letters; the sets are represented by the bold upper-case letters.

2.1 Problem Definition

This work aims to discover discriminative intrinsic properties in an SPD matrix and compute the distance of SPD matrices considering different roles of discovered properties. The property discovery can be regarded as a feature extraction process that projects an original SPD matrix to multiple low-dimensional SPD manifolds to form a set of the low-dimensional SPD matrices. We propose a set-to-set metric to consider different roles of intrinsic properties. Individual sub-metrics are assigned on low-dimensional manifolds and summarized discriminatively. Consequently, our metric learning method is composed of two components, multiple low-dimensional manifold projections and a set-to-set metric. Given two SPD matrices X_i and X_j , the distance $D^\Theta(X_i, X_j)$ is

$$\begin{aligned}
 D^\Theta(X_i, X_j) &= D_s(\mathbf{X}_i, \mathbf{X}_j) \\
 &= D_s\left(\{f_W^1(X_i), \dots, f_W^m(X_i)\}, \{f_W^1(X_j), \dots, f_W^m(X_j)\}\right) \quad (1) \\
 &= h_M\left(g_A^1(f_W^1(X_i), f_W^1(X_j)), \dots, g_A^m(f_W^m(X_i), f_W^m(X_j))\right),
 \end{aligned}$$

where $f_W^k(\cdot)$ is the low-dimensional manifold projection, and $\mathbf{X}_i = \{f_W^k(X_i)\}_{k=1}^m$ is the set containing low-dimensional SPD matrices. The distance $D^\Theta(X_i, X_j)$ between original SPD matrices X_i and X_j is transformed into a set-to-set distance $D_s(\mathbf{X}_i, \mathbf{X}_j)$, where the sub-metric on the k -th low-dimensional manifold

is calculated by $g_A^k(\cdot, \cdot)$ and all sub-metrics of properties are summarized by $h_M(\cdot)$. W, A, M are the projection parameter, the sub-metric parameter, and the summarization parameter, respectively. We exploit a learnable parameter set $\Theta = \{W, A, M\}$ to represent the parameters. The framework of our metric learning method for the SPD matrix is shown in Fig. 1.

The goal of metric learning is to learn the metric parameter Θ from an SPD matrix similar pair set \mathcal{S} , a dissimilar pair set \mathcal{D} , and their labels Y , where $y_{ij} = 1$ means X_i and X_j are similar, otherwise $y_{ij} = 0$. The metric parameter Θ can be learned by optimizing the loss function $\mathcal{L}(\Theta, \mathcal{S}, \mathcal{D}, Y)$ which is the punishment of both far similar sample pairs and close dissimilar sample pairs. We define $\mathcal{L}(\Theta, \mathcal{S}, \mathcal{D}, Y)$ in the following subsection. Moreover, we impose the manifold constraints on W and M to obtain a more robust metric.

2.2 Multiple Low-Dimensional Manifold Projections

For an SPD matrix sample $X_i \in \mathbb{R}^{n \times n}$, we project X_i to m low-dimensional manifolds to discover the intrinsic properties,

$$\begin{aligned} X_i^1 &= f_W^1(X_i) = W_1^\top X_i W_1 \\ &\dots \\ X_i^m &= f_W^m(X_i) = W_m^\top X_i W_m, \end{aligned} \tag{2}$$

where $X_i^k \in \mathbb{R}^{p \times p}$ is the k -th low-dimensional SPD matrix, $k \in \{1, 2, \dots, m\}$. An SPD matrix X_i is projected to a set $\mathbf{X}_i = \{X_i^k\}_{k=1}^m$, which contains several low-dimensional SPD matrices.

We expect that each low-dimensional matrix X_i^k is guaranteed to be still an SPD matrix having the ability of capturing desirable feature distribution, and any two low-dimensional SPD manifolds are unrelated to preserve as much information as possible in the low-dimensional SPD matrix set. The learnable parameter W_k needs to be a column full rank matrix to make X_i^k be an SPD matrix as well. Based on the affine invariance [3, 7] of the alpha-beta divergence, we relax the column full rank constraint of W_k to the semi-orthogonal constraint, *i.e.*, $W_k^\top W_k = I_p$. In order to preserve more information in the $\mathbf{X}_i = \{X_i^k\}_{k=1}^m$ set, we expect that any two low-dimensional manifolds have a low relevance. For any $k \neq l$, we set $W_k^\top W_l = \mathbf{0}$, where $\mathbf{0} \in \mathbb{R}^{p \times p}$ is a matrix whose elements are all "0"s, to reduce relevance between X_i^k and X_i^l . These low-dimensional SPD manifolds can be seen as analogies of different PCA subspaces. A total projection matrix W is composed of all W_k , $W = [W_1, W_2, \dots, W_m] \in \mathbb{R}^{n \times mp}$, in which W_k is a partitioned matrix of W containing p columns. Note that, W is a semi-orthogonal matrix, *i.e.*, $W^\top W = I_{mp}$, which is on the non-Euclidean Stiefel manifold [1].

2.3 The Set-to-Set Metric

Based on multiple manifold projections, the distance $D^\Theta(X_i, X_j)$ of two SPD matrices is transformed into the set-to-set distance $D_s(\mathbf{X}_i, \mathbf{X}_j)$. Firstly

$\{g_A^k(\cdot, \cdot)\}_{k=1}^m$ is exploited to compute sub-metrics on m low-dimensional SPD manifolds, and then $h_M(\cdot)$ is utilized to summarize the m sub-metrics, where A and M are learnable parameters. We use the flexible alpha-beta divergence [3, 4, 22] as the sub-metric $g_A^k(\cdot, \cdot)$. For two SPD sets $\mathbf{X}_i = \{X_i^k\}_{k=1}^m$, $\mathbf{X}_j = \{X_j^k\}_{k=1}^m$, the distance d_{ij}^k between X_i^k and X_j^k is computed by the k -th alpha-beta divergence,

$$d_{ij}^k = g_A^k(X_i^k, X_j^k) = D^{(\alpha_k, \beta_k)}(X_i^k \| X_j^k) = \frac{1}{\alpha_k \beta_k} \sum_{u=1}^p \log \left(\frac{\alpha_k (\lambda_{iju}^k)^{\beta_k} + \beta_k (\lambda_{iju}^k)^{-\alpha_k}}{\alpha_k + \beta_k} \right), \tag{3}$$

where λ_{iju}^k is the u -th eigenvalue of $X_i^k (X_j^k)^{-1}$, and (α_k, β_k) is the individual parameter of the k -th alpha-beta divergence. We denote all alpha-beta divergence parameters as a matrix $A = [(\alpha_1, \beta_1), (\alpha_2, \beta_2), \dots, (\alpha_m, \beta_m)] \in \mathbb{R}^{m \times 2}$, and a distance vector between \mathbf{X}_i and \mathbf{X}_j as $\mathbf{d}_{ij} = [d_{ij}^1, d_{ij}^2, \dots, d_{ij}^m] \in \mathbb{R}^{m \times 1}$. Since (α_k, β_k) needs to be adaptive to the k -th low-dimensional manifold, we exploit a learnable strategy to update (α_k, β_k) , which is detailed in the next subsection. After computing all sub-metrics, the distance metric $D^\theta(X_i, X_j)$ between two original SPD matrices X_i and X_j is formulated as

$$\begin{aligned} D^\theta(X_i, X_j) &= D_s(\mathbf{X}_i, \mathbf{X}_j) = h_M(d_{ij}^1, d_{ij}^2, \dots, d_{ij}^m) = \mathbf{d}_{ij}^\top M \mathbf{d}_{ij} \\ &= \sum_{k=1}^m \sum_{l=1}^m \left(D^{(\alpha_k, \beta_k)}(W_k^\top X_i W_k \| W_k^\top X_j W_k) \cdot M_{kl} \cdot D^{(\alpha_l, \beta_l)}(W_l^\top X_i W_l \| W_l^\top X_j W_l) \right), \end{aligned} \tag{4}$$

where $M \in \mathbb{R}^{m \times m}$ is the metric parameter, and M_{kl} is an element of M in the k -th row and l -th column, reflecting the significance and relationship of properties. If $X_i = X_j$, then \mathbf{d}_{ij} is a zero vector, and $D^\theta(X_i, X_j) = 0$. If $X_i \neq X_j$, then \mathbf{d}_{ij} is a non-zero vector, and $D^\theta(X_i, X_j)$ should be larger than 0. The nonnegativity of the metric forces M to be an SPD matrix and $M \in Sym_m^+$.

To learn the parameter θ , we formulate loss function $\mathcal{L}(\theta, \mathcal{S}, \mathcal{D}, Y)$ as

$$\begin{aligned} \min_{\theta} \mathcal{L}(\theta, \mathcal{S}, \mathcal{D}, Y) &= \frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} y_{ij} \cdot \max(D^\theta(X_i, X_j) - \zeta_s, 0)^2 \\ &\quad + \frac{1}{|\mathcal{D}|} \sum_{i,j \in \mathcal{D}} (1 - y_{ij}) \cdot \max(\zeta_d - D^\theta(X_i, X_j), 0)^2 \tag{5} \\ &\quad + \xi \cdot \gamma(M, M_0). \end{aligned}$$

We expect that the distance between similar samples is smaller than a threshold ζ_s , and the distance between dissimilar samples is larger than a threshold ζ_d . We add two coefficients $\frac{1}{|\mathcal{S}|}$ and $\frac{1}{|\mathcal{D}|}$ to solve the imbalance issue of similar and dissimilar sample pairs, where $|\mathcal{S}|$ and $|\mathcal{D}|$ are the pair numbers of sets \mathcal{S} and \mathcal{D} . In addition, we add a regularization term $\xi \cdot \gamma(M, M_0)$ on M in Eq. (5). $\gamma(M, M_0) = Tr(MM_0^{-1}) - \log \det(MM_0^{-1}) - m$ is the burgman divergence [5, 8, 10], where $Tr(\cdot)$ is the trace of a matrix, M_0 is the prior information, and ξ is the trade-off coefficient.

2.4 Optimization

$\mathcal{L}(\Theta, \mathcal{S}, \mathcal{D}, Y)$ in Eq. (5) is not a convex function with respect to W , A , and M . Accordingly, we apply the gradient descent to learn Θ . The gradients are computed as follows.

(1) The gradient of \mathcal{L} with respect to M

The gradient of \mathcal{L} with respect to M can be computed by

$$\nabla_M(\mathcal{L}) = \frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} \mathbf{d}_{ij} \nabla_{D_{ij}^\Theta}(\mathcal{L}) \mathbf{d}_{ij}^\top + \frac{1}{|\mathcal{D}|} \sum_{i,j \in \mathcal{D}} \mathbf{d}_{ij} \nabla_{D_{ij}^\Theta}(\mathcal{L}) \mathbf{d}_{ij}^\top + \xi \cdot \nabla_M(\gamma(M, M_0)), \quad (6)$$

where $\nabla_{D_{ij}^\Theta}(\mathcal{L})$ is the gradient of \mathcal{L} with respect to $D^\Theta(X_i, X_j)$,

$$\nabla_{D_{ij}^\Theta}(\mathcal{L}) = 2 \cdot y_{ij} \cdot \max(D_{ij}^\Theta - \zeta_s, 0) + 2 \cdot (y_{ij} - 1) \cdot \max(\zeta_d - D_{ij}^\Theta, 0), \quad (7)$$

and $\nabla_M(\gamma(M, M_0))$ is the gradient of $\gamma(M, M_0)$ with respect to M ,

$$\nabla_M(\gamma(M, M_0)) = M_0^{-1} - M^{-1}. \quad (8)$$

(2) The gradient of \mathcal{L} with respect to A

The gradients of \mathcal{L} with respect to α_k and β_k in A are

$$\nabla_{\alpha_k}(\mathcal{L}) = \frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} \nabla_{d_{ij}^k}(\mathcal{L}) \cdot \nabla_{\alpha_k}(d_{ij}^k) + \frac{1}{|\mathcal{D}|} \sum_{i,j \in \mathcal{D}} \nabla_{d_{ij}^k}(\mathcal{L}) \cdot \nabla_{\alpha_k}(d_{ij}^k), \quad (9)$$

$$\nabla_{\beta_k}(\mathcal{L}) = \frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} \nabla_{d_{ij}^k}(\mathcal{L}) \cdot \nabla_{\beta_k}(d_{ij}^k) + \frac{1}{|\mathcal{D}|} \sum_{i,j \in \mathcal{D}} \nabla_{d_{ij}^k}(\mathcal{L}) \cdot \nabla_{\beta_k}(d_{ij}^k). \quad (10)$$

$\nabla_{d_{ij}^k}(\mathcal{L})$ is the k -th element of $\nabla_{\mathbf{d}_{ij}}(\mathcal{L})$ which is the gradient of \mathcal{L} with respect to \mathbf{d}_{ij} ,

$$\nabla_{\mathbf{d}_{ij}}(\mathcal{L}) = \nabla_{D_{ij}^\Theta}(\mathcal{L}) \cdot \nabla_{\mathbf{d}_{ij}}(D_{ij}^\Theta) = \nabla_{D_{ij}^\Theta}(\mathcal{L}) \mathbf{d}_{ij}^\top (M^\top + M). \quad (11)$$

$\nabla_{\alpha_k}(d_{ij}^k)$ and $\nabla_{\beta_k}(d_{ij}^k)$ are the gradients of d_{ij}^k with respect to α_k and β_k , respectively,

$$\begin{aligned} \nabla_{\alpha_k}(d_{ij}^k) = & \frac{1}{\alpha_k^2 \beta_k} \sum_{u=1}^p \left(\frac{\alpha_k (\lambda_{iju}^k)^{\beta_k} - \alpha_k \beta_k (\lambda_{iju}^k)^{-\alpha_k} \log \lambda_{iju}^k}{\alpha_k (\lambda_{iju}^k)^{\beta_k} + \beta_k (\lambda_{iju}^k)^{-\alpha_k}} \right. \\ & \left. - \frac{\alpha_k}{\alpha_k + \beta_k} - \log \frac{\alpha_k (\lambda_{iju}^k)^{\beta_k} + \beta_k (\lambda_{iju}^k)^{-\alpha_k}}{\alpha_k + \beta_k} \right), \end{aligned} \quad (12)$$

$$\nabla_{\beta_k}(d_{ij}^k) = \frac{1}{\alpha_k \beta_k^2} \sum_{u=1}^p \left(\frac{\beta_k (\lambda_{iju}^k)^{-\alpha_k} - \alpha_k \beta_k (\lambda_{iju}^k)^{\beta_k} \log \lambda_{iju}^k}{\alpha_k (\lambda_{iju}^k)^{\beta_k} + \beta_k (\lambda_{iju}^k)^{-\alpha_k}} - \frac{\beta_k}{\alpha_k + \beta_k} - \log \frac{\alpha_k (\lambda_{iju}^k)^{\beta_k} + \beta_k (\lambda_{iju}^k)^{-\alpha_k}}{\alpha_k + \beta_k} \right). \quad (13)$$

(3) The gradient of \mathcal{L} with respect to W

The gradient of \mathcal{L} with respect to each W_k is

$$\nabla_{W_k}(\mathcal{L}) = \sum_i^N ((X_i)^\top W_k \nabla_{X_i^k}(\mathcal{L}) + X_i W_k \nabla_{X_i^k}(\mathcal{L})^\top), \quad (14)$$

where N is the number of training samples, and $N = 2 \times (|\mathcal{S}| + |\mathcal{D}|)$. $\nabla_{X_i^k}(\mathcal{L})$ is the gradient of \mathcal{L} with respect to the low-dimensional SPD matrix X_i^k . The eigenvalue decomposition of $X_i^k (X_j^k)^{-1}$ is $X_i^k (X_j^k)^{-1} = U_{ij}^k \Sigma_{ij}^k (U_{ij}^k)^\top$. Σ_{ij}^k is the diagonal matrix eigenvalues, and λ_{iju}^k is the u -th eigenvalue. The gradients $\nabla_{X_i^k}(\mathcal{L})$ and $\nabla_{X_j^k}(\mathcal{L})$ are

$$\nabla_{X_i^k}(\mathcal{L}) = U_{ij}^k \nabla_{\Sigma_{ij}^k}(\mathcal{L}) (U_{ij}^k)^\top (X_i^k)^{-\top}, \quad (15)$$

$$\nabla_{X_j^k}(\mathcal{L}) = (-1) \cdot (X_j^k)^{-\top} (X_i^k)^\top U_{ij}^k \nabla_{\Sigma_{ij}^k}(\mathcal{L}) (U_{ij}^k)^\top (X_j^k)^{-\top}, \quad (16)$$

where $\nabla_{\Sigma_{ij}^k}(\mathcal{L})$ is the gradient of Σ_{ij}^k with respect to \mathcal{L} . $\nabla_{\Sigma_{ij}^k}(\mathcal{L})$ is a diagonal matrix, and the u -th element is

$$\begin{aligned} \nabla_{\lambda_{iju}^k}(\mathcal{L}) &= \nabla_{d_{ij}^k}(\mathcal{L}) \cdot \nabla_{\lambda_{iju}^k}(d_{ij}^k) \\ &= \nabla_{d_{ij}^k}(\mathcal{L}) \cdot \frac{1}{\alpha_k \beta_k} \frac{\alpha_k \beta_k (\lambda_{iju}^k)^{\beta_k - 1} - \alpha_k \beta_k (\lambda_{iju}^k)^{-\alpha_k - 1}}{\alpha_k (\lambda_{iju}^k)^{\beta_k} + \beta_k (\lambda_{iju}^k)^{-\alpha_k}}. \end{aligned} \quad (17)$$

Since the gradients $\nabla_W(\mathcal{L})$, $\nabla_M(\mathcal{L})$, and $\nabla_A(\mathcal{L})$ are obtained, the metric parameter set Θ can be updated. A is optimized by the standard gradient descent, $A := A - \eta \nabla_A(\mathcal{L})$, where η is the learning rate. W and M are updated by the Riemannian optimization algorithm [1, 6, 20]. The computation details are presented below,

$$\begin{cases} \nabla_{W_R}(\mathcal{L}) = \nabla_W(\mathcal{L}) - W \frac{1}{2} (W^\top \nabla_W(\mathcal{L}) + \nabla_W(\mathcal{L})^\top W) \\ W := q(W - \eta \nabla_{W_R}(\mathcal{L})) \end{cases}, \quad (18)$$

and

$$\begin{cases} \nabla_{M_R}(\mathcal{L}) = M \frac{1}{2} (\nabla_M(\mathcal{L}) + \nabla_M(\mathcal{L})^\top) M \\ M := M^{\frac{1}{2}} \expm(-\eta M^{-\frac{1}{2}} \nabla_{M_R}(\mathcal{L}) M^{-\frac{1}{2}}) M^{\frac{1}{2}}, \end{cases} \quad (19)$$

where $\nabla_{W_R}(\mathcal{L})$ and $\nabla_{M_R}(\mathcal{L})$ are the Riemannian gradients with respect to W and M . In Eq. (18), $q(\cdot)$ is the retraction operation mapping the data back to the Stiefel manifold. $q(W)$ denotes the Q matrix of the QR decomposition to a matrix W , *i.e.*, for the matrix $W \in \mathbb{R}^{n \times p}$, $W = QR$, where $Q \in \mathbb{R}^{n \times p}$ is a semi-orthogonal matrix and $R \in \mathbb{R}^{p \times p}$ is a upper triangular matrix. In Eq. (19), $\expm(\cdot)$ is the matrix exponential function. We summarize the learning process of our method in Algorithm 1, w.

Algorithm 1. Training Process of Our Method

Input: Training SPD sample pairs \mathcal{S} and \mathcal{D} , labels Y . The initial projection matrix W . The initial metric matrix M . The initial alpha-beta divergence parameter A . Learning rate η .

Output: The learned W , M , and A .

- 1: **while** not converge **do**
- 2: For each SPD matrix, compute subspaces by Eq.(2).
- 3: For each sample pairs, compute the distance between their sets by Eq.(3) and Eq.(4).
- 4: Compute the loss \mathcal{L} by Eq.(5).
- 5: Compute the gradient $\nabla_M(\mathcal{L})$ by Eq.(7), Eq.(8), and Eq.(6).
- 6: Compute the gradient $\nabla_A(\mathcal{L})$ by Eq.(12), Eq.(13), Eq.(9), and Eq.(10).
- 7: Compute the gradient $\nabla_W(\mathcal{L})$ by Eq.(17), Eq.(15), Eq.(16), and Eq.(14).
- 8: Update the parameter W by Eq.(18).
- 9: Update the parameter A by $A := A - \eta \nabla_A(\mathcal{L})$.
- 10: Update the parameter M by Eq.(19).
- 11: **end while**
- 12: **return** W , M and A

3 Experiments

In order to test the efficiency of our method, we conduct experiments on the object recognition, video-based face recognition, action recognition, and texture classification tasks. Four datasets are utilized: the ETH-80 [15], the MSR-Action3D [17], the YouTube Celebrities (YTC) [13], and the UIUC [18] datasets.

3.1 Datasets and Settings

The ETH-80 is an object image dataset, which contains 80 image sets of eight categories. Each category consists of 10 image sets, and each set includes 41 images captured under different views. In our experiment, all the images of the ETH-80 are resized to 20×20 and denoted by the intensity features. The YTC is a video-based face dataset, collecting 1910 videos of 47 persons. Face regions are detected from each frame by a cascaded face detector and resized to 30×30 , followed by the histogram equalized operation, and represented by

the gray values. The MSR-Action3D is a 3D action dataset, containing totally 567 videos of 20 actions. There are 20 skeleton joints in the body of actions. In the experiments, each frame is represented by a 120-dimensional feature, which is the 3D coordinate differences of skeleton joints between this frame and its two neighborhood frames. The UIUC material dataset contains 216 samples of 18 categories. We resize each image to 400×400 . Then 128-dimensional dense SIFT features are extracted from each image with 4-pixel space concatenated by 27-dimensional RGB color features from 3×3 patches centered at the locations of dense SIFT features.

On the ETH-80, YTC, and UIUC datasets, we compute a covariance matrix C to represent each sample and add a small ridge δI to avoid the matrix singularity, where $\delta = 0.001 \times \text{Tr}(C)$. On the MSR-Action3D dataset, we first compute the covariance matrix C with size of 120×120 , then transform it to a 121×121 Gaussian distribution SPD matrix, $C = |C|^{-\frac{1}{121}} \begin{bmatrix} C + \frac{mm^T}{m} & m \\ m^T & 1 \end{bmatrix}$ as the sample representation, where \mathbf{m} is the mean vector of 120-dimensional features. Following the standard protocols [7, 11, 24, 29], for each category, we randomly select half of the samples for training and the rest for testing on the ETH-80, MSR-Action3D, and UIUC datasets. On the YTC dataset, for each person, three videos are randomly selected as the gallery, and six as the probe. In experiments, we set $\xi = 0.01$, $M_0 = I_m$, $\zeta_s = 5$, and $\zeta_d = 100$.

3.2 Evaluation

We exploit the 1-NN classifier to evaluate the performance of all metric learning methods. The following methods are evaluated in our experiments: AIM [19], Stein Divergence [21], LEM [2], SPD-DR [7], CDL [29], RSR-ML [9], LEML [11], and α -CML [31]. AIM, Stein Divergence, and LEM are the basic SPD matrix metrics, measuring the geodesic distance between SPD matrices. SPD-DR implements the dimensionality reduction on the SPD matrix and then applies the AIM or Stein Divergence between samples. CDL is a Riemannian kernel discriminative learning approach on the SPD manifold. RSR-ML employs sparse coding and dictionary learning scheme on the SPD manifold. LEML and α -CML are two LEM based SPD matrix metric learning methods which project SPD matrices to the tangent space and utilize the LEM to compute the distance between them.

Table 1 shows the comparisons of the four visual tasks. In the object recognition task, we set the dimensionality of the low-dimensional manifolds is 10×10 and the number of them is 20, *i.e.*, $m = 20$. We find that LEM has a better performance than AIM, 93.0 vs 85.0, showing that the point on the tangent space is more discriminative. If the manifold point is projected to a low-dimensional discriminative space, *i.e.*, the SPD-DR method, the performance can be improved to 96.0, 0.5 better than LEML. Compared with SPD-DR, our method achieves 97.5, 1.5 higher than it, which shows the power of discovering discriminative properties and their roles.

In the video-based face recognition task, the dimensionality of projected manifolds is 10×10 , and the number of them is 40. We achieve 49.2 in this task, 2.5

higher than SPD-DR and 10 percent higher than the basic SPD matrix metrics approximately. However, due to the large variable faces caused by posture, illumination, scale, and occlusion, the performance of linear metric learning methods is far less than it of the nonlinear kernel method CDL. The reason we think is that the samples in the original space are not separable, a more higher-dimensional RKHS space can relieve this problem.

In the action recognition task, the dimensionality of the low-dimensional manifolds is 8×8 and the number of them is 15. Nonlinear kernel methods CDL and RSR-ML achieve 95.4 and 95.0 respectively and have a better performance than the existing metric methods [7, 11, 31]. In this case, our linear method obtains the comparable performance with CDL and RSR-ML, achieving 95.8. Besides, Wang *et al.* [26] shows that the nonlinear kernel matrix representation has a better performance than the linear SPD representation, while our accuracy is 3.1 higher than α -CML whose performance is based on the kernel matrix [26] rather than the Gaussian distribution SPD matrix.

In the texture classification task, in our method, we set the dimensionality of the low-dimensional manifolds is 8×8 , and there are totally 18 low-dimensional manifolds. We can see that, the three basic SPD matrix metrics *i.e.*, AIM, Stein Divergence, and LEM achieve comparable performance in the UIUC dataset, 35.6, 35.8 and 36.7 respectively. Meanwhile, metric learning methods can bring a remarkable improvement. CDL achieves 54.9, and the accuracy of LEML is 53.9. SPD-DR achieves a better performance 58.3, showing that there are too much noise and information redundancy in the original SPD representation. Our method further improves the result to 60.8 showing that our method can not only remove the noise and information redundancy but also bring the benefits of discovering discriminative intrinsic properties and their different roles.

Table 1. Accuracies (%) on the four visual tasks. Our method is bold in the last line.

Method	Eth-80	YTC	MSR-Action3D	UIUC
AIM [19]	85.0	38.2	84.7	35.6
Stein [21]	-	-	83.5	35.8
LEM [2]	93.0	40.8	84.7	36.7
AIM-DR [7]	96.0	46.7	93.1	58.3
Stein-DR [7]	-	-	94.6	58.1
CDL [29]	94.5	67.5	95.4	54.9
RSR-ML [9]	94.8	-	95.0	-
LEML [11]	95.5	-	92.3	53.9
α -CML [31]	-	-	92.7	-
Ours	97.5	49.2	95.8	60.8

4 Conclusions

In this paper, we have proposed a novel metric learning method on the SPD manifold, which can discover discriminative intrinsic properties and computes the metric considering their different roles. We can formulate the SPD manifold metric learning process as the multiple projections and a set-to-set metric joint optimization problem. Moreover, we force the projection matrix and the metric matrix on manifolds, obtaining a robust metric. Extensive experiments have shown that our method outperforms existing metric learning methods on the SPD manifold. As our method is differentiable in the whole process, in the future, we will endow it with deep learning for the desirable nonlinearity.

Acknowledgements. This work was supported by the Natural Science Foundation of China (NSFC) under Grants No. 61702037 and No. 61773062, and Beijing Municipal Natural Science Foundation under Grant No. L172027, in part by Beijing Institute of Technology Research Fund Program for Young Scholars.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2009)
2. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Reson. Med.* **56**(2), 411 (2006)
3. Cherian, A., Stanitsas, P., Harandi, M., Morellas, V., Papanikolopoulos, N.: Learning discriminative $\alpha\beta$ -divergences for positive definite matrices. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4270–4279 (2017)
4. Cichocki, A., Cruces, S., Amari, S.: Log-determinant divergences revisited: alpha-beta and gamma log-det divergences. *Entropy* **17**(5), 2988–3034 (2015)
5. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 209–216 (2007)
6. Harandi, M., Fernando, B.: Generalized backpropagation, Étude de cas: Orthogonality. arXiv preprint [arXiv:1611.05927](https://arxiv.org/abs/1611.05927) (2016)
7. Harandi, M., Salzmann, M., Hartley, R.: Dimensionality reduction on SPD manifolds: the emergence of geometry-aware methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 48–62 (2017)
8. Harandi, M., Salzmann, M., Hartley, R.: Joint dimensionality reduction and metric learning: a geometric take. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 1404–1413 (2017)
9. Harandi, M.T., Sanderson, C., Hartley, R., Lovell, B.C.: Sparse coding and dictionary learning for symmetric positive definite matrices: a kernel approach. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 216–229. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_16
10. Hoffman, J., Rodner, E., Donahue, J., Kulis, B., Saenko, K.: Asymmetric and category invariant feature transformations for domain adaptation. *Int. J. Comput. Vis.* **109**(1–2), 28–41 (2014)

11. Huang, Z., Wang, R., Shan, S., Li, X., Chen, X.: Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 720–729 (2015)
12. Huang, Z., Wang, R., Van Gool, L., Chen, X., et al.: Cross Euclidean-to-Riemannian metric learning with application to face recognition from video. *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**(99), 1 (2018)
13. Kim, M., Kumar, S., Pavlovic, V., Rowley, H.: Face tracking and recognition with visual constraints in real-world videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
14. Kulis, B., Sustik, M.A., Dhillon, I.S.: Low-rank kernel learning with Bregman matrix divergences. *J. Mach. Learn. Res.* **10**(1), 341–376 (2009)
15. Leibe, B., Schiele, B.: Analyzing appearance and contour based methods for object categorization. In: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. II-409 (2003)
16. Li, P., Xie, J., Wang, Q., Zuo, W.: Is second-order information helpful for large-scale visual recognition? In: IEEE International Conference on Computer Vision, pp. 2089–2097 (2017)
17. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 9–14 (2010)
18. Liao, Z., Rock, J., Wang, Y., Forsyth, D.: Non-parametric filtering for geometric detail extraction and material representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 963–970 (2013)
19. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66**(1), 41–66 (2006)
20. Roy, Kumar, S., Mhammedi, Z., Harandi, M.: Geometry aware constrained optimization techniques for deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1 (2018)
21. Sra, S.: A new metric on the manifold of kernel matrices with application to matrix geometric means. In: Advances in Neural Information Processing Systems, pp. 144–152 (2012)
22. Thiyam, D.B., Cruces, S., Olias, J., Cichocki, A.: Optimization of Alpha-Beta Log-Det divergences and their application in the spatial filtering of two class motor imagery movements. *Entropy* **19**(3), 89 (2017)
23. Vemulapalli, R., Jacobs, D.W.: Riemannian metric learning for symmetric positive definite matrices. arXiv preprint [arXiv:1501.02393](https://arxiv.org/abs/1501.02393) (2015)
24. Vemulapalli, R., Pillai, J.K., Chellappa, R.: Kernel learning for extrinsic classification of manifold features. In: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1782–1789 (2013)
25. Wang, H., Wang, Q., Gao, M., Li, P., Zuo, W.: Multi-scale location-aware kernel representation for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1 (2018)
26. Wang, L., Zhang, J., Zhou, L., Tang, C., Li, W.: Beyond covariance: feature representation with nonlinear kernel matrices. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4570–4578 (2015)
27. Wang, Q., Li, P., Zhang, L.: G2DeNet: global Gaussian distribution embedding network and its application to visual recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6507–6516 (2017)

28. Wang, Q., Li, P., Zuo, W., Zhang, L.: RAID-G: robust estimation of approximate infinite dimensional Gaussian with application to material recognition. In: Computer Vision and Pattern Recognition, pp. 4433–4441 (2016)
29. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: a natural and efficient approach to image set classification. In: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2496–2503 (2012)
30. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Li, Y.: Deep manifold-to-manifold transforming network for action recognition. arXiv preprint [arXiv:1705.10732](https://arxiv.org/abs/1705.10732) (2017)
31. Zhou, L., Wang, L., Zhang, J., Shi, Y., Gao, Y.: Revisiting metric learning for SPD matrix based visual representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3241–3249 (2017)



Structure Fusion and Propagation for Zero-Shot Learning

Guangfeng Lin^(✉), Yajun Chen, and Fan Zhao

Xi'an University of Technology, Xi'an 710048, Shaanxi Province,
People's Republic of China
{lgf78103, chenyajun, vcu}@xaut.edu.cn

Abstract. The key of zero-shot learning (ZSL) is how to find the information transfer model for bridging the gap between images and semantic information (texts or attributes). Existing ZSL methods usually construct the compatibility function between images and class labels with consideration of the relevance on the semantic classes (the manifold structure of semantic classes). However, the relationship of image classes (the manifold structure of image classes) is also very important for the compatibility model construction. It is difficult to capture the relationship among image classes due to unseen classes, so that the manifold structure of image classes often is ignored in ZSL. To complement each other between the manifold structure of image classes and that of semantic classes information, we propose structure fusion and propagation (SFP) for improving the performance of ZSL for classification. SFP can jointly consider the manifold structure of image classes and that of semantic classes for approximating to the intrinsic structure of object classes. Moreover, the SFP can describe the constraint condition between the compatibility function and these manifold structures for balancing the influence of the structure fusion and propagation iteration. The SFP solution provides not only unseen class labels but also the relationship of two manifold structures that encodes the positive transfer in structure fusion and propagation. Experiments demonstrate that SFP can attain the promising results on the AwA, CUB, Dogs and SUN datasets.

Keywords: Structure fusion and propagation · Manifold structure
Zero-shot learning · Transfer learning

1 Introduction

Although deep learning [32] depending on large-scale labeled data training has been generally used for visual recognition [31], a daunting challenge still exists to recognize visual object “in the wild”. In fact, in specific applications it is

Supported by NSFC (Program No. 61771386, Program No. 61671376 and Program No. 61671374), Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2016JM6045, Program No. 2017JZ020).

impossible to collect all class data for training deep model, so training (seen classes) and testing classes (unseen classes) are often disjoint. The main idea of ZSL is to handle this problem by exploiting the transfer model from the redundant relevance of the semantic description. To recognize unseen classes from seen classes, ZSL needs face to two challenges [3]. One is how to utilize the semantic information for constructing the relationship between unseen classes and seen classes, and other is how to find the compatibility among all kinds of information for obtaining the optimal discriminative characteristics on unseen classes.

ZSL can bridge the gap among the different domains to recognize unseen class objects by semantic embedding of class labels. These semantic embeddings can come from vision (attributes [11]) and language information (text [25]) by the manual annotation, machine learning [29] or data mining [5]. In term of the transformation relationship of different embedding, recent ZSL methods mainly fall into linear embedding, nonlinear embedding and similarity embedding. Linear embedding [1, 2, 7, 13, 24] implements the linear transformation method among different embedding spaces for learning the relevance between unseen class objects and class labels. Nonlinear embedding [23, 25, 28] can realize the nonlinear mapping of the embedding space for building the compatibility function or classifier, which can be learned by deep networks [14, 30]. Similarity embedding [3, 9, 15, 19, 33] builds the classifier by the similarity metrics, which mostly include structure learning or class-wise similarities. In our approach, the similarity metric is extended from semantic space to image space, we attempt to find the relationship of similarities (manifold structure in the different space) for constraining the compatibility function, and further capture to the positive structure propagation for the significantly improvement of the unseen object classification.

In this paper, our motivation is inspired by structure fusion [16–18] for jointly dealing with two challenges. The intrinsic manifold structure is crucial for object classification. However, in fact, we only can attain the observation data of the manifold structure, which can represent different aspects of the intrinsic manifold structure. For recovering or approximating the intrinsic structure, we can fuse various manifold structures from observation data. Based on the above idea, we try to capture different manifold structures in image and semantic space for improving the recognition performance of unseen classes in ZSL. Therefore, we expect to construct the compatibility function for predicting labels of unseen classes by building the manifold structure of image classes. On the other end, we attempt to find the relevance between the manifold structure of semantic classes and that of image classes in model space for encoding the influence between the negative and positive transfer, and further make the better compatibility function for classifying unseen class objects. Model space corresponding to visual appearances is the jointed projection space of semantic space and image space, and can preserve the respective manifold structure. Figure 1 illustrates the idea of the proposed method conceptually. SFP considers not only semantic and image structures but also the positive structure propagation for ameliorating unseen

objects classification, while SynC [3] only focus on manifold structure in semantic space for combining the base classifier in ZSL.

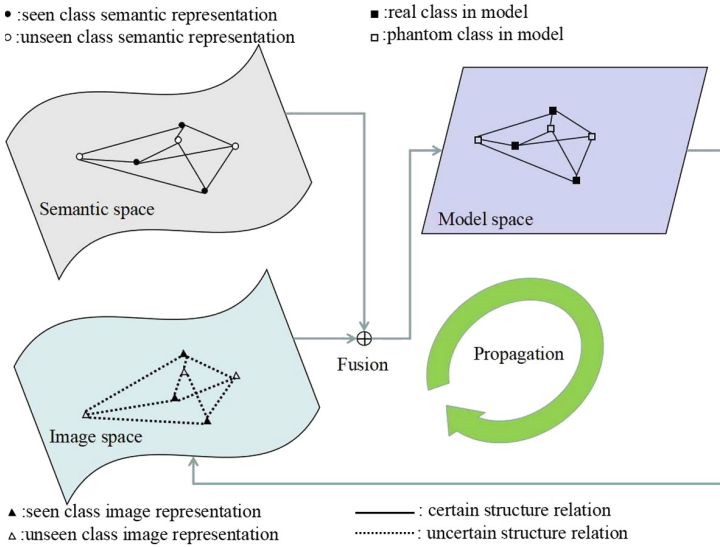


Fig. 1. The illustration of structure fusion and propagation for zero-shot learning. Phantom object classes (the coordinates of classes in the model space are optimized to achieve the best performance of the resulting model for the real object classes in discriminative tasks [3].) and real object classes corresponding to all classes in model space.

In our main contribution, a novel idea have tow aspects to recover or approximate the intrinsic manifold structure from seen classes to unseen classes by fusing the different space manifold structure for handling the challenging unseen classes recognition. Specifically, one constructs the projected manifold structure for real and phantom class in model space, another constrains the compatibility function and the relationship of the manifold structure for the positive structure propagation.

2 Structure Fusion and Propagation

In ZSL, we have training data set $\mathcal{D} = \{(x_n \in R^D, y_n)\}_{n=1}^N$, in which x_n is image representation (it can be extracted based on deep model, and the detail is described in Table 1) and $y_n (n = 1, \dots, N)$ is the class label in the seen class set $\mathcal{S} = \{s|s = 1, \dots, S\}$. We can denote the unseen class set as $\mathcal{U} = \{u|u = S+1, \dots, S+U\}$. $a_c \in R_D$ is the linear transformation vector of the $c \in \{\mathcal{S} \cup \mathcal{U}\}$ class.

2.1 Classification Model and Manifold Structure

We construct a pair-wise linear classifier [3] in the visual image feature space, and determinate a estimated label \hat{y} to a feature x by the following formula.

$$\hat{y} = \arg \max_c a_c^T x, \tag{1}$$

here, $a_c \in R^D$ is not only the transformation vector of the feature x , but also the representation of the class c in model. In other words, the above formula can describe the pair-wise linear relation between the feature space and the class label space for characterizing the class representation in the model.

To measure the manifold structure, we can compute the similarity of the related representation in the homogeneous space, which has the same scale and metric. To this end, we respectively build a bipartite graph between unseen classes and seen classes in semantic space and image space (this space includes all image representations). In these bipartite graphs, nodes are corresponding to unseen classes or seen classes, and weights of these nodes connect unseen classes with seen classes. Because we focus on the transfer relation between unseen classes and seen classes, no connection exists in unseen classes or seen classes. Supposing $G_b < V_b, E_b >$ can denote the manifold structure of semantic classes. Here, $V_b = V_{bs} \cup V_{bu}$ and $\emptyset = V_{bs} \cap V_{bu}$. E_b includes connections between V_{bs} (seen classes set in semantic space) and V_{bu} (unseen classes set in semantic space); $G_x < V_x, E_x >$ for the manifold structure of image classes. Here, $V_x = V_{xs} \cup V_{xu}$ and $\emptyset = V_{xs} \cap V_{xu}$. E_x includes the connections between V_{xs} (seen classes set in image space) and V_{xu} (unseen classes set in image space). Therefore, the similarity of semantic and image space is respectively regarded as the weight between nodes, which can be defined as following.

$$w_{su}^{(b)} = \frac{\exp(-d(b_s, b_u))}{\sum_{u=1}^U \exp(-d(b_s, b_u))}, w_{su}^{(x)} = \frac{\exp(-d(x_s, x_u))}{\sum_{u=1}^U \exp(-d(x_s, x_u))}, \tag{2}$$

here, b_s and x_s are respectively the semantic and image representation (the detail is described in Table 1) of the seen class s , while b_u and x_u are respectively the semantic and image representation of the unseen class u . $w_{su}^{(b)}$ and $w_{su}^{(x)}$ are respectively the weight (the similarity) between the seen class s and the unseen class u in semantic and image representation space. $d(b_s, b_u)$ and $d(x_s, x_u)$ are respectively the distance metric [3] of each space, and can be defined as following.

$$d(b_s, b_u) = (b_s - b_u)^T \Sigma_b^{-1} (b_s - b_u), d(x_s, x_u) = (x_s - x_u)^T \Sigma_x^{-1} (x_s - x_u), \tag{3}$$

here, $\Sigma_b = \sigma_b I$ can be learned from the semantic representation by cross-validation (We alternately divide the training classes set into two part in according with the proportion between the training classes set and the test classes set. One part is to learn the model, and another is to validate the model. We give the range of σ_b , which is form 2^{-5} to 2^5 , and select the parameter corresponding to the best result as the value of σ_b .) $\Sigma_x = \sigma_x I$ can be learned from the image representation by cross-validation (It is the same procedure like σ_b learning.).

In image space, the differentiation compared with the semantic space is that x_u is not determined because of unseen classes, while x_s can be obtained from training data by computing the mean value of the seen class. The way to produce the center of the class as a representation is simple for convenient computation, and it is reasonable to preserve the base characteristic of image representation according with the distribution of the same class. x_u can be attained by pre-classification of unseen classes (the detail in the next section).

In (1), a_c is the transformation vector, and also is the class representation in model space. In (2), b_s and b_u is the class representation in semantic space, while x_s and x_u is the class representation in image space. We expect to construct the link among these space by v_s and v_u , which are respectively the phantom class of seen or unseen classes in model. For preserving the manifold structure of two bipartite graphs and aligning the image, the semantic and the model space, we build the optimization formula under the condition of the distortion error minimization, which is defined as following.

$$\begin{aligned}
 (a_c, v_u, \beta) = \arg \min_{a_c, v_u, \beta} & \|a_c - \sum_{u=1}^U \beta^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(b)} \end{bmatrix}^T v_u - \sum_{s=1}^S \gamma^T \begin{bmatrix} w_{ss}^{(x)} & w_{ss}^{(b)} \end{bmatrix}^T v_s\|_2^2, \\
 \text{s.t. } & \beta^T \mathbf{1} = 1, \gamma^T \mathbf{1} = 1, 0 \leq \beta_i \leq 1, 0 \leq \gamma_i \leq 1 \quad (i = 1, 2)
 \end{aligned} \tag{4}$$

here, $\beta = [\beta_1 \ \beta_2]^T$, $\gamma = [\gamma_1 \ \gamma_2]^T$, and $\mathbf{1} = [1 \ 1]^T$. Because no connection exists between unseen classes or seen classes in tow bipartite graphs, $w_{ss}^{(b)} = 0$ and $w_{ss}^{(x)} = 0$. The analytical solution of (4) can find the relation between a_c and v_u .

$$\begin{aligned}
 a_c = \sum_{u=1}^U & \beta^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(b)} \end{bmatrix}^T v_u, \\
 \text{s.t. } & \beta^T \mathbf{1} = 1, 0 \leq \beta_i \leq 1 \quad (i = 1, 2)
 \end{aligned} \tag{5}$$

here, $\forall c \in \{1, 2, \dots, S + U\}$.

2.2 Phantom Classes and Structure Relation Learning

For obtaining phantom class $v_u (u = 1, \dots, U)$ and the manifold structure of the weight coefficient vector β , we further reformulate the optimization formula for one-versus-other classifier [3].

$$\begin{aligned}
(v_1, \dots, v_U, \beta) = \arg \min_{v_1, \dots, v_U, \beta} & \sum_{c=1}^S \sum_{n=1}^N \ell(x_n, \mathbb{I}_{y_n, c}, a_c) \\
& + \frac{\lambda}{2} \sum_{c=1}^S \|a_c\|_2^2 + \frac{\gamma}{2} \|\beta_1 W^x - \beta_2 W^b\|_2^2, \\
s.t. \quad a_c = & \sum_{u=1}^U \beta^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(b)} \end{bmatrix}^T v_u, \\
\beta^T \mathbf{1} = & 1, 0 \leq \beta_i \leq 1 \quad (i = 1, 2)
\end{aligned} \tag{6}$$

here, $w_{su}^{(x)}$ is the element of the matrix W^x , and $w_{su}^{(b)}$ is the element of the matrix W^b . The first term of formula (6) is the squared hinge loss, which can be defined as $\ell(x_n, \mathbb{I}_{y_n, c}, a_c) = \max(0, 1 - \mathbb{I}_{y_n, c} a_c x_n)$. $\mathbb{I}_{y_n, c} \in \{-1, 1\}$ determines whether or not $y_n = c$. The second term of formula (6) is a_c of a regularization term, which avoids over-fitting problem on the pair-wise linear classifier for modeling the relationship between the class label and the image representation. The third term of formula (6) is the constraint of the manifold structure similarity for preventing the negative structure propagation in image space. The alternating optimization can be implemented for minimizing the formula (6) with respect to $\{v_u\}_{u=1}^U$ and β by solving the quadratic programming problem.

To depict the whole process of the structure fusion and propagation mechanism, we show the pseudo code of the proposed SFP algorithm in Algorithm 1.

Algorithm 1. The pseudo code of the SFP algorithm

Input: $\mathcal{D} = \{(x_n \in R^D, y_n)\}_{n=1}^N, b_s$ and b_u (input data)

Output: y_P^* (P is the total iteration number)

- 1: Computes the similarity matrix $W_{(b)}$ on the semantic representation by (2)
 - 2: Setting the similarity matrix $W_{(x)}$ to zero matrix on the image representation
 - 3: **for** $1 < t < P$ **do**
 - 4: Solving $\{v_u\}_{u=1}^U$ and β by alternately optimizing (6)
 - 5: Computing a_c according to (5)
 - 6: Computing \hat{y} by (1) and obtaining the class label y_t^* of the unseen class corresponding to the semantic class
 - 7: Computing the mean value of each image class as the image class representation x_s and x_u
 - 8: Computing and updating the similarity matrix $W_{(x)}$ on the image representation by (2)
 - 9: **end for**
-

2.3 Complexity Analysis

Formula (6) can be solved by alternately quadratic programming, which of the complexity includes two parts. In the first part, when β is fixed, formula (6) is

related to $\{v_u\}_{u=1}^U$ of a quadratic programming problem, which of the complexity is $O(U^3)$ for the worst. In the second part, while $\{v_u\}_{u=1}^U$ is fixed, formula (6) is corresponding to β of a quadratic programming problem, which of the complexity is $O(k^3)$ (k is the dimension of β) for the worst. Given the proposed algorithm SFP needs P iterations, it's complexity is $O(PU^3 + Pk^3)$.

3 Experiment

3.1 Datasets

For evaluating the proposed algorithm SFP¹, we carry out the experiment in four challenging datasets, which are Animals with Attributes (AwA) [12], CUB-200-2011 Birds (CUB) [27], Stanford Dogs (Dogs) [4], and SUN Attribute (SUN) [21]. These datasets can be used for fine-grained recognition (CUB and Dogs) or non-fine-grained recognition (AwA and SUN) in ZSL. In semantic space, AwA and CUB respectively are described by att [6], w2v [20], glo [22] and hie [1], while Dogs is represented by w2v [20], glo [22] and hie [1]. SUN is only depicted by att [6]. Table 1 provides the statistics and the extracted features for these datasets. In addition, for conveniently comparing with the state-of-art methods, we adopt image feature provided by [1].

Table 1. Datasets statistics and the extracted feature in experiments.

Datasets	Number of seen classes	Number of unseen classes	Total number of images	Semantic feature/dimension	Image feature/dimension
AwA	40	10	30473	att/85, w2v/400, glo/400, hie/about 200	Deep feature based on GoogleNet [26]/1024
CUB	150	50	11786	att/312, w2v/400, glo/400, hie/about 200	Deep feature based on GoogleNet [26]/1024
Dogs	85	28	19499	N/A, w2v/400, glo/400, hie/about 200	Deep feature based on GoogleNet [26]/1024
SUN	645	72	14340	att/102, N/A, N/A, N/A	Deep feature based on GoogleNet [26]/1024

¹ Source code: <https://github.com/lgf78103/Structure-propagation-for-zero-shot-learning>.

3.2 Comparison with the Baseline Methods

In this paper, there are three methods as the baseline for comparing with the proposed SFP method because of the semantic structure mining. The first method is structured joint embedding (SJE) [1], which can build the bilinear compatibility function with consideration of the structured output space for predicting the label of the unseen class. The second method is latent embedding model (LatEm) [28], which can construct the pair-wise bilinear (nonlinear) compatibility function according to model number selection for recognizing unseen classes. The third method is synthesized classifiers (SynC) [3], which can make nonlinear compatibility function with manifold structure in semantic space for combining the base classifier in ZSL. Table 2 shows the performance of the structure fusion and propagation (the proposed SFP method) greatly outperforms that of other three methods.

3.3 Classification and Validation Protocols

Classification accuracy is average value of all test class accuracy in each database. Because the learned model involves four parameters, which are $\lambda, \gamma, \sigma_b$ and σ_x (respectively are in formula (3) in formula (6)). We alternately divide the training classes set into two part in according with the proportion between the training classes set and the test classes set. One part is to learn the model, and another is to validate the model. Firstly, we set σ_b and σ_x to 1, and obtain γ and λ corresponding to the best result in γ (form 2^{-24} to 2^{-9}) and λ (form 2^{-24} to 2^{-9}) by cross validation. Secondly, we learn σ_b and σ_x corresponding to the best result in σ_b and σ_x (form 2^{-5} to 2^5) by cross validation.

3.4 Structure Fusion and Propagation with the Iteration

The main idea of the proposed SFP method shows three contents. In the first content, the manifold structure of images is considered for constructing the compatibility function between the class label and the visual feature. In the second content, the relationship between multi-manifold structures is found for booting the influence of the positive structure. In the last content, it is the most important to propagate the positive structure and fuse multi-manifold structures by the iteration computation. Therefore, we carry out the related experiment for evaluating the effect of the iteration on the structure evolution in AWA. The recognition accuracy can show the approximation degree of the class manifold structure. In other word, the better recognition accuracy is proportional to the more similar relationship between the reconstruction manifold structure and the intrinsic manifold structure of classes. Figure 2 demonstrates the recognition accuracy change with the iteration. In the beginning, the recognition accuracy rapidly increases with the iteration, and then reaches a stable state. It means that structure fusion and propagation with the iteration can advance the recognition accuracy and finally obtain the best state.

Table 2. Comparison of SFP method with SJE [1], LatEm [28] and SynC [3] in each semantic space, average per-class Top-1 accuracy (%) of unseen classes is reported based on the same data configurations, same images and semantic features in AwA. w: the fusion includes att, w2v, glo and hie, while w/o: the fusion contains w2v, glo and hie.

Datasets	Semantic feature	SJE	LatEm	SynC	SFP
AwA	att	66.7	71.9	69.3	84.3
	w2v	51.2	61.1	52.9	77.4
	glo	58.8	62.9	53.4	70.5
	hie	51.2	57.5	52.0	62.1
	w	73.9	76.1	78.0	85.4
	w/o	60.1	66.2	69.1	81.4
CUB	att	50.1	45.5	47.5	51.8
	w2v	28.4	31.8	32.3	32.5
	glo	24.2	32.5	32.8	33.3
	hie	20.6	24.2	22.7	24.3
	w	51.7	47.4	48.8	54.1
	w/o	29.9	34.9	35.2	35.3
Dogs	att	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
	w2v	19.6	22.6	27.6	33.3
	glo	17.8	20.9	21.9	33.4
	hie	24.3	25.2	31.1	32.4
	w	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
	w/o	35.1	36.3	36.3	48.1
SUN	att	56.1	57.6	62.8	67.6

3.5 Comparison with State-of-the-Arts

In term of the image data utilization of unseen classes in testing, we can divide ZSL methods into two categories, which are inductive ZSL and transductive ZSL. Inductive ZSL methods can serially process unseen samples without the consideration of the underlying manifold structure in unseen samples [1, 3, 28, 33], while transductive ZSL can usually use the manifold structure of unseen samples to improve ZSL performance [8, 10, 15]. SFP can find the structure of unseen classes in image feature space to enhance the transfer model between seen and unseen classes, so SFP belongs to a transductive ZSL method. For a fair comparison, we use deep feature of images based on GoogleNet [26] in contrasting methods, which include our method, one transductive ZSL method (DMaP [15]), and three inductive ZSL methods (SJE [1], LatEm [28] and SynC [3]). To the best of our knowledge, these methods are state-of-the-art methods for ZSL. Table 3 shows their results for ZSL on three benchmark datasets. SFP mostly outperforms the state-of-the-art methods except DMaP on CUB. DMaP focuses on the manifold

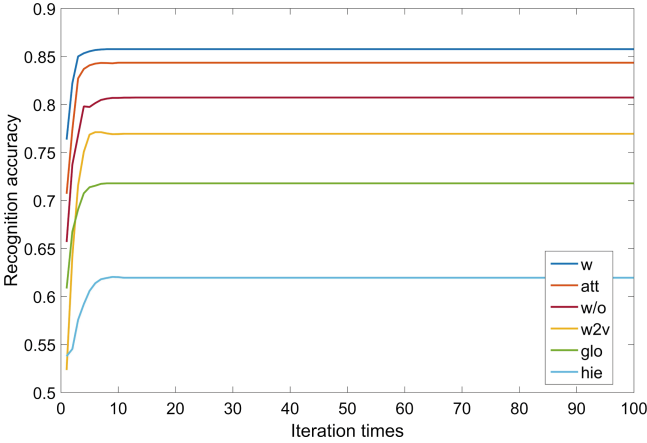


Fig. 2. Average per-class Top-1 accuracy (%) of unseen classes is reported with structure fusion and propagation iteration times on AwA. w: the fusion includes att, w2v, glo and hie, while w/o: the fusion contains w2v, glo and hie

structure consistency between the semantic representation and the image feature, and can better distinguish fine-grained classes. SFP can complement the manifold structure between the semantic representation and the image feature, and better recognize coarse-grained classes. Therefore, integrating two ideas is expected to further improve the ZSL performance in future work.

3.6 Experimental Result Analysis

From the above experiments, we can attain the following observations.

- The semantic description have the different contribution for classifying unseen classes. The supervised attribute tend to obtain the better recognition performance than the unsupervised semantic representation (w2v, glo and hie) in AwA and CUB. In the unsupervised semantic representation, the recognition accuracy of w2v or glo is better than that of hie in AwA and CUB, but the performance of hie is superior to that of w2v or glo in Dogs. This is mainly due to the flexibility and uncertainty of the semantic representation in the unsupervised way.
- The performance of SFP is better than that of other three methods, which are SJE, LatEm, and SynC. However, the performance improvement is different in the various datasets. The obvious improvement can be found in AwA, Dogs and SUN, while the slight improvement can be shown in CUB. The main reason of this situation is related to whether or not effectively to propagate the positive structure in the optimization computation in term of data differences.
- SFP emphasizes on the different manifold structure complement, while DMaP focuses on the various manifold structure consistency. Therefore, the performance of SFP is superior to that of DMaP because the structure complementarity plays the important role for learning transfer model in AwA and

Table 3. Comparison of SFP method with state-of-the-art methods for ZSL, average per-class Top-1 accuracy (%) of unseen classes is reported based on the same data configurations. ‘+’ indicates fusion operation.

Method	Semantic feature	T/I	AwA	CUB	Dogs
SJE	att	I	66.7	50.1	N/A
	w2v	I	51.2	28.4	19.6
LatEm	att	I	71.9	45.5	N/A
	w2v	I	61.1	31.8	22.6
SynC	att	I	69.3	47.5	N/A
	w2v	I	52.9	32.3	27.6
DMaP	att	T	74.9	61.8	N/A
	w2v	T	67.9	31.6	38.9
	att+w2v	T	78.6	59.6	N/A
SFP	att	T	84.3	51.8	N/A
	w2v	T	77.4	32.5	33.3
	att+w2v	T	84.7	52.5	N/A
	att+w2v+glo+hie	T	85.4	54.1	N/A
	w2v+glo+hie	T	81.4	35.3	48.1

Dogs, and the performance of DMaP is better than that of SFP because the structure consistency is a key point for classifying unseen classes in CUB.

- SFP performs better with the positive structure fusion and propagation. SFP has demonstrated great promise in above experiments due to multi-manifold structure consideration and alternated optimization between the weight computation and the manifold structure estimation for ZSL.
- The proposed fusion method can attain the better performance than the non-fusion method because of appropriate complementing each other. w or w/o always performs better on AwA, CUB and Dogs.

4 Conclusion

We have proposed a new ZSL method, which called structure fusion and propagation (SFP). This method can not only directly model the relevance among the manifold structures in semantic and image space, but also dynamically propagate the positive structure by the crossing iteration. Specifically, the proposed SFP method mainly includes four parts. First, nonlinear model constructs the mapping relationship between the class label and the visual image representation. Second, graph describes the relevance between seen classes and unseen classes in semantic or image space. Three, loss function indicates the constrains relationship of multi-manifold structure to balance the structure dependance. Last, structure fusion and propagation is implemented by the crossing iteration computation between phantom classes and weights solving. For evaluating the

proposed SFP, we carry out the experiment on AwA, CUB, Dogs and SUN. Experimental results show that SFP can obtain the promising results for ZSL.

References

1. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2927–2936 (2015)
2. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(7), 1425–1438 (2016)
3. Changpinyo, S., Chao, W.L., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5327–5336 (2016)
4. Deng, J., Krause, J., Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587 (2013)
5. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: zero-shot learning using purely textual descriptions. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2584–2591 (2013)
6. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1778–1785 (2009)
7. Frome, A., et al.: DeViSE: a deep visual-semantic embedding model. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 2121–2129 (2013)
8. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Transductive multi-view zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(11), 2332–2345 (2015)
9. Fu, Z., Xiang, T.A., Kodirov, E., Gong, S.: Zero-shot object recognition by semantic manifold distance. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2635–2644 (2015)
10. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Unsupervised domain adaptation for zero-shot learning. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2452–2460 (2015)
11. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 951–958 (2009)
12. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(3), 453–465 (2014)
13. Li, X., Guo, Y., Schuurmans, D.: Semi-supervised zero-shot classification with label representation learning. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 4211–4219 (2016)
14. Li, Y., Zhang, J., Zhang, J., Huang, K.: Discriminative learning of latent features for zero-shot recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7463–7471 (2018)
15. Li, Y., Wang, D., Hu, H., Lin, Y., Zhuang, Y.: Zero-shot recognition using dual visual-semantic mapping paths. *arXiv preprint [arXiv:1703.05002](https://arxiv.org/abs/1703.05002)* (2017)
16. Lin, G., Fan, C., Zhu, H., Miu, Y., Kang, X.: Visual feature coding based on heterogeneous structure fusion for image classification. *Inf. Fusion* **36**, 275–283 (2017)

17. Lin, G., Fan, G., Kang, X., Zhang, E., Yu, L.: Heterogeneous feature structure fusion for classification. *Pattern Recognit.* **53**, 1–11 (2016)
18. Lin, G., Liao, K., Sun, B., Chen, Y., Zhao, F.: Dynamic graph fusion label propagation for semi-supervised multi-modality classification. *Pattern Recognit.* **68**, 14–23 (2017)
19. Mensink, T., Gavves, E., Snoek, C.G.M.: Costa: co-occurrence statistics for zero-shot classification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2441–2448 (2014)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 3111–3119 (2013)
21. Patterson, G., Xu, C., Su, H., Hays, J.: The sun attribute database: beyond categories for deeper scene understanding. *Int. J. Comput. Vis.* **108**(1), 59–81 (2014)
22. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014)
23. Qi, G.J., Liu, W., Aggarwal, C., Huang, T.S.: Joint intermodal and intramodal label transfers for extremely rare or unseen classes. *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**(99), 1 (2016). <https://doi.org/10.1109/TPAMI.2016.2587643>
24. Romera-Paredes, B., Torr, P.H.: An embarrassingly simple approach to zero-shot learning. In: *International Conference on Machine Learning (ICML)*, pp. 2152–2161 (2015)
25. Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 935–943 (2013)
26. Szegedy, C., et al.: Going deeper with convolutions. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9 (2015)
27. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds200-2011 dataset. California Institute of Technology (2011)
28. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 69–77 (2016)
29. Yu, F.X., Cao, L., Feris, R.S., Smith, J.R., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 771–778 (2013)
30. Zhang, C., Peng, Y.: Visual data synthesis via GAN for zero-shot video classification. *arXiv preprint arXiv:1804.10073* (2018)
31. Zhang, E., Chen, W., Zhang, Z., Zhang, Y.: Local surface geometric feature for 3D human action recognition. *Neurocomputing* **208**, 281–289 (2016)
32. Zhang, Y., Zhang, E., Chen, W.: Deep neural network for halftone image classification based on sparse auto-encoder. *Eng. Appl. Artif. Intell.* **50**, 245–255 (2016)
33. Zhang, Z., Saligrama, V.: Zero-shot learning via joint latent similarity embedding. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6042 (2016)



A Hierarchical Cluster Validity Based Visual Tree Learning for Hierarchical Classification

Yu Zheng^{1,2}, Jianping Fan³, Ji Zhang⁴, and Xinbo Gao²(✉)

¹ School of Cyber Engineering, Xidian University, Xi'an 710071, People's Republic of China

² School of Electronic Engineering, Xidian University, Xi'an 710071, People's Republic of China
xbgao@mail.xidian.edu.cn

³ Department of Computer Science, University of North Carolina, Charlotte, NC 28223, USA

⁴ Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, People's Republic of China

Abstract. For hierarchical learning, one open issue is how to build a reasonable hierarchical structure which characterize the inter-relation between categories. An effective approach is to utilize hierarchical clustering to build a visual tree structure, however, the critical issue of this approach is how to determine the number of clusters in hierarchical clustering. In this paper, a hierarchical cluster validity index (HCVI) is developed for supporting visual tree learning. Before clustering of each level begins, we will measure the impact of different numbers of clusters on visual tree building and select the most suitable number of clusters. The proposed HCVI will control the structure of visual tree neither too flat nor too deep. Based on this visual tree, a hierarchical classifier can be trained for achieving more discriminative capability. Our experimental results have demonstrated that the proposed hierarchical cluster validity index (HCVI) can guide the building of a more reasonable visual tree structure, so that the hierarchical classifier can achieve better results on classification accuracy.

Keywords: Hierarchical cluster validity · Number of clusters
Visual tree · Hierarchical classification

1 Introduction

Recently, hierarchical classification has received enough attention in the field of machine learning [19, 30, 38, 39], and also has been applied successfully in many applications [3, 9, 40]. In general, hierarchical classification has three advantages: (1) Hierarchical classification has higher classification efficiency. In the testing phase, hierarchical classifier only need to go through fewer node classifiers than

flat classifiers [6, 37]. (2) Hierarchical classification can effectively deal with the imbalanced data. (3) The structural characteristics of the hierarchical classifier make it possible to obtain higher classification accuracy when dealing with structured data. For hierarchical learning, one open problem is how to build a reasonable hierarchical structure which characterize the inter-relation between categories.

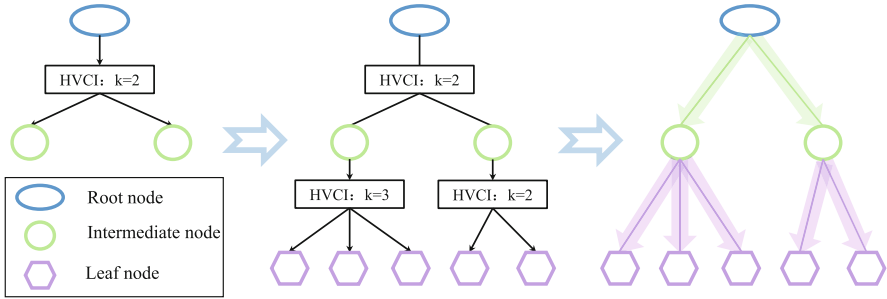


Fig. 1. The framework of hierarchical classifier training. The blue elliptical nodes represent the root nodes; the green circular nodes represent the intermediate nodes; the violet hexagonal nodes represent the leaf nodes. Before clustering, HCVI was utilized to select the optimal number of clusters. This process is applied recursively until the leaf node is reached, and then the hierarchical classifier can be trained over the visual tree from top to bottom. (Color figure online)

In general, the existing approaches for building hierarchical structure can be roughly divided into three types: (1) Semantic tree [12, 22, 33]. It builds an hierarchical structure by leveraging the semantic ontology in the real world. However, it cannot characterise the inter-relation between categories in the feature space. (2) Label tree [9, 23]. To learn a label tree, we need to train a flat one-versus-rest (OVR) binary classifiers first, and then utilize the classification results to build the visual tree. However, the label tree structure always suffer from the imbalanced data and training efficiency. (3) Visual tree [13, 30, 41]. In general visual tree learning, a large number of categories can be organized hierarchically in a coarse-to-fine fashion with hierarchical clustering. Because the feature space is the common space for classifier training and classification, the visual tree can provide a good environment to characterize the inter-relation between categories. However, the number of cluster centers will profoundly influence the structure of the visual tree. Thus, how to determine the number of clusters is a critical issue.

Therefore, the suitable clustering number of hierarchical clustering is the key to building a reasonable visual tree and training a more discriminative classifier. It is necessary to find a way to effectively evaluate the goodness of clustering in order to select the suitable number of clusters. It is worth noting that the cluster validity index (CVI) is often used to evaluate the success of clustering

applications [24, 25]. Cluster validity index can be roughly divided into two categories: external cluster validity index and internal cluster validity index. The main difference is whether the external information is used in the cluster validity. Usually, the external information refers to the category labels. For visual tree, the objects of clustering are categories instead of samples, so there is no external information available for visual tree structure. Therefore, only internal cluster validity index can be used to guide the visual tree building.

Based on these observations, in this paper, an hierarchical cluster validity index (HCVI) is developed for supporting visual tree learning. The HCVI will consider both the clustering results of each level and the structural rationality of the visual tree. In hierarchical clustering, we will measure the impact of different numbers of clusters on visual tree building and select the most suitable number of clusters before clustering of each level begins. Based on the visual tree, a hierarchical classifier can be trained from top to bottom. Figure 1 illustrates the framework of hierarchical classifier training.

This paper is organized as follows. In Sect. 2, we review some relevant work. In Sect. 3, we present the proposed HCVI algorithm for visual tree learning. In Sect. 4, we present our experiments for algorithm evaluation. Section 5 provides some conclusions.

2 Related Work

The existing approaches for building hierarchical structure can be divided into three groups: (a) semantic tree; (b) label tree; (c) visual tree. Some researchers utilize the semantic ontology to organize large numbers of categories hierarchically [8, 12, 22, 26, 27, 33]. Marszalek et al. employ the affiliation between nouns of WordNet to build a semantic tree for visual recognition [26]. Li et al. utilize both image and tag information to discover the semantic image hierarchy, and then employ this hierarchy to encode the inter-categories relations [22]. Fan et al. integrate semantic ontology and multi-task learning to complete the multi-level image annotation [12]. Some researchers build the label tree structure in the feature space [1, 9, 15, 29, 36]. Bengio et al. propose a label embedding tree for multi-class tasks [1]. Griffin et al. automatically generate useful taxonomies for learning hierarchical relationships between categories [15]. However, the label tree structure always suffer from the imbalanced data and training efficiency. Therefore, other researchers learn the visual tree by hierarchical clustering directly [28, 30, 40, 41]. Zheng et al. utilize hierarchical affinity propagation clustering and active learning to build the visual tree [40]. Nister et al. built a vocabulary tree by employing hierarchical clustering [28].

Cluster validity index can be roughly divided into two categories: external cluster validity index and internal cluster validity index. External cluster validity is a measure for evaluating the quality of a clustering by employing the ground truth partition [21, 24, 25]. At present, many external cluster validity indexes have been proposed, such as: Rand Index (RI) [31], Adjusted Rand Index (ARI) [17], Fowlkes and Mallow index (FM) [14], Jaccard Index (JI) [18]. However, in visual

tree learning, no ground truth information is available, so the internal cluster validity index should be used. Internal cluster validity index has been widely used in selecting the number of clusters. Calinski et al. proposed the Calinski-Harabasz index (CH), and it defined as the average between- and within- cluster sum of squares [4, 24]. Davies et al. proposed the Davies-Bouldin index (DB), and it defined as the sum ratio of within-cluster scatter to between-cluster separation [7]. Rousseeuw proposed the Silhouette index (Si), and it is utilized to evaluate the consistency within clusters of data [32]. Tibshirani et al. focused the well separated clusters and developed a Gap index (Gap) [34]. Dunn proposed the Dunn index (Dunn), and it defined as the ratio between the inter-cluster separation to the intra-cluster compactness [11]. Hartigan proposed the Hartigan index (Har) [5, 16].

3 Hierarchical Cluster Validity Index for Visual Tree Learning

In general, both external cluster validity index and internal cluster validity index are used to evaluate the performance of clustering. If we want to use CVI to guide visual tree learning, the most direct method is to find a reasonable CVI and use it to select the suitable number of clusters before each level clustering starts. This approach is appropriate for hierarchical clustering alone. However, although the visual tree is built by employing the hierarchical clustering, its purpose is not to get a good clustering result, but to train a discriminative hierarchical classifier based on it. No matter what CVI is used, it can only select the optimal number of clusters for a single clustering. However, one hierarchical clustering contains many sub-clustering. As computer scientists often say: local greed does not guarantee the global optimum, a satisfactory visual tree structure cannot be obtained through traditional internal CVI guidance. For example: according to CVI, one hierarchical clustering tends to choose fewer clusters at each level, the obtaining visual tree will be deep and narrow, and then the more node classifiers will be trained on one path of hierarchical classifier. Unfortunately, at some times, the more node classifiers passed, the lower the classification accuracy will be.

Based on these understanding, we propose a hierarchical cluster validity index that can measure the clustering validity while taking care of visual tree learning. The vast majority of CVIs are designed based on two key criteria: compactness and separation. The compactness measures the distance between the cluster center and samples in one cluster. Separation measures the pairwise distances between cluster centers. The existing methods have done a good job on these two criteria. Therefore, our hierarchical cluster validity index (HCVI) mainly focuses on visual tree learning. Specifically, we design a parameter based on the clustering results to measure whether the current cluster is suitable for building a visual tree. After that, we combined this parameter with the common CVIs to construct HCVI and employing the HCVI to guide the visual tree building.

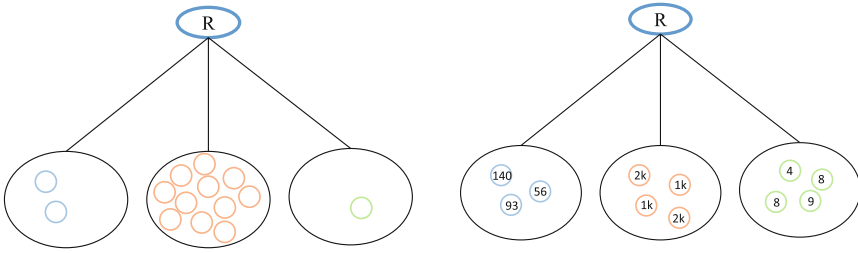


Fig. 2. The overly imbalanced structure of the visual tree. For sub-figure (a), most categories are grouped into one cluster, which leads to category imbalance. For sub-figure (b), Although the categories are relatively balanced, the huge difference in the number of samples in different categories leads to data imbalance.

In the real-world, large numbers of categories are usually imbalanced in the feature space (e.g., some of them have strong inter-category similarities, while others may have weaker inter-category similarities). Therefore, hierarchical clustering also generates an imbalanced visual tree. However, an overly imbalanced structure can also have negative effects on the training of hierarchical classifiers. Figure 2(a) illustrates an overly imbalanced structure. In this figure, each circle represents one category, and one can observe that most categories are grouped into one cluster. It will lead to imbalanced data problems when training hierarchical classifiers over this visual tree. In order to solve this problem, we have developed a parameter to evaluate the category balance, it defined as:

$$\sum_{k=1}^q \left[\left(\frac{r_k - r_E}{r_E} \right)^2 + 1 \right] \tag{1}$$

where parameter q indicates number of clusters. r_E indicates the average number of categories for one cluster. r_k indicates the number of categories contained in the k th cluster.

This parameter indicates the category balance in the clustering. The larger the parameter, the more imbalanced it is. On the other hand, in visual tree learning, the clustering objects are categories. However, when training hierarchical classifiers over the visual tree, every sample needs to be used for training. Therefore, we also need to consider the sample balance. Figure 2(b) illustrates an overly sample imbalance. One can observe that the number of categories in each cluster is almost equal, but the number of samples in each category varies greatly, which can seriously affect the training of hierarchical classifiers. Therefore, we have developed another parameter to evaluate the sample balance, it defined as:

$$\sum_{k=1}^q \left[\left(\frac{m_k - m_E}{m_E} \right)^2 + 1 \right] \tag{2}$$

where m_E indicates the average number of samples for one cluster. m_k indicates the number of samples contained in the k th cluster.

In order to measure the category and sample balance simultaneously, we combine these two parameters to generate a balance parameter, it defined as:

$$\delta(q) = \frac{1}{q} \sum_{k=1}^q \left(\left(\frac{r_k - r_E}{r_E} \right)^2 + 1 \right) \left(\left(\frac{m_k - m_E}{m_E} \right)^2 + 1 \right) \tag{3}$$

This parameter measures the balance of the visual tree learning. The smaller the value, the better the balance. We employ the balance parameter in combination with common CVIs as HCVI to measure the clustering effect so that the optimal number of clusters for hierarchical clustering can be selected. Some CVIs are the bigger the better, such as: CH [4], we denote HCVI as $CH/\delta(q)$, meanwhile, others are the smaller the better, such as: DB [7], we denote the HCVI as $DB \cdot \delta(q)$.

4 Experimental Results

4.1 Notation and Definitions

In this section, we introduce the notations used in the experiment [5], and then provide the definitions about HCVIs and internal CVIs, such as: CH [4], DB [7], Si [32], Dunn [11] and Har [16].

In the following, we denote:

- n = number of samples;
- p = number of variables;
- q = number of clusters;
- $X = \{x_{ij}\}, i = 1, \dots, n, j = 1, \dots, p;$
- \bar{x} = centroid of data matrix X ;
- C_k = the k -th clusters;
- n_k = number of objects in cluster C_k ;
- c_k = centroid of cluster C_k ;
- $d(x, y)$ = distance between x and y ;
- x_i = p -dimensional vector of samples of the i -th object in cluster C_k ;
- $\|x\| = (x^T x)^{1/2};$
- $W_q = \sum_{k=1}^q \sum_{i \in C_k} (x_i - c_k)(x_i - c_k)^T$ is the within-class dispersion matrix;
- $B_q = \sum_{k=1}^q n_k (c_k - \bar{x})(c_k - \bar{x})^T$ is the between-class dispersion matrix;
- $N_t = n(n - 1)/2;$
- $N_w = \sum_{k=1}^q n_k(n_k - 1)/2;$
- $N_b = N_t - N_w;$
- $S_w = \sum_{k=1}^q \sum_{i, j \in C_k, i < j} d(x_i, x_j)$ is sum of the within-cluster distances;

Table 1. Definitions of cluster validity index.

Method	Notation	CVI definition	HCVI definition
Calinski-Harabasz index	CH	$CH = \frac{\text{trace}(B_q)/(q-1)}{\text{trace}(W_q)/(n-q)}$	$CH/\delta(q)$
Davies-Bouldin index	DB	$DB = \frac{1}{q} \sum_{k=1}^q \max_{k \neq l} \left\{ \frac{\chi_k + \chi_l}{d(c_k, c_l)} \right\},$ $\chi_k = \frac{1}{n_k} \sum_{x \in C_k} d(x, c_k),$ $\chi_l = \frac{1}{n_l} \sum_{x \in C_l} d(x, c_l)$	$DB \cdot \delta(q)$
Silhouette index	Si	$Si = \frac{1}{n} \sum_{x=1}^n \frac{b(x) - a(x)}{\max\{a(x), b(x)\}},$ $a(x) = \frac{1}{n_i - 1} \sum_{y \in C_i, y \neq x} d(x, y),$ $b(x) = \min_{j, j \neq i} \left[\frac{1}{n_j} \sum_{y \in C_j} d(x, y) \right]$	$Si/\delta(q)$
Dunn index	Dunn	$Dunn = \frac{\min_{1 \leq i < j \leq q} (\min_{x \in C_i, y \in C_j} d(x, y))}{\max_{1 \leq k \leq q} (\max_{x, y \in C_k} d(x, y))}$	$Dunn/\delta(q)$
Hartigan index	Har	$Har = \left(\frac{\text{trace}(W_q)}{\text{trace}(W_{q+1})} - 1 \right) (n - q - 1)$	$Har \cdot \delta(q)$

$S_b = \sum_{k=1}^{q-1} \sum_{l=k+1}^q \sum_{i \in C_k, j \in C_l} d(x_i, y_j)$ is sum of the between-cluster distances.

Based on these notations, Table 1 shows 5 widely used internal cluster validity index and its corresponding hierarchical cluster validity index. The ‘‘Method’’ column gives the full name of these indices, and the ‘‘Notation’’ column gives the abbreviation. The ‘‘CVI Definition’’ column gives the computation formulas of these indices and the ‘‘HCVI Definition’’ column gives the corresponding hierarchical forms.

4.2 Experimental Settings

In order to verify the effectiveness of the proposed balance parameter, we compare the common CVIs and the balance parameter based HCVIs through experiments. We employ K-means as the clustering algorithm for experiment. All the experiments are carried out on Matlab 2015a. In our experiments, DB, CH, Si can be implemented by using the Statistics and Machine Learning Toolbox of Matlab. We implement the Har index by employed part of CVAP toolbox [35]. Our experimental environment is: a single machine with 4 cores and 16GB memory.

4.3 Experiment for Balance Parameter

In this experiment, we evaluate the proposed approach on *Fisheriris* data set. *Fisheriris* data set is one of available data set at UCI machine learning repository [2]. It has 150 samples with 50 samples in each category. The dimension of the original data is 4. To facilitate visualization, we use the first two dimensions of

Table 2. Criterion values of CVIs.

Number of clusters	2	3	4	5	6	7	8
CH	730.92	873.13	1009.93	1190.18	1230.11	1196.25	1173.77
DB	0.36	0.58	0.70	0.61	0.67	0.69	0.72
Si	0.84	0.69	0.67	0.70	0.67	0.64	0.63
Dunn	0.0207	0.0282	0.0340	0.0654	0.0504	0.0548	0.0504
Har	-184.47	-31.60	60.72	-64.90	318.50	-112.69	63.88

Table 3. Balance parameter of Fisheriris data set.

Number of clusters	2	3	4	5	6	7	8
Balance parameter	1.89	1.57	1.24	1.06	1.40	1.23	1.22

each sample as one sample and the last two dimensions as another. In this way, there are 300 samples in total. In the experiment, we use 5 common CVIs and their corresponding HCVis to evaluate the clustering results with different number of clusters. It is worth noting that each sample in this experiment represents only one single sample, so HCVis can only evaluate the sample balance.

Table 2 shows the criterion values of CVIs. The bold value is the optimal criterion values of different CVIs. One can observe that the optimal number of clusters derived from different CVIs is not identical, even though the clustering data is the same. It shows the criteria of different CVIs vary widely. Since we have reconstructed the data set, the original labels has been invalidated, so we cannot evaluate which CVI is better. However, our main purpose is not to find the optimal indicators, but to verify the effectiveness of the balance parameters. Table 3 shows the balance parameter of different number of clusters. From the result, it is obvious that the clustering result is the most balanced when the number of clusters $q=5$. From Tables 2 and 3, we can observe that most common CVIs do not pay attention to the balance of clustering, which is the precisely concern of building a visual tree. Therefore, HCVI is a reasonable choice for considering both clustering goodness and balance. Figure 3 illuminates the cluster assignments and the criterion values of CVIs and their corresponding HCVis. The first two columns show the results of the CVI, and the last two columns show the results of the corresponding HCVI. We can observe that the common CVIs tend to choose fewer clusters, while HCVis tend to choose more clusters. In particular, the DB and Si indices both consider $q = 2$ as the optimal number of clusters, however, it results in a very imbalanced clustering result. After using the balance parameter, the HCVis of DB and Si have selected a reasonable cluster number that makes the clustering results more balanced. It is worth noting that the balance parameter does not improve balance of Har index. It shows that the Har index hardly considers the clustering balance as a criterion. In summary, we can conclude that the proposed balance parameter can effectively improve the performance of CVIs in terms of clustering balance.

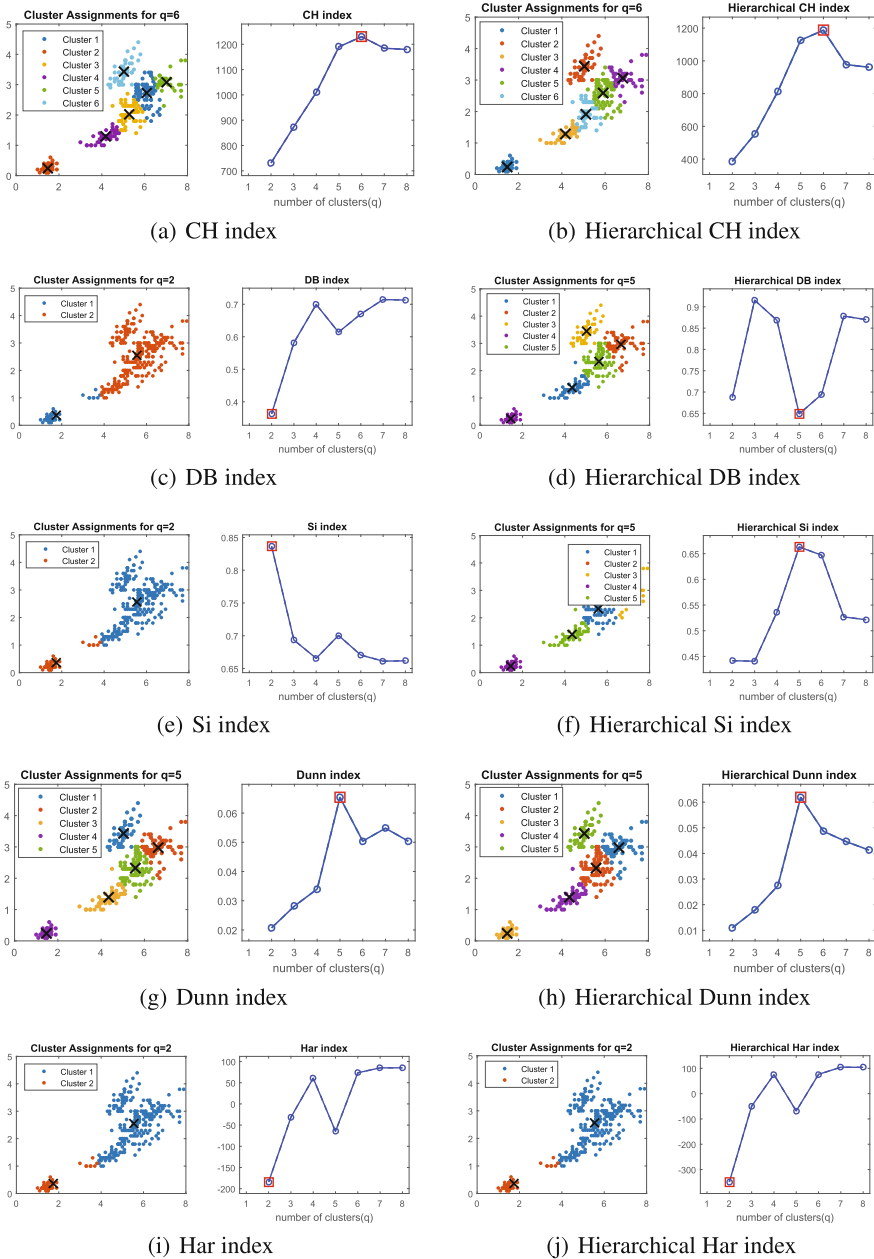


Fig. 3. The cluster assignments and the criterion values of CVIs and their corresponding HCVis.

4.4 Experiment for Hierarchical Classification

In this section, we evaluate the proposed HVCIs comparing the classification accuracy of different visual tree structures. In the experiment, we employ the proposed HCVI to build visual trees and train hierarchical classifiers based on these visual trees. Our experiment are carried out on two data sets: *CIFAR-100* [20], the *ILSVRC-2012* [8]. CIFAR-100 has 100 image categories and each category contains 600 images. We randomly select 10,000 images, half for training and half for testing. ILSVRC-2012 data set is a subset of ImageNet. It contains 1000 image categories and each category has over 1,000 images. We randomly select 20,000 images, half for training and half for testing. In the experiment, we employ DeCAF features as the image representation [10], and then use PCA to reduce the dimensionality of the DeCAF from 4096 to 128.

Table 4. Classification results on CIFAR-100 and ILSVRC-2012 image set.

Approaches	CIFAR-100	ILSVRC-2012
Semantic tree	22.86	26.39
Label tree	24.51	28.11
Visual tree	25.07	28.16
EVT	28.28	28.59
CH-VT	36.98	33.96
DB-VT	25.46	32.21
Si-VT	30.90	39.12
Dunn-VT	28.08	30.02
Har-VT	29.34	28.10
HCH-VT	38.00	35.03
HDB-VT	29.04	34.46
HSi-VT	29.30	38.35
HDunn-VT	34.74	32.71
HHar-VT	31.54	28.58

In this experiments, we compare the proposed HCVI-visual tree structure with two types of tree structure: CVI-visual tree structure and traditional hierarchical structure. In particular, traditional hierarchical structures contains: semantic tree [27], label tree [15], visual tree [40] and EVT [40]. We train hierarchical classifiers based on these tree structure and compare their classification results. We employ K-means as the clustering algorithm for experiment and the SVM classifier as the node classifiers. The Mean Accuracy (%) is used as the criterion to evaluate the performance of all approaches. The experimental results are shown in Table 4. We can observe that the hierarchical classifiers based on visual trees which utilizing cluster validity indices can achieve better results.

The reason is that the cluster validity indices allows us to get better clustering results, so as to get more discriminative visual trees. In addition, most of HCVis-based methods have achieved better results, which illustrates the effectiveness of the proposed balance parameter. It's worth noting that CH index based method achieve higher classification accuracy compared to HCH index based method. One possible reason is that the HCH index considers balance too much and ignores the compactness and the separation of clustering. In general, we can obtain more reasonable visual tree structures through the guidance of HCVis to help train more discriminative hierarchical classifiers.

5 Conclusion

In this paper, a hierarchical cluster validity index (HCVI) is developed to achieve more discriminative solution for visual tree learning, where the hierarchical classifiers can be trained over the visual trees. Our HCVI integrate the proposed balance parameter and the common CVIs. Both the balance of visual tree and the effectiveness of clustering are leveraged to learn more discriminative hierarchical structure. Therefore, the hierarchical classifier can achieve better results. The experimental results have demonstrated that our hierarchical cluster validity index has superior performance as compared with other cluster validity indices on both the clustering balance and the classification accuracy.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under (Grant No. 61432014, No. U1605252, No. 61772402, No. 61671339, No. 61571347, and No. 61603233), in part by National High-Level Talents Special Support Program (Leading Talent of Technological Innovation of Ten-Thousands Talents Program) (No. CS31117200001), in part by the National Key Research and Development Program of China (No. 2016QY01W0200), in part by the Key Industrial Innovation Chain Project in Industrial Domain (Grant No. 2016KTZDGY04-02), in part by the Shaanxi Basic Research Projects in Natural Sciences (No. 2017JQ6076).

References

1. Bengio, S., Weston, J., Grangier, D.: Label embedding trees for large multi-class tasks. In: NIPS, pp. 163–171 (2010)
2. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA, vol. 55 (1998). <http://www.ics.uci.edu/~mllearn/mlrepository.html>
3. Bruse, J.L., et al.: Detecting clinically meaningful shape clusters in medical image data: metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches. *IEEE Trans. Biomed. Eng.* **64**, 2373–2383 (2017)
4. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **3**(1), 1–27 (1974)
5. Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A.: NbClust package: finding the relevant number of clusters in a dataset. *J. Stat. Softw.* (2012)

6. Chen, S., Yang, J., Luo, L., Wei, Y., Zhang, K., Tai, Y.: Low-rank latent pattern approximation with applications to robust image classification. *IEEE Trans. Image Process.* **26**, 5519–5530 (2017)
7. Davies, D.L., Bouldin, D.W.: A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **2**, 224–227 (1979)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: *CVPR*, pp. 248–255 (2009)
9. Deng, J., Satheesh, S., Berg, A.C., Li, F.: Fast and balanced: efficient label tree learning for large scale object recognition. In: *NIPS*, pp. 567–575 (2011)
10. Donahue, J., et al.: DeCAF: a deep convolutional activation feature for generic visual recognition. In: *ICML*, pp. 647–655 (2014)
11. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **4**(1), 95–104 (1974)
12. Fan, J., Gao, Y., Luo, H.: Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE Trans. Image Process.* **17**(3), 407–426 (2008)
13. Fan, J., Zhou, N., Peng, J., Gao, L.: Hierarchical learning of tree classifiers for large-scale plant species identification. *IEEE Trans. Image Process.* **24**(11), 4172–4184 (2015)
14. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **78**(383), 553–569 (1983)
15. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: *CVPR*, pp. 1–8 (2008)
16. Hartigan, J.A.: *Clustering algorithms* (1975)
17. Hubert, L., Arabie, P.: Comparing partitions. *J. Classif.* **2**(1), 193–218 (1985)
18. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull. Soc. Vaudoise Sci. Nat.* **37**, 547–579 (1901)
19. Kalantarian, H., Sideris, C., Sarrafzadeh, M.: A hierarchical classification and segmentation scheme for processing sensor data. *IEEE J. Biomed. Health Inform.* **21**(3), 672–681 (2017)
20. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, U. Toronto (2009)
21. Lei, Y., Bezdek, J.C., Romano, S., Vinh, N.X., Chan, J., Bailey, J.: Ground truth bias in external cluster validity indices. *Pattern Recogn.* **65**, 58–70 (2017)
22. Li, L.J., Wang, C., Lim, Y., Blei, D.M., Fei-Fei, L.: Building and using a semantivisual image hierarchy. In: *CVPR*, pp. 3336–3343 (2010)
23. Liu, B., Sadeghi, F., Tappen, M., Shamir, O., Liu, C.: Probabilistic label trees for efficient large scale image classification. In: *CVPR*, pp. 843–850 (2013)
24. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: *ICDM*, pp. 911–916 (2010)
25. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., Wu, S.: Understanding and enhancement of internal clustering validation measures. *IEEE Trans. Cybern.* **43**(3), 982–994 (2013)
26. Marszałek, M., Schmid, C.: Constructing category hierarchies for visual recognition. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008*. LNCS, vol. 5305, pp. 479–491. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88693-8_35
27. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
28. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: *CVPR*, vol. 2, pp. 2161–2168 (2006)

29. Phan, H., Hertel, L., Maass, M., Koch, P., Mertins, A.: Label tree embeddings for acoustic scene classification. In: ACM MM, pp. 486–490 (2016)
30. Qu, Y., et al.: Joint hierarchical category structure learning and large-scale image classification. *IEEE Trans. Image Process.* **26**(9), 4331–4346 (2017)
31. Rand, W.M.: Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**(336), 846–850 (1971)
32. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
33. Tang, J., Chang, S., Qi, G.J., Tian, Q., Rui, Y., Huang, T.S.: LEGO-MM: learning structured model by probabilistic logic ontology tree for multimedia. *IEEE Trans. Image Process.* **26**(1), 196–207 (2017)
34. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **63**(2), 411–423 (2001)
35. Wang, K., Wang, B., Peng, L.: CVAP: validation for cluster analyses. *Data Sci. J.* **8**, 88–93 (2009)
36. Wu, Q., Tan, M., Song, H., Chen, J., Ng, M.K.: ML-Forest: a multi-label tree ensemble method for multi-label classification. *IEEE Trans. Knowl. Data Eng.* **28**(10), 2665–2680 (2016)
37. Yao, C., Liu, Y.F., Jiang, B., Han, J., Han, J.: LLE score: a new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition. *IEEE Trans. Image Process.* **26**, 5257–5269 (2017)
38. Zhang, L., Shah, S., Kakadiaris, I.: Hierarchical multi-label classification using fully associative ensemble learning. *Pattern Recogn.* **70**, 89–103 (2017)
39. Zhao, T., et al.: Deep mixture of diverse experts for large-scale visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* (2018)
40. Zheng, Y., Fan, J., Zhang, J., Gao, X.: Hierarchical learning of multi-task sparse metrics for large-scale image classification. *Pattern Recogn.* **67**, 97–109 (2017)
41. Zhou, N., Fan, J.: Jointly learning visually correlated dictionaries for large-scale visual recognition applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(4), 715–730 (2014)



Robust Shapelets Learning: Transform-Invariant Prototypes

Huiqi Deng^{1,3}, Weifu Chen¹, Andy J. Ma^{2,3}, Qi Shen^{1,4}, Pong C. Yuen³,
and Guocan Feng¹(✉)

¹ School of Mathematics, Sun Yat-sen University, Guangzhou, China
denghq7@mail2.sysu.edu.cn, {chenwf26,mcsfg}@mail.sysu.edu.cn

² School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
majh8@mail.sysu.edu.cn

³ Department of Computer Science, Hong Kong Baptist University,
Hong Kong, China
pcyuen@comp.hkbu.edu.hk

⁴ Department of Genetics and Genomic Sciences, Icahn School of Medicine
at Mount Sinai, New York, USA
qi.shen@mssm.edu

Abstract. *Shapelets* are discriminative local patterns in time series, which maximally distinguish among different classes. Instead of considering full series, shapelet transformation considers the existence or absence of local shapelets, which leads to high classification accuracy, easy visualization and interpretability. One of the limitation of existing methods is robustness. For example, *Search-based* approaches select sample subsequences as shapelets and those methods intuitively may be not accurate and robust enough. *Learning-based* approaches learn shapelets by maximizing the discriminative ability. However, those methods may not preserve basic shape for visualization. In practice, shapelets are subjected to various geometric transformations, such as translation, scaling, and stretching, which may result in a confusion of shapelet judgement. In this paper, robust shapelet learning is proposed to solve above problems. By learning transform-invariant representative prototypes from all training time series, rather than just selecting samples from the sequences, each time series sample could be approximated by the combination of the transformations of those prototypes. Based on the combination, samples could be easily classified into different classes. Experiments on 16 UCR time series datasets showed that the performance of the proposed framework is comparable to the state-of-art methods, but could learn more representative shapelets for complex scenarios.

Keywords: Robustness · Transform-invariant
Representative prototype

1 Introduction

Time series classification has attracted a lot of attention in many applications, such as finance (e.g. stock market), medical diagnosis (e.g. EEG and ECG),

motion capture and speech recognition. Since the order of series and the dependence of close time stamps are crucial in finding the most discriminative features and patterns, it raises great difficulty for algorithms to classify time series data. Recently, a new primitive named *Shapelets* has been generating increasing interests in time series classification (TSC). Shapelets are discriminative continuous snippets of full series, which maximally distinguish among different classes. Hence, shapelets can be treated as representative of some class, and time series classification turns out to be the problem of presence or absence of some shapelets for representing time series data.

Shapelets for time series have attracted many researchers' interest for two main reasons. First, shapelets focus on local variation rather than global variation as traditional algorithms did, which could be more robust to noise and available for early time series prediction. Second, shapelets reveal the inherent attention mechanism of data so that allow for easier summarization and visualization, which provides explanatory insights to the classification problem.

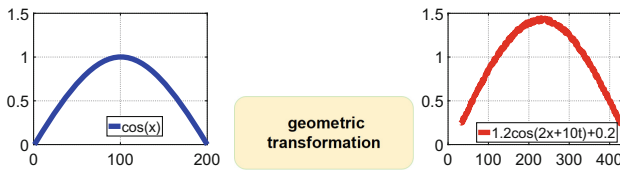


Fig. 1. Demo of deformation. The left column is a cosine curve; the right column is the transformed cosine curve obtained by imposing shift, translation, scaling, stretching transformations and Gaussian noise on the original cosine curve.

Despite the above advantages of shapelets, current shapelets are a little clumsy. We observe that samples in the same class always share a basic shape, while in practice, the basic shape may subject to individual difference and go through various deformations, such as shift, translation, scaling, stretching and so on. As shown in Fig. 1, the cosine curve is actually similar to the transformed cosine curve in term of shape, despite individual noises and slight differences in phase (shift), amplitude (scaling), offset (translation), uniform scaling (stretching).

In this paper, we introduce a new conception, “shapelets prototypes” and propose a transform-invariant robust shapelet learning framework based on dictionary learning theory. The proposed framework aims to learn representative basic shapes that are learned from the transformed subsequences by alternative iteration. In each iteration, robust shapelet learning performs two steps: a) in the alignment step, the best transformation operators are automatically obtained by minimizing the average least square error between the transformed subsequences and current shapelets prototype; b) in the refinement step, the dictionaries are updated to reflect the basic shapes from transformed training series. Figure 2 shows a real-world example from the GunPoint dataset. *S1* and

S_1 and S_2 are learned shapelet prototypes, T_1 – T_3 are three time series from gun class, and T_4 – T_6 are examples from no gun class. Here S_1 represents an action of drawing guns from holsters, and S_2 represents returning guns to the holsters, which are critical patterns for classifying the two classes. While such actions are subject-dependent, and the corresponding patterns from samples shows a high variety due to individual factors. As seen from Fig. 2, our proposed method can align probe sample to learned prototypes and reveals the inherent knowledge of data. Our contributions can be summarized as:

- We propose a robust shapelet learning framework based on dictionary learning theory, which is invariant to various deformations;
- The discovered shapelet prototypes can explore intrinsic shapes which are more general and expressive.
- Shapelet prototypes well preserve the basic shapes and hence have better interpretability.

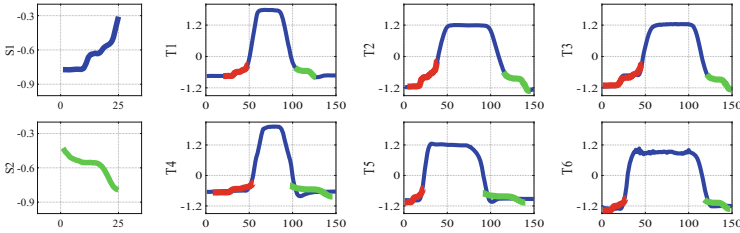


Fig. 2. Illustration of two shapelet prototypes S_1, S_2 (leftmost plots) learned from *GunPoint* dataset. First row (T_1, T_2, T_3) and second row (T_4, T_5, T_6), are instances from Gun class and NoGun class respectively. Each sample matched prototypes to all segments and projected corresponding segments (red and green) to prototypes by optimal transformation parameters. It can be observed that series in Gun have a better match to prototypes than series in NoGun. (Color figure online)

2 Related Work

Shapelets are discriminative shapes that can be used to classify time series data effectively. Shapelet learning algorithms make classification decision based on the presence or absence of shapelets in representing a time series. Existing shapelet learning methods can be categorized into *search-based* algorithms [7, 8, 10, 13, 14] and *learning-based* algorithms [4, 12]. The original *search-based* algorithm adopts a brute-force strategy to select shapelets from a large pool of candidate segments and select the most expressive subsequences by various quality criteria, which is time-consuming [14]. In order to reduce the searching time, several algorithms have been proposed to speed up the algorithm by skipping similar segments so

that the number of candidates are greatly reduced [7, 8, 10, 13]. Instead of learning the shapelets first, learning-based algorithms try to learn shapelets and classify time series data simultaneously. Grabocka et al. proposed a classification logistic loss function to jointly learn the shapelets and the logistic regression classifier through stochastic gradient descent approach. *Search-based* approaches choose existing subsequences from training time series, while *learning-based* approach sometimes may not preserve the basic shape of shapelets so that may lack of interpretability, as it mainly considered classification ability. In other words, the above methods lack of the ability of learning a transform-invariant prototypes from training series.

Invariance of transformations is important for time-series domain because sequences are easily distorted and always show high variety due to geometric transformations. For examples, scaling (amplitude) and translation (offset) invariance might benefit for seasonal variations of markets with inflation motion caption [9], and shift invariance [15] is essential for the case where time series share similar patterns but in different phase. In addition, uniform scaling invariance is required for heartbeats with measurement periods of different sampling frequency.

Dictionary learning has been proposed to learn a set of basis for compact representation. In dictionary learning frameworks, each sample T can be approximated by a sparse linear combination of learned basis $\{D_k\}_{k=1}^K$.

$$T = \sum_{k=1}^K \alpha_k D_k + \epsilon \quad \text{s.t. } \|\alpha\|_1 \leq c_0, \quad (1)$$

Such models always employ reconstruction error as objective loss function, where basis and representation coefficients are optimized by alternate iteration. Dictionary learning has been proved feasible and desirable in image classification [6], signal reconstruction and representation [1, 11]. Further, it has been shown great power in scalable data mining and has strong interpretability and generalization capability [16].

3 Learning Transform Invariant Shapelet Prototypes

Formulations. In this section we adopt a conception of “prototype” D , representative shapes learned from training samples. For a set of subsequences without transformation, the prototype is usually computed based on Euclidean space:

$$\begin{aligned} \arg \min_{\substack{\alpha^i \in \mathbb{R}^K \\ D_k \in \mathbb{R}^q}} \sum_{i=1}^N \|T^i - \sum_{k=1}^K \alpha_k^i D_k\|^2 + \lambda \sum_{i=1}^N \|\alpha^i\|_1 \\ \text{s.t. } \|D_k\|^2 \leq d_0, \quad \text{for } k = 1 \dots, K \end{aligned} \quad (2)$$

While it’s not appropriate to utilize Euclidean distance by $L2$ norm in many cases, as we often pay more attention to the shape similarity. Even tiny operation in scaling, translation, and stretching may rapidly swamp shape similarity measured by Euclidean distance.

For subsequences with various transformations, we introduce a conception of transformation operator τ , and τ may be a compound of multiple transformations. An intuitive observation is that each sample subsequences is transformed from shared “prototype” dictionary bases D by a corresponding transformation operator:

$$\tau^i(T^i) = \sum_{k=1}^K \alpha_k D_k + \epsilon^i \quad \text{s.t. } \|\alpha\|_1 \leq c_0, \quad (3)$$

where each sample has specific transformation operator τ^i . Therefore, the learned dictionaries, removing the effect of transformations, is defined as “prototypes” and could be estimated by minimizing the reconstruction error between the sample subsequence and the aligned reconstruction samples. The reformulation is as follows:

$$\begin{aligned} \arg \min_{\substack{\alpha^i \in R^K \\ D_k \in R^q \\ \tau^i \in \omega}} \sum_{i=1}^N \|\tau^i(T^i) - \sum_{k=1}^K \alpha_k^i D_k\|^2 + \lambda \sum_{i=1}^N \|\alpha^i\|_1 \\ \text{s.t. } \|D_k\|^2 \leq d_0, \quad \text{for } k = 1, \dots, K. \end{aligned} \quad (4)$$

Here prototypes D for series and transformation operator τ^i and codes α^i for each sample need to be optimized.

Transform Definition

The proposed robust shapelet learning framework aims to find the intrinsic local patterns of time series. For a prototype $w \in R^p$, a vector ordered by time stamps, most classical linear transformations can be represented as:

$$T^i = \tau^i(w) + \epsilon_i = a_i w(\mu_i t) + c_i + \epsilon_i, \quad (5)$$

where a_i is a scaling factor, c_i is a translation factor and μ_i is a stretching factor. Here τ_i defines a general transformation, where scaling, translation and stretching operators are its special cases:

- **Scaling operator:** Scaling transformation describes differences in amplitude between $w(t)$ and transformed sequence $T_i(t)$:

$$T_i(t) = a_i w(t) \quad t = 1, \dots, p \quad (6)$$

- **Translation operator:** Translation transformation describes a translation along y axis from $w(t)$ to transformed sequence $T_i(t)$:

$$T_i(t) = w(t) + c_i \quad t = 1, \dots, p \quad (7)$$

- **Stretch operator** [3]: Uniform scaling transformation is used for matching sequences with different lengths. Subsequences with different length may require a stretching or shrinking operation due to different tempo or sampling frequency:

$$T_i(t) = w\left(\left\lceil \frac{t}{\mu} \right\rceil\right) \quad t = 1, \dots, \lceil \mu p \rceil \quad (8)$$

Note that stretching operator defined is motivated by uniform scaling operation through ceiling function [3].

Figure 3 shows the above defined transformations. Similarly, we can define the inverse operators for the above transformations.

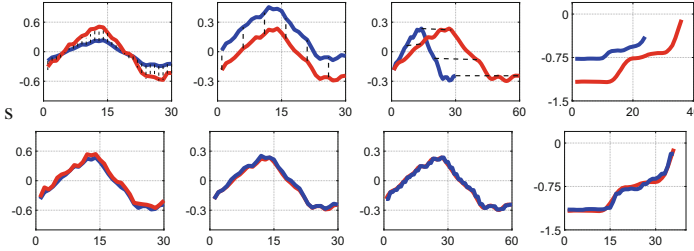


Fig. 3. Difference between transformed subsequence (plotted in red) and original subsequence (blue): Scale operator $a = 1.5$ (leftmost). Translation operator $c = 0.2$ (second column). Stretch operator $\mu = 2$ (third column). Multiple transformations on GunPoint dataset (rightmost). (Color figure online)

4 Model Inference

Optimization objective function is a non-convex problem with respect to the prototypes and transformation operators. We adopt a coordinate descent approach for iteration alternatively. And in each iteration, we perform two steps shown in Algorithm 1: (i) Alignment step: Given current dictionary, updating corresponding transformation operator for each sample subsequences. (ii) Refinement step: Given all transformation operators, the dictionary is updated to reflect the basic shapes learnt from training series.

Alignment Step. Fix the dictionary and sparse coding, i.e., D and α , reconstruction sequence has been determined. We are going to find the optimal alignment (deformation) between original series T^i and reconstruction sequence. To learn the transform parameters, we minimize the mean square error between the reconstruction sequence and subsequence of time series by all possible conversion. It’s remarkable that for input segments T_j^i with stretching factor μ , it has an unequal length to the dictionary. So it needs to do an inverse operation μ^{-1} . For a given sparse coding and dictionary, problems can be solve as:

$$\arg \min_{\tau^i} \|\tau^i(T^i) - \sum_{k=1}^K \alpha_k^i D_k(t)\|^2 \tag{9}$$

where $\tau = \{a, \mu, c\}$, superscript and subscript are dropped for clarify. Obviously, a and c are scaling and translation factors with continuous values, μ is a stretch

Algorithm 1. Transform Invariant shapelet prototypes learning

Input: Initial segments $\{T_i\}_{i=1}^N$, Initial dictionary $D \in R^{K \times q}$.

- 1: **for** $Iter = 1, \dots, MaxIter$ **do do**
- 2: **for** $i = 1, \dots, N$ **do do**
- 3: Alignment step:
- 4: $\tau_i = \text{E.q (9)}$;
- 5: $T'_i = \tau^i(T^i)$;
- 6: **end for**
- 7: Refinement step:
- 8: $\{\alpha, D, \tau\} = \text{E.q (10)}$
- 9: **end for**
- 10: **return** α, D, τ ;

factor with discrete factor. So fixed μ , optimal a, c can be derived by Least Square methods. Then a grid search will be conducted by all possible μ . With obtained parameters, original subsequence $\tau(T)$ at transformed location are exacted, as well as aligned sequence $S = \tau(T) \in R^q$.

Refinement Step. Fix the transform parameters, i.e., $\tau = \{a, \mu, c\}$, optimize α, D . If transform parameters τ are known, it's easy to get updated segments, i.e., updated $\{\tau^i(T^i)\}_{i=1}^N$, which is transformed with fixed length. The problem then reduces to a traditional dictionary learning problem:

$$\begin{aligned} \arg \min_{\alpha, D} \sum_{i=1}^N \|\tau^i(T^i) - \sum_{k=1}^K \alpha_k^i D_k\|^2 + \sum_{i=1}^N \lambda \|\alpha^i\|_1 \\ \text{s.t. } \|D_k\|^2 \leq c, \quad \text{for } k = 1 \dots, K \end{aligned} \quad (10)$$

Coordinate descent is used to solve for dictionary D and sparse coding α .

5 Experiment

5.1 Experiment Setting

Since there are enormous shapelet candidates, we calculated the average sequence to decide the discriminative ability by information gain. And shapes with low discriminative power have been removed so that the volume of candidates would be much smaller (similar, redundant candidates have been discarded), and only valuable candidates were selected as input segments. In training phase, the algorithm was initialized by subsequence matching, where we aligned the initial candidate to all possible segment and extracted the most similar projection. Then, these segments T_i were fed into our transform invariant prototype learning framework and after iterations, a transform-invariant dictionary could be derived. Lastly in testing phase, the sparse representation were conducted based on learned transform-invariant dictionary learning.

Table 1. Statistics of the benchmark time series datasets

	Train/test	Length	Class
Adiac	390/391	176	37
Beef	30/30	470	5
Coffee	28/28	286	2
Diatom	16/306	345	4
ECGFiveDays	23/861	136	2
FaceFour	24/88	350	4
GunPoint	50/150	150	2
ItalyPower	67/1029	24	2
Lighting7	70/73	319	7
MedicalImages	381/760	99	10
MoteStrain	20/1252	84	2
Sony	20/601	70	2
Symbols	25/995	398	6
SytheticC	300/300	60	6
Trace	100/100	275	4
TwoLeadECG	23/1129	82	2

During candidate selection process, it requires the tuning of hyper-parameters, which were found through a grid search approach using cross-validation over the training data. The initial number of shape set was searched in a range of $K_1 \in \{0.05, 0.1\} * Q$, and the length of shapelets $q \in \{0.125, 0.25, 0.375, 0.5\} * Q$, where Q is the full series length. During dictionary learning process, the number of atoms in dictionary $K_2 \in \{5, 10, 15\}$, while the sparsity parameter $\lambda \in \{0.1, 1\}$. For efficiency, the operation of the above deformation could be selected flexibly.

5.2 Complexity Analysis

The time complexity of shapelet prototypes learning consists of two parts, one for initialization segments exaction, and the other for transform invariant dictionary learning. For initialization segments exaction, we need to compute a robust shape similarity, instead of optimal transformation operators, between shapelets with length q candidates and all possible segments. Therefore, we adopted a Z-normalization distance with all range of stretching factors, and the cost is $O(Cq)$, where C is the number of stretching factors. Here, the complexity is similar to the Euclidean distance computation, with a constant multiplier.

For transform invariant dictionary learning, the cost part of alignment step is the computing for τ^i , which takes $O(CNq)$, including the complexity of least square solution of scaling, translation factor and grid search on stretching factors.

Refinement step takes $O(KNq)$, including sparse coding and dictionary learning. Therefore, the total complexity for one call to Algorithm 1 is $O(M((C+K)Nq))$, where M is the maximum number of iterations allowed in Algorithm 1. M can be set quite small. In practice, 20 iterations would be sufficient.

5.3 Classification Accuracy

Experiments were performed on the 16 commonly used UCR time series benchmark which could be downloaded from UCR website [2], and the information of the those datasets were listed in Table 1. For the sake of equivalent comparison, we used the same training and testing split for all the methods. And in the experiments, SVM was chosen as the classifier.

There are many shapelet learning algorithms proposed in the last decades. Ref. [5] compared the performance of some popular shapelet learning algorithms. In this work, we compared our algorithm with three state-of-art baselines IGSVM, LTS and FLAG:

- IGSVM [7]: Shapelet-transform algorithm, which uses the linear SVM as classifier and information as shapelet quality measurement.
- LTS [4]: learning time series shapelets algorithm, which learns the shapelets and logistic classifier automatically and jointly.
- FLAG [5]: Learning position of shapelets, which maximizes the ratio of projected data variances between classes by fused lasso constraints efficiently.

Table 2 shows the classification accuracy of baseline and the proposed method. Our method shows a superiority to IGSVM, FLAG and a comparable performance to LTS, even better prediction accuracy in several dataset. In addition, dictionary learning is desirable for scalable analysis, as well as interpretability.

5.4 Exploratory Data Analysis

One of the strengths of using shapelets as a classification tool is that they provide an easy interpretation and summarization behind data that other classification approaches simply do not. It helps for mining inherent structure. One of the key motivation of our work is to capture an intrinsic structure and knowledge behind data, which is interpretable and unified for data.

To verify the power of transform invariant shapelet learning, we briefly analyze a classical problem in time series data mining domain. On the Gun/NoGun motion capture time series dataset, there are 100 instances from each class. In the Gun class, the actors have their hands by their sides, draw a gun from a hip-mounted holster, point it at a target for approximately one second, and then return the gun to the holster and their hands to their sides. In contrast, in the NoGun class, actors do the similar hands-down, point, hold, and return motion without the gun in their hands and therefore are pointing to a target using the index finger. The classification problem is to distinguish between above two very

Table 2. Classification accuracy on 16 commonly used dataset.

Dataset	Compared methods			
	IGSVM	LTS	FLAG	Ours
Adiac	23.5(4)	51.9(3)	75.2(1)	61.8(2)
Beef	90.0(1)	76.7(3)	83.2(3)	76.7(3)
Coffee	100.0(1)	100.0(1)	100.0(1)	100.0(1)
Diatom	93.1(4)	94.2(3)	96.4(1)	95.4(2)
ECGFiveDays	99.0(3)	100.0(1)	92.0(4)	100.0(1)
FaceFour	97.7(1)	94.3(3)	90.9(4)	96.6(2)
GunPoint	100.0(1)	99.6(2)	96.7(4)	100.0(1)
ItalyPower	93.7(4)	95.8(1)	94.6(3)	95.3(2)
Lighting7	63.0(4)	79.0(1)	76.7(3)	78.1(2)
MedicalImages	52.2(4)	71.2(2)	71.4(1)	70.5(3)
MoteStrain	88.7(2)	90.0(1)	88.7(2)	88.7(2)
Sony	92.7(1)	91.0(4)	92.9(1)	91.5(3)
Symbols	84.6(4)	94.5(2)	87.5(3)	97.1(1)
SytheticC	87.3(4)	97.3(3)	99.7(1)	99.0(2)
Trace	98.0(4)	100.0(1)	99.0(3)	100.0(1)
TwoLeadECG	100.0(1)	100.0(1)	99.0(4)	100.0(1)
AverageRank	2.69	2	2.43	1.78

similar action. Moreover, the dataset consists of instances from two actors, who differ in baseline height about 12 in. (translation) and motion ‘style’, including different movement range (scaling), tempo (stretching). To sum up, the challenges relied on the similarity between two action and the intra-class diversity due to geometric transformation.

As shown in Fig. 4, the original top shapelet trained by [14], represents a “overshot” phenomenon at the end of series. It contains an action corresponding to the arm being lowered back into position. However, at the begin of series, Gun class has a specific shape found by proposed method before a consistent action of raising the arm. That’s because action from Gun class has to draw a gun from a holster. Intuitively, it’s one of intrinsic feature of Gun class, which differ from NoGun class. However, owing to the intra-class diversity, instances differ in offset, scale and stretching factor. DTW and Z-normalization fail to deal with such variance, so that they failed to explore the latent discriminative power of $S1$, because of the complex transformation. While our transform invariant framework achieved meaningful detection of shapelets. The graphs in Fig. 4 shows a best alignment for $S1$ by projecting it through optimal deformation factors.

Therefore, shapelet found by proposed method reflect the essential difference between classes. Interestingly, once instance take a gun from a holster, we can achieve a earlier judgement or prediction for Gun/NoGun classification problem.

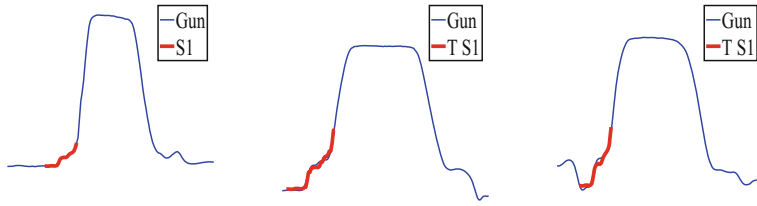


Fig. 4. An illustration for complex transformation from the Gun class, where $TS1$ represents the optimal projection (transformation) from $S1$ to samples.

Acknowledgement. This work is partially supported by the NSFC under grants Nos. 61673018, 61272338, 61703443 and Guangzhou Science and Technology Founding Committee under grant No. 201804010255 and Guangdong Province Key Laboratory of Computer Science.

References

1. Chen, X., Du, Z., Li, J., Li, X., Zhang, H.: Compressed sensing based on dictionary learning for extracting impulse components. *Sign. Process.* **96**, 94–109 (2014)
2. Chen, Y., et al.: The UCR time series classification archive, July 2015
3. Fu, W.C., Keogh, E., Lau, L.Y., Ratanamahatana, C.A., Wong, C.W.: Scaling and time warping in time series querying. *VLDB J.* **17**(4), 899–921 (2008)
4. Grabocka, J., Schilling, N., Wistuba, M., Schmidt-Thieme, L.: Learning time-series shapelets. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 392–401. ACM (2014)
5. Hou, L., Kwok, J.T., Zurada, J.M.: Efficient learning of timeseries shapelets. In: *Thirtieth AAAI Conference on Artificial Intelligence* (2016)
6. Kong, S., Wang, D.: A dictionary learning approach for classification: separating the particularity and the commonality. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012*. LNCS, vol. 7572, pp. 186–199. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33718-5_14
7. Lines, J., Davis, L.M., Hills, J., Bagnall, A.: A shapelet transform for time series classification. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 289–297. ACM (2012)
8. Mueen, A., Keogh, E., Young, N.: Logical-shapelets: an expressive primitive for time series classification. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1154–1162. ACM (2011)
9. Paparrizos, J., Gravano, L.: K-shape: efficient and accurate clustering of time series. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1855–1870. ACM (2015)
10. Rakthanmanon, T., Keogh, E.: Fast shapelets: a scalable algorithm for discovering time series shapelets. In: *Proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 668–676. SIAM (2013)
11. Rubinstein, R., Zibulevsky, M., Elad, M.: Double sparsity: learning sparse dictionaries for sparse signal approximation. *IEEE Trans. Sig. Process.* **58**(3), 1553–1564 (2010)

12. Shah, M., Grabocka, J., Schilling, N., Wistuba, M., Schmidt-Thieme, L.: Learning DTW-shapelets for time-series classification. In: IKDD Conference on Data Science, p. 3 (2016)
13. Wistuba, M., Grabocka, J., Schmidt-Thieme, L.: Ultra-fast shapelets for time series classification. arXiv preprint [arXiv:1503.05018](https://arxiv.org/abs/1503.05018) (2015)
14. Ye, L., Keogh, E.: Time series shapelets: a new primitive for data mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 947–956. ACM (2009)
15. Zhao, R., Schalk, G., Ji, Q.: Temporal pattern localization using mixed integer linear programming
16. Zheng, G., Yang, Y., Carbonell, J.G.: Efficient shift-invariant dictionary learning. In: KDD, pp. 2095–2104 (2016)



A Co-training Approach for Multi-view Density Peak Clustering

Yu Ling^{1,2}, Jinrong He^{1,2}(✉), Silin Ren^{1,2}, Heng Pan^{1,2}, and Guoliang He³

¹ College of Information Engineering, Northwest A&F University, Yangling 712100, Shaanxi, China

hejinrong@nwfau.edu.cn

² Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling 712100, Shaanxi, China

³ School of Computer Science, Wuhan University, Wuhan 430072, China

Abstract. In this paper, we propose a multi-view clustering algorithm based on fast search and find of density peaks. We combined the original clustering algorithm with co-training to handle multi-view data and implement self-adapting cluster center selecting through cluster fusion. Based on the assumption that a point would be assigned to the same cluster in all views, we search for the clustering result that agree across the views by continually modifying one view with the clustering from another view. We demonstrate the efficacy of the proposed algorithm on several test cases.

Keywords: Peak clustering · Multi-view learning · Co-training
Cluster center · Adaptive clustering

1 Introduction

Unlabeled data exist in nature widely, and labeling each sample in a big-scale data in multi-view learning costs a lot of time and work. Thus, we focus on unsupervised learning. Clustering algorithms are widely used in unsupervised learning, which aim to partition elements based on their similarity. Many clustering algorithms have been proposed such as K-means clustering algorithm seeking to minimize the average squared distance between points in the same cluster [1], spectral clustering [2] dividing the graph up into several subgraphs exploiting the properties of the Laplacian of the original graph and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [3] viewing clusters as high-density areas. In 2014, a clustering algorithm based on fast search and find of

This work was partially supported by the National Natural Science Foundation of China (61876136), China Postdoctoral Science Foundation (2018M633585), Natural Science Basic Research Plan in Shaanxi Province of China (2018JQ6060), the Doctoral Starting up Foundation of Northwest A&F University (2452015302), and Students Innovation Training Project of China (201710712064).

density (DPC) was proposed in [4], which was formed by the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. The DPC algorithm has attracted attention by its good performance on automatically excluding outliers and recognizing clusters irrespective of their shape and of the dimensionality of the space.

In real world, we have access to lots of features from single object, and limited information can be obtained through an individual view. Hence, we attempt to obtain more information through observing an object in multiple views. For examples, we can take a photo of an object in different angles or even by different sensors. Different views make up for the lack of information in single-view learning. Motivated by this factor, many multi-view learning methods have been proposed. In [5], Laplacian support vector machines (SVMs) [6] is extended from supervised learning to multi-view semi-supervised learning. Canonical Correlational Analysis (CCA) [7–9], Bilinear Model (BLM) [10] and Partial Least Squares (PLS) [8, 11, 12] are popular unsupervised approaches in multi-view learning [13]. In 2015, Later Multi-View Linear Discriminant Analysis (MLDA) [14] was proposed through combining CCA and Linear Discriminant Analysis (LDA) [15]. Linear Discriminant Analysis is a single-view learning method seeking an optimal linear transformation that maps data into a subspace. Multi-View Intact Space Learning (MISL) proposed in [16] aims to find a space from several views, which assumes that different views are generated from an intact view. Differing from many multi-view approaches, MISL focuses on the insufficiency of each view. However, we do not pay attention to whether each view is sufficient or not, but focus on how to combine the information of multiple views. Therefore, we focus on co-training [17] which is widely used in multi-view learning.

Recently, many clustering methods are applied in multi-view learning. In 2013, a multi-view method, which combines spectral clustering with co-training is proposed in [18]. In 2015, a Co-Spectral Clustering Based Density Peak is proposed in [19], which replaces k-means in spectral clustering with DPC and combines the extended spectral clustering with co-training. In 2016, a Multi-View Subspace Clustering is proposed in [20], which performs subspace clustering on each view simultaneously, meanwhile guarantees the consistence of the clustering structure among different views.

Some clustering methods demand preset number of clusters such as k-means and spectral clustering. In this paper, we extend the cluster centers selection of the original DPC with cluster fusion to implement self-adaptive cluster centers selection which remains unsolved in [4]. We propose an adjusted co-training framework for DPC which varies weights of views according to views' aggregation. Combining the extended DPC and adjusted co-training, the proposed approach is runed without sensitive parameters.

2 Related Work

2.1 Co-training

Co-training [17] was proposed for problems of semi-supervised learning setting, in which we have access to both labeled and unlabeled samples in two distinct views. It considered the problem of using a small set of labeled samples to boost the performance of unsupervised learning. It has its basis on two assumptions: each view is sufficient for classification independently, and the views are conditionally independent given the labels.

Given the labeled training set L and the unlabeled training set U , here we outline the process of co-training:

- Create a pool U' of examples with u examples chosen randomly from U
- Loop for k iterations:
 - Use L to train a classifier h_1 that considers only the x_1 portion of x
 - Use L to train a classifier h_2 that considers only the x_2 portion of x
 - Allow h_1 to label p positive and n negative examples from U'
 - Allow h_2 to label p positive and n negative examples from U'
 - Add these self-labeled examples to L
 - Randomly choose $2p + 2n$ examples from U to replenish U' .

2.2 Clustering by Fast Search and Find of Density Peaks

Given the distance between data points, density peaks clustering (DPC) [4] chooses data points surrounded by neighbours with lower local density as cluster centers. For data point p_i , two quantities ρ_i and δ_i need to be calculated. ρ_i indicates the number of points that distances between point p_i and these points are less than the cutoff distance d_c . δ_i indicates the distance between point p_i and its nearest neighbour with higher local density, and δ_i is defined as

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij} \quad (1)$$

One can choose d_c so that the average number of neighbors is around 1% to 2% of the total number of points in the data set.

For the point with highest density, δ_i is defined as $\delta_i = \max_j(d_{ij})$. Expect the point with highest density, each point and its nearest neighbour with higher local density are assigned to the same cluster temporarily.

Data points with high ρ and high δ or with high γ defined as $\gamma = \rho\delta$ are selected as cluster center.

To exclude outliers, for each cluster, the algorithm finds a border region, defined as the set of points assigned to that cluster but being within a distance d_c from data points belonging to other clusters. Then the algorithm finds the point with highest density within its border region for each cluster. Its density is denoted by ρ_b . A point is considered part of the cluster core (robust assignment), if their density is higher than ρ_b of its cluster. Otherwise, it is considered part of the cluster halo (suitable to be considered as noise).

3 A Co-training Approach for Multi-view Density Peak Clustering

3.1 Adjusted Co-training Framework

The main idea of the standard co-training is training several classifiers through results produced by themselves. Thus, in the proposed approach, views are modified with their clustering results. In a modified view v'_a , distances between two data points belonging to the same cluster in another view v_b are supposed to decrease according to the aggregation of v_b denoted by A_b , and other distances maintain unchanged. Specifically, given the adjacency matrix D_b of view v_b , we first obtain labels L_b by clustering and calculate modification weight matrix W_b defined as:

$$W_{bij} = \begin{cases} A_b & L_{bi} = L_{bj} \\ 1 & L_{bi} \neq L_{bj} \end{cases} \quad (2)$$

$$A_b = \max \frac{\sum_{L_{bi}=L_{bj}} \frac{D_{bij}}{\max_{D_{boxy}}}}{Size(L_{bi})} \quad (3)$$

In Eq. (3), $Size(L_{bi})$ denotes the size of the cluster which includes data point p_i in view v_b .

The modified view v'_a is defined as

$$v'_{aij} = W_{bij} D_{aij} \quad (4)$$

Similar with the standard co-training, we modify each view with another view's clustering result through some iterations. The modification will be ended when all views' clustering results are the same or $\max_i A_i$ is less than a preset threshold T . The brief process of the proposed approach is shown in Fig. 1.

3.2 Cluster Center Selection and Cluster Fusion

A problem remains unsolved in the original DPC is how to select cluster centers automatically and accurately. To help select cluster centers, the author introduced a quantity γ defined as $\gamma_i = \delta_i \rho_i$ for each data point i , whose value is enormously large for cluster centers [4]. Since we attempt to produce the clustering result through iterations in our adjusted co-training framework, DPC doesn't have to perform perfectly in cluster centers selection during each iteration. Thus, we simply select points whose γ is higher than the average value of γ as temporary cluster centers to ensure that the expected cluster centers are included in the set of chosen points. After this step, we fuse some excessive clusters based on the border region of cluster center defined in [4].

The border region of a cluster is originally used to find the cluster halo which can be regarded as outliers [4]. We discard its function for excluding outliers, and instead we apply it in merging excessive clusters produced by the cluster centers selection. In the process of calculating border densities, for each cluster

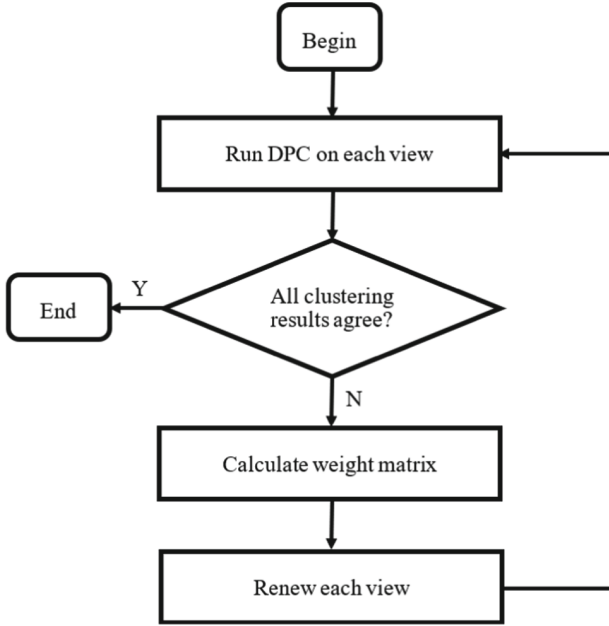


Fig. 1. The brief process of the proposed approach

C_i in we record its border cluster denoted by BC_i within whose border region the border density ρ_{Bi} is obtained, where ρ_{Bi} and BC_i are defined as

$$\rho_{Bi} = \max_{CL_x \neq CL_y, CL_x = C_i} \frac{\rho_x + \rho_y}{2} \tag{5}$$

$$BC_i = arg \max_{CL_y} \frac{\rho_x + \rho_y}{2} \quad (CL_x \neq CL_y, CL_x = C_i) \tag{6}$$

where CL_x denotes the cluster which data point p_x belongs to, and ρ_x denotes the local density of data point p_x .

If the local density of the cluster center in cluster C_i is less than ρ_{Bi} , cluster C_i will be merged with BC_i and the center of new cluster will be the original center of BC_i .

4 Experiment

4.1 Experiment Setup

To demonstrate the efficiency of the proposed approach, we compare our co-trained density peak clustering approach with following baselines:

- **Best Single View (BSV)** Selecting most informative view where clustering result achieving the highest scores.

- **Feature Concatenation (FC)** Concatenating the features from each view, and then running a clustering algorithm on the joint features.
- **Kernel Addition (KA)** Combining different kernels by adding them. As suggested in [21], this seemingly simple approach often leads to near optimal results as compared to more sophisticated approaches for classification. It can be noted that kernel addition reduces to feature concatenation for the special case of linear kernel. In general, kernel addition is same as concatenation of features in the Reproducing Kernel Hilbert Space [18].
- **Kernel Product (element-wise) (KP)** Multiplying the corresponding entries of kernels and applying a clustering algorithm on the resultant matrix. For the special case of Gaussian kernel, element-wise kernel product would be same as simple feature concatenation if both kernels use same width parameter σ [18].

In the section of experiments, we compare performances of DPC with Density Peak Spectral Clustering (DPSC) proposed in [19] combined with above baselines and co-training. DPSC replaces k-means in spectral clustering with DPC to determine number of clusters without preset parameters. The self-adaptive cluster selection is the advantage of the proposed approach as well. Therefore, we compare the proposed approach with DPSC and co-trained DPC instead of spectral clustering or other clustering algorithms requiring sensitive parameters.

4.2 Dataset

– Synthetic Dataset

Our synthetic data consists of 3 views. Each view consists of 2000 data points in two-dimension space ($x_0, x_1, x_2 \in \mathbb{R}^2$) and four central points ($p_0 = (1, 1), p_1 = (-1, -1), p_2 = (1, -1), p_3 = (-1, 1)$). The distribution of data points follows

$$\|x_i - p_{(i \bmod 4)}\|_\infty \leq r \quad (7)$$

where r is a given range for generating data points randomly. We define the true label of data point x_i as $L_i = i \bmod 4$. We evaluate the proposed approach with a synthesis dataset containing three views as shown in Fig. 2.

– MNIST Handwritten Digit

One real-world dataset is taken from the handwritten digits (0–9) data from the MNIST dataset (Modified National Institute of Standards and Technology database). The dataset is consisted of 1000 examples. Digit images are described in two ways: Histogram of Oriented Gradient (HOG) [22] (view-1) and binaryzation (view-2). This dataset will exam the proposed approach’s performance on features extracted with different methods from the same samples.

– IXMAS Actions Dataset

The IXMAS dataset contains recordings of 14 actions from different angles. Images from each angle are regarded as samples in one view. HOG is applied for describing features in views of different angles. This dataset will exam the proposed approach’s performance on samples taken from different angles.

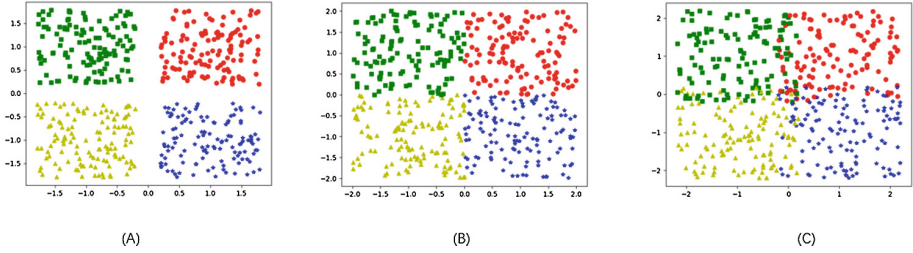


Fig. 2. Three images showing distribution of data points in three views. The range r in view (A) is 0.8; in view (B) is 1.0; and in view (C) is 1.2. Each shape or colour represents one expected cluster.

4.3 Results

The clustering results are evaluated with adjusted rand score (adj-RI) [23] and normalized mutual information score (NMI) [24].

Table 1. Results for synthetic dataset

Method	adj-RI	NMI
BSV DPSC	0.1468	0.6094
BSV DPC	0.4322	0.7328
FC DPSC	0.7828	0.8498
FC DPC	0.7902	0.8681
KA DPSC	0.3850	0.6508
KA DPC	0.4759	0.7359
KP DPSC	0.3306	0.5769
KP DPC	0.5036	0.7484
Co-trained DPSC	0.2298	0.5457
Co-trained DPC	0.9683	0.9712

Table 1 shows the clustering result on synthetic dataset. Our approach outperforms all baselines by a significant margin. The feature concatenation is the second best one among remaining baselines. Compared with DPSC, the proposed approach integrates information in three views and avoids degradation of performance.

Table 2 shows the clustering result on MNIST digit dataset. Our approach outperforms all the baselines in adj-RI score and its NMI score is close to the best one. Performances of kernel addition and kernel product are close to that of the best single view.

Table 2. Results for MNIST dataset

Method	adj-RI	NMI
BSV DPSC	0.3480	0.6271
BSV DPC	0.3633	0.5665
FC DPSC	0.4118	0.6395
FC DPC	0.4164	0.6649
KA DPSC	0.3966	0.5931
KA DPC	0.3511	0.5796
KP DPSC	0.3238	0.5315
KP DPC	0.3421	0.6011
Co-trained DPSC	0.3498	0.5637
Co-trained DPC	0.4797	0.6456

Table 3 shows the clustering results on IXMAS action dataset. On this dataset, our approach outperforms all baselines by a significant margin. Except the co-trained DPC, other baselines combined with DPC perform worse than the Best Single View combined with DPC do.

Table 3. Results for IXMAS dataset

Method	adj-RI	NMI
BSV DPSC	0.3479	0.6491
BSV DPC	0.3841	0.6550
FC DPSC	0.4214	0.6957
FC DPC	0.3398	0.6250
KA DPSC	0.3960	0.6723
KA DPC	0.3429	0.6353
KP DPSC	0.3238	0.5315
KP DPC	0.3554	0.6422
Co-trained DPSC	0.3746	0.6772
Co-trained DPC	0.5178	0.7495

Figures 3, 4 and 5 show adj-RI scores in different datasets with increase of the number of iterations. The proposed approach complete clustering by few steps of iteration.

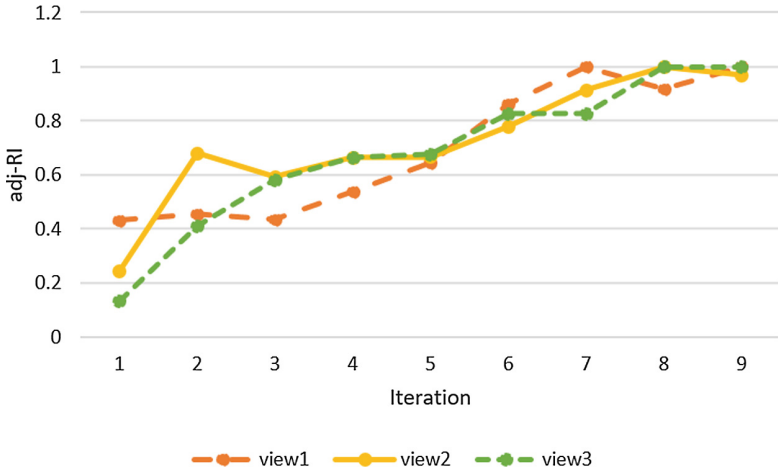


Fig. 3. adj-RI scores in different views vs number of iterations of co-trained DPC for Synthetic dataset

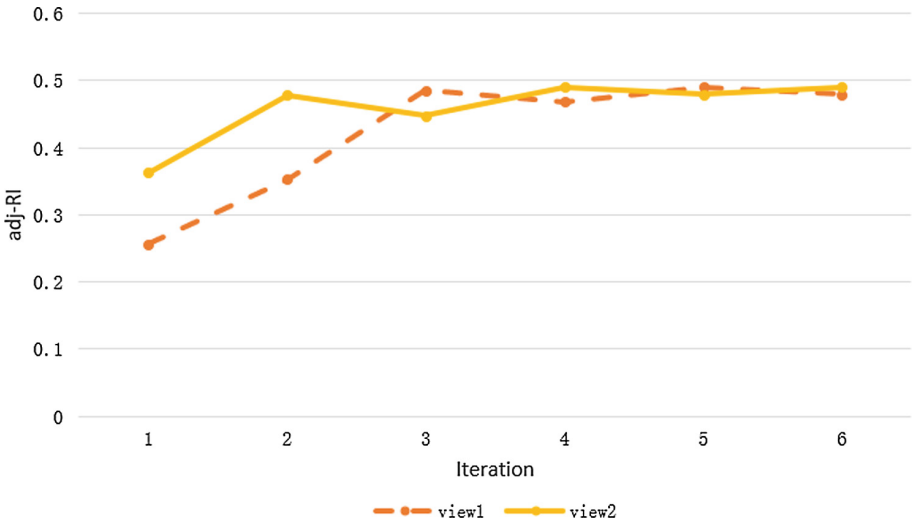


Fig. 4. adj-RI scores in different views vs number of iterations of co-trained DPC for MNIST dataset

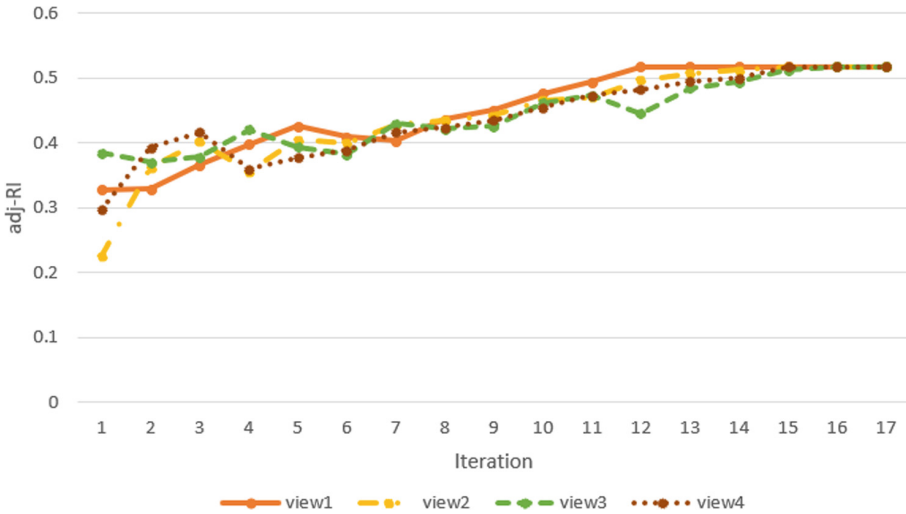


Fig. 5. adj-RI scores in different views vs number of iteration of co-trained DPC for IXMAS action dataset

5 Conclusion

We extend the original density peak clustering method from single-view learning to multi-view learning with the idea of co-training. In our adjusted co-training framework, distances between data points belonging to the same cluster decrease during iteration according to the clustering result for another view. In our adjusted density peak clustering method, cluster centers are selected simply, and then excessive clusters produced by the simple cluster center selection are merged according to densities of points in the border area of clusters. Based on these extensions, the co-trained density peak clustering method outperforms other baselines in experiments. The proposed approach has the ability to integrating information in views and avoiding degradation of performance through few steps of iteration.

References

1. Arthur, D., Vassilvitskii, S.: K-means++: the advantages of careful seeding. In: Eighteenth ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, pp. 1027–1035 (2007)
2. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 888–905 (2000)
3. Ester, M., Kriegel, H.P., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press (1996)
4. Rodriguez, A., Laio, A.: Machine learning. Clustering by fast search and find of density peaks. *Science* **344**(6191), 1492 (2014)

5. Sun, S.: Multi-view laplacian support vector machines. *Appl. Intell.* **41**(4), 209–222 (2013)
6. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)
7. Cohen, J., Cohen, P., West, S.G., et al.: *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd edn, pp. 227–229. L. Erlbaum Associates (2003)
8. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. *Publ. Am. Stat. Assoc.* **101**(476), 1730–1730 (2004)
9. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**(12), 2639–2664 (2014)
10. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. *Neural Comput.* **12**(6), 1247–1283 (2014)
11. Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) *SLSFS 2005*. LNCS, vol. 3940, pp. 34–51. Springer, Heidelberg (2006). https://doi.org/10.1007/11752790_2
12. Sharma, A., Jacobs, D.W.: Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In: *Computer Vision and Pattern Recognition*, pp. 593–600. IEEE (2011)
13. Sharma, A., Kumar, A., Daume, H., et al.: Generalized multiview analysis: a discriminative latent space. In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, pp. 2160–2167 (2012)
14. Sun, S., Xie, X., Yang, M.: Multiview uncorrelated discriminant analysis. *IEEE Trans. Cybern.* **46**(12), 3272 (2016)
15. Hotelling, H.: *Relations Between Two Sets of Variates*. *Breakthroughs in Statistics*, pp. 321–377. Springer, New York (1992)
16. Xu, C., Tao, D., Xu, C.: Multi-view intact space learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(12), 2531–2544 (2015)
17. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: *Eleventh Conference on Computational Learning Theory*, pp. 92–100. ACM (1998)
18. Kumar, A., Daumé III, H.: A co-training approach for multi-view spectral clustering. In: *International Conference on International Conference on Machine Learning*, pp. 393–400. Omnipress (2011)
19. Li, Y., Liu, W., Wang, Y., et al.: Co-spectral clustering based density peak. In: *IEEE International Conference on Communication Technology*, pp. 925–929. IEEE (2015)
20. Gao, H., Nie, F., Li, X., et al.: Multi-view subspace clustering. In: *IEEE International Conference on Computer Vision*, pp. 4238–4246. IEEE (2016)
21. Cortes, C., Mohri, M., Rostamizadeh, A.: Learning non-linear combinations of kernels. In: *International Conference on Neural Information Processing Systems*, pp. 396–404. Curran Associates Inc. (2009)
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, pp. 886–893. IEEE (2005)
23. Manning, C.D., Raghavan, P., Schütze, H.: An introduction to information retrieval. *J. Am. Soc. Inf. Sci. Technol.* **61**(4), 852–853 (2008)
24. Hubert, L., Arabie, P.: Comparing partitions. *J. Classif.* **2**(1), 193–218 (1985). Assortative pairing and life history strategy - a cross-cultural study. *Hum. Nat.* **20**, 317–330



Boosting Sparsity-Induced Autoencoder: A Novel Sparse Feature Ensemble Learning for Image Classification

Rui Shi, Jian Ji^(✉), Chunhui Zhang, and Qiguang Miao

School of Computer Science and Technology, Xidian University,
Xi'an 710071, China
jji@xidian.edu.cn

Abstract. As a model of unsupervised learning, autoencoder is often employed to perform the pre-training of the deep neural networks. However, autoencoder and its variants have not taken the statistical characteristics and the domain knowledge of training set into the design of deep neural networks and have abandoned a lot of features learned from different levels at the pre-training process. In this paper, we propose a novel sparse feature ensemble learning method for natural image classification, named boosting sparsity-induced autoencoder, to fully utilize hierarchical and diverse features. Firstly, a sparsity encourage method is introduced by adding an extra sparsity-induced layer to exploit the representative and intrinsic features of the input. And then, the ensemble learning is taken into consideration of the construction of the model to improve and boost the accuracy and stability of a single model. The classification results on three datasets demonstrate the effectiveness of the proposed method.

Keywords: Sparse representation · Sparsity-induced method
Ensemble learning · Image classification

1 Introduction

The performance of a generic learning algorithm, especially adopted to the classification problem, extremely relies on the quality of learned feature representation of raw input data. Good features not only could remove irrelevant or redundant features coexisting in the original input space, but preserve the essential information for the target tasks. A good feature extractor built for input space, especially using unsupervised learning methods, can be further utilized for computer vision tasks. Deep hierarchical features produced by stacked unsupervised models have been demonstrated to be a powerful tool and appeal to emerging focus [1, 2].

In recent years, the study found that deep learning constructed by the multiple non-linear transformations can be a powerful feature learning tool. Deep learning has already been broadly used to address image classification tasks [3–6]. As a tool of deep learning with a special architecture, the autoencoder has already been stacked to pre-train a deep neural network using a greedy layer-wise means [7], where each layer is separately initialized by unsupervised pre-training method, and then a fine-tuning way

based on backpropagation is used by a supervised learning algorithm [8, 9], leading to solving the lack of expression ability of shallow network.

By restricting the output of the model identical to the input data, autoencoder can be regarded as an identity function which could reconstruct the raw input data composed of an encoding phase and a decoding phase. Meanwhile, sparse representation has proven its significant impact on computer vision [10–12]. The performance of an image classifier can be improved if the input image can be represented by a sparse representation. Ghifary [12] demonstrated that, in most cases, sparse network structures have better classification performance than dense structures. In recent years, the sparse deep model is proposed based on the sparse encoding strategy, sparse regularization term and sparse filtering that have taken the input samples into sparse depth related neural network model.

However, autoencoder and its variants have not taken the statistical characteristics and the domain knowledge of training set into the design of deep networks, and they have abandoned a lot of features learned from different levels. Therefore, autoencoder can only provide a relatively coarse parameters setting and serves as a pre-training method because of the large variance and low generalization ability on the unknown testing dataset. So, how to fully utilize the features existed in the input is one of the most important points in our work. It is well known that an ensemble of multiple classifiers is considered as a practical technique for improving accuracy and stability with comparisons to a single classifier. Ensemble learning employs some weak classifiers, according to some combination rule, to construct a stronger one to obtain significantly reduced generalization error than any weak one. But, two key issues, namely the diversity and accuracy of each classifier and the combination rules of fusion rules [13], are required to be taken into consideration to ensure a better performance.

In this paper, we introduce a novel sparsity-induced autoencoder that can further exploit the representative and intrinsic features of the input. Then, to benefit the ability of the ensemble learning, an ensemble sparse feature learning algorithm based on the novel sparse autoencoder mentioned above, named BoostingAE, is proposed. On the one hand, the completion of the pre-training sparsity-induced autoencoder can obtain a plurality of different levels of abstraction of sparse features; on the other hand, ensemble learning could effectively improve and enhance the recognition rate and stability of single classifier. Experimental results on three different datasets show that the proposed ensemble feature learning method can significantly improve the overall performance.

2 Related Work

2.1 Sparse Representation

Sparse Coding. Sparse coding provides a family of methods for acquiring the condense features in the input. Given only the unlabeled dataset, it can discover the basic functions aimed to capture the higher-level features in the data itself. Despite its close relationship to the traditional sparse coding techniques on image denoising, the main

drawback of sparse coding is its high computation cost. Moreover, it is well-known that the sparse coding is not “smooth” [14, 15], which means a tiny variation in input space might result in a significant difference in code space.

Sparse Filtering [16]. In contrast to many existing feature learning models, one of the important properties of sparse filtering is that it only requires one hyper-parameter rather than extensive hyper-parameters tuning for its very simple cost function:

$$\min \sum_{i=1}^M \|\hat{f}^{(i)}\|_1 = \sum_{i=1}^M \left\| \frac{\tilde{f}^{(i)}}{\|\tilde{f}^{(i)}\|_2} \right\|_1 \tag{1}$$

where f represents the learned feature value for input sample, \tilde{f} is defined by ℓ_2 norm of f , and M indicates the sample’s number.

Sparse Regularization. Compared with sparse coding, sparse regularization needs to perform an extra separate stage to induce sparsity and encourage sparse representations of input. Various methods of sparsity regularization either employed in deep belief network or autoencoder [17], similar to sparse coding, each of which has been proved the beneficial effects for some particular scene.

2.2 Softmax Regression

Softmax regression is a generalized version of logistic regression applied to classification problems where the class label y can be chosen from more than two values. Assume that there are k labels and m training samples: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(i)}, y^{(i)}), \dots, (x^{(m)}, y^{(m)})\}$ ($i = 1, 2, \dots, m$), where x is the input sample, and $y \in \{1, 2, \dots, k\}$ is the corresponding label.

For every input, the output probability function can be defined as follows:

$$h_{\theta}(x^i) = \begin{bmatrix} p(y^{(i)} = 1|x^{(i)}; \theta) \\ p(y^{(i)} = 2|x^{(i)}; \theta) \\ \vdots \\ p(y^{(i)} = k|x^{(i)}; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^k e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_j^T x^{(i)}} \end{bmatrix} \tag{2}$$

where θ is the parameter of the softmax model. For each input, the probability of its category is estimated to be:

$$p(y = k|x^{(i)}; \theta) = \frac{e^{\theta_k^T x^{(i)}}}{\sum_k e^{\theta_k^T x^{(i)}}} \tag{3}$$

2.3 Ensemble Learning

According to certain combination rule, ensemble learning employs some weak classifiers to construct a stronger one to obtain significantly reduced generalization error

than any weak one. Weak learner refers to whose generalization performance on the unknown testing dataset is only slightly better than random guessing. From mathematics, ensemble learning can significantly reduce the variance to achieve more stable performance. In order to get a better integration result, it is necessary to make the individual learner as different as possible, that is to say, there is a high degree of diversity between the base learners, which will be helpful to the performance of ensemble learning.

Boosting method is a widely used method for statistical learning, and serves as an important means of ensemble learning. By changing the weights of training samples, boosting method trains a group of individual learners and gets final decision results with a combination rule of voting.

3 Boosting Sparsity-Induced Autoencoder

To learn more representative and intrinsic features of input, a novel sparsity encourage method is first introduced to build a new autoencoder, called sparsity-induced autoencoder (SparsityAE). Based on SparsityAE and ensemble learning, we further proposed a boosting sparsity-induced autoencoder (BoostingAE), which is capable of utilizing the hierarchical and diverse features, ensuring the accuracy and diversity, and boosting the performance of the single SparsityAE on computer vision tasks.

3.1 Sparsity-Induced Autoencoder

Inspired by the assumptions of the sparse representation and the efficient reconstruction of low-dimension feature representation obtained in the encoding phase of deep models, SparsityAE is proposed, whose structure is shown in Fig. 1.

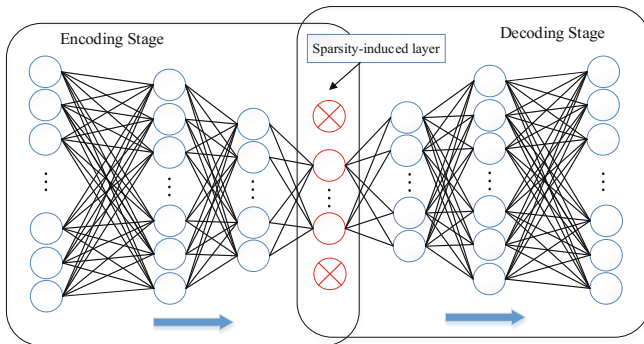


Fig. 1. The topology of the proposed sparsity-induced autoencoder

We feed the encoder by the dataset of high-dimension space. The length of the codes learned by each layer gets less along with the deepening of the encoder. At the end of the encoding phase, we employ a sparsity-induced layer to generate sparse

codes. Conversely, the decoding phase deals with the compressed and sparse codes given by sparsity-induced layer. In sparsity-induced layer, the neurons without significant activation value will be set to zero, which could decrease the number of neurons, remove the correlation between attributes and compress the raw inputs.

Let $y_i (i = 1, 2, \dots, N)$ be the original data and x_i be its degraded version, so the input can be mapped to a hidden representation by the formulations as follows:

$$\hat{y}(x_i) = \sigma(W'h(x_i) + b') \tag{4}$$

where \hat{y}_i is an approximation of y_i , and $\sigma(\bullet)$ is the mapping function.

To benefit both from the virtues of sparse representation and deep neural networks, we optimize the reconstruction loss regularized by a weight decay and a sparsity-inducing term. The cost function can be designed as follows:

$$L(X, Y; \theta) = \|y_i - \hat{y}(x_i)\|_2^2 / N + \beta \bullet KL(\hat{\rho} || \rho) \tag{5}$$

where $\theta = (W, b, W', b')$ represents weights and bias, $KL(\hat{\rho} || \rho)$ is the sparse regularization to extract sparse representation, and $\hat{\rho}$ is the average output of hidden neurons:

$$KL(\hat{\rho} || \rho) = \sum_{j=1}^{|\hat{\rho}|} \rho \log\left(\frac{\rho}{\hat{\rho}_j}\right) + (1 - \rho) \log\left(\frac{1 - \rho}{1 - \hat{\rho}_j}\right) \tag{6}$$

$$\hat{\rho} = (1/N) \sum_i^N h(x_i) \tag{7}$$

In the training process, for the intractability of the whole image, the model is provided with the original overlapping patches $y_i (i = 1, 2, \dots, N)$ as the reconstruction, and their corrupted image patches x_i as the polluted input. As long as the training is completed, the learned model could reconstruct the corresponding clean image given any polluted observation. The detailed process is shown in Algorithm 1.

Algorithm 1 SparsityAE

Notation: Ω_i is reconstruction error for x_i , S_i is the reconstruction coefficient for x_i , and $\{y_i\}_1^N$ is the hidden representation for every input.

Input: training set $D = \{x_i\}_1^N$, parameters k (constant) and $\theta = \{W, b, W', b'\}$.

Process:

- (1) Compute the S_i for each input $\{x_i\}_1^N$;
- (2) Minimize the cost function by the stochastic gradient descent and update θ ;
- (3) Compute the hidden representation $\{y_i\}_1^N$ for each input, keep the k biggest activation value and others are set to zero, and update S_i and Ω_i ;
- (4) Repeat the step (2) and (3) until convergence.

Output: reconstruction representation of the input.

3.2 Feature Ensemble Method

Multiple sparse features with different abstraction levels will be obtained using the SparsityAE introduced above; ensemble feature learning could effectively improve the accuracy and stability of a single classifier. Together two points above, a BoostingAE is proposed that uses hierarchical feature obtained in pre-training stage to train multiple classifiers, and integrates the outputs of classifiers with specific fusion rules to get the final prediction of image classification.

To make the whole structure easy to understand, Fig. 2 gives a clear and detailed understanding of BoostingAE, which indicates that by cascading multiple SparsityAEs, BoostingAE is theoretically possible to obtain N compressed sparse features derived from the output of SparsityAEs.

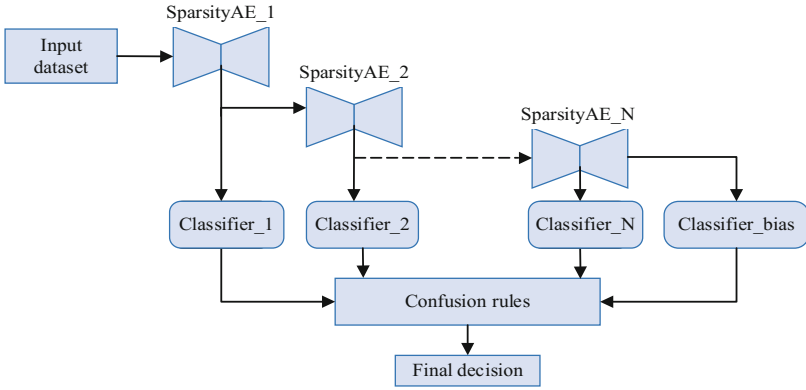


Fig. 2. The topology of the proposed BoostingAE

In our work, training three SparsityAEs and utilizing softmax regression to perform the classification task. First, we train SparsityAE_1 in Fig. 2. Assuming that the original input sample is x , the weight matrix connecting the input layer and the hidden layer is $W^{(1)}$, the bias vector is $b^{(1)}$, so its output can be mapped by Eq. (4):

$$\hat{y}_1 = \sigma(x \bullet W^{(1)} + b^{(1)}) \quad (8)$$

Regard \hat{y}_1 as the input of the second SparsityAE, thus further train SparsityAE_2. Its weight matrix connecting the input layer and the hidden layer is $W^{(2)}$, and bias vector is $b^{(2)}$. With reference to the above operation, SparsityAE_2's output can be further obtained, which also be used as the input of next SparsityAE:

$$\hat{y}_2 = \sigma(\hat{y}_1 \bullet W^{(2)} + b^{(2)}) \quad (9)$$

Along with the cascaded sparsity-induced autoencoder network, the characteristic attributes which are trained from the current layer will be passed to the next layer by

above process, and therefore three SparsityAEs can be trained. At the same time, in the longitudinal direction, the trained classifier model and optimal parameters of each classifier are obtained by training the characteristic attribute at the encoding stage. Further, three classifiers are obtained.

3.3 Combination Method of Voting

After training all base classifiers, final prediction is given by results of three classifiers after integrating with some fusion rules. Here, the Naïve Bayes combination rules [18] are applied which assume that individual classifiers are mutually independent.

We adopt three Naïve Bayes combination methods, namely MAX, MIN, and AVG rules. Given a sample x , and its label y has C possible values. Assuming that the current BoostingAE model consists of N base classifiers, $P_{nj}(x)$ is the probability that the category of x is j in the n th classifier. So label y can be defined as follows:

- MAX rule: $y = \arg \max_{j=1,2,\dots,C} \max_{n=1,2,\dots,N} P_{nj}(x)$,
- MIN rule: $y = \arg \max_{j=1,2,\dots,C} \min_{n=1,2,\dots,N} P_{nj}(x)$,
- AVG rule: $y = \arg \max_{j=1,2,\dots,C} \frac{1}{N} \sum_{n=1}^N P_{nj}(x)$.

3.4 BoostingAE Algorithm

From Fig. 2, a multi-layer architecture based on ensemble learning consists of an input layer, some hidden layers and an output layer to carry out specific tasks.

Here, how to measure the importance of each layer's feature and how to select the optimal models for each layer are two key issues, which will directly influence the performance of the model. As a main contribution of this paper, we employ Adaboost to supervise the adjustment of parameters and weight coefficients. Algorithm 2 gives the detailed process of the proposed BoostingAE.

Algorithm 2 BoostingAE

Notation: T is the number of base sparse autoencoder, and the algorithm of base learner is SparsityAE.

Input: training set $D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$.

Process:

Initialize: assign the default parameters.

Loop: for $k = 1, 2, \dots, k$

Generate the new dataset D' by bootstrapping the original dataset;

Obtain the trained individual classifier, $g_k = \text{SparsityAE}(D')$.

Output: Obtain ensemble classifier $D' = \text{Uniform}(\{g_k\})$.

Compared with traditional sparse stacked autoencoder, BoostingAE’s characteristics are shown in following aspects:

- whose construction of base learners is similar to that of AdaBoost, BoostingAE utilizes the cascade serialization mechanism among the base learners, which makes the individual learners are related to each other and also maintain the difference;
- the subsequent layer takes the output of the previous layer as the input to obtain rich feature representation, which makes each learners receive various “training input” at the same time and avoids the waste of computing and storage resources;
- when design individual learners, the topology of each model can be specified separately rather than by a unified model topology. This makes it possible to further increase the diversity of base learners while maintaining the homogeneity of them.

4 Experiments

First of all, we verify the performance of SparsityAE in sparse feature learning. Next, to unbiasedly and accurately show the performance and the stability of BoostingAE on real-world image classification, the experiments are carried out on three widely employed datasets, i.e., MNIST, CIFAR-10, and SVHN. Moreover, some state-of-art methods are employed to provide the comparable results on the same datasets.

4.1 The Sparse Feature Learning of SparsityAE

To validate the performance of SparsityAE in sparse feature learning, we mainly focus on denoising of grey-scale images. From <http://decsai.ugr.es/cvg/dbimagenes>, a set of natural images are employed as the training set, and a set of standard natural images as the testing set which has been widely used in the image processing.

When it comes to the training process, we randomly pick a clean image y from the dataset and generate its corresponding noisy patch x by corrupting it with a specific strength of additive white Gaussian noise. The training performed, the learned model will be capable of reconstructing the corresponding clean image given any noisy observation. To avoid the local minimum, we adopt the layer-wise pre-training procedure introduced in [7].

Figure 3 shows the comparison between SparsityAE and classic image denoising methods: KSVD [19], BM3D [20], on standard testing images degraded by various noise levels. We tell that when $\sigma = 25$, SparsityAE (magenta line) is competitive, while corrupts for other different, i.e., higher noise strengthens, which is owing to that our model knowing nothing about the noise of level but other methods were provided with such information. The green line shows that if we train the proposed model on several different noise levels, our SparsityAE is more robust to the change of noise levels which means that it can generalize significantly better to higher noise levels.

What’s more, we compared the SparsityAE with several state-of-art denoising methods: WNNM [21] and two training based methods: MLP [22], TNRD [23]. The numerical results are shown in Table 1, which is measured by the peak signal to noise ratio (PSNR in dB). The best PSNR result for each image is highlighted in bold.

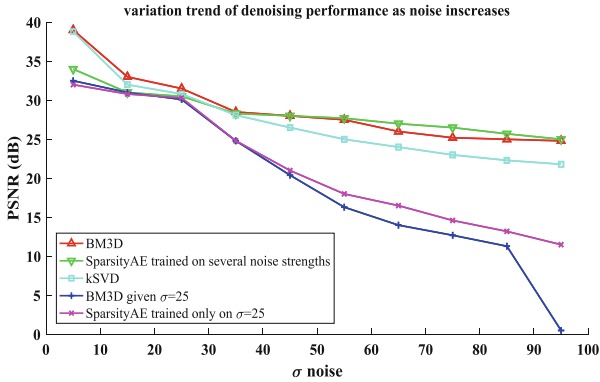


Fig. 3. Denoising performance comparison of various methods with various noise levels (Color figure online)

Table 1. Comparison of the various methods' denoising performance measured by PSNR.

Image	KSVD	BM3D	WNNM	MLP	TNRD	Ours
Lena	31.35	32.08	32.24	32.25	32.00	32.46
House	32.14	32.86	33.23	32.56	32.53	32.96
C.man	28.72	29.45	29.64	29.61	29.72	29.98
Moarch	28.81	29.25	29.85	29.61	29.85	29.65
Couple	28.84	29.71	29.82	29.76	29.69	29.68
Man	29.09	29.61	29.76	29.88	30.11	29.78
Babara	29.60	30.72	31.24	29.54	29.41	29.20
Boat	29.32	29.91	30.03	29.97	30.21	29.91
Pepper	29.71	30.16	30.42	30.30	30.57	30.37

Although images with a lot of repeating structure are ideal for both KSVD and BM3D, we do outperform KSVD, BM3D, and WNNM on every image except Barbara. It is also shown that our SparsityAE is able to compete with MLP and TNPd.

The results illustrated that SparsityAE can not only project the original high dimensional space to a lower dimensional and more intrinsic space from the perspective of dimension-reduction, but capture the more representative sparse feature from multiple layers to make the best use of the information contained in original space.

4.2 BoostingAE for Classification on MNIST

MNIST is a large dataset of handwritten digits that is widely used for image processing and computer vision tasks. It contains 60,000 training images and 10,000 testing images with labels, and the size of a single image is 28×28 .

The topology of the SparsityAE on MNIST with three hidden layers is first determined, i.e., $784 \times 500 \times 250 \times 100 \times 10$, which 784 is the size of the image and

10 is the label number. Then, Classifier_1 in Fig. 2 is obtained by pre-training and fine-tuning SparsityAE_1. After that, SparsityAE_2 takes the feature representations learned from SparsityAE_1 as input to get Classifier_2. With the process above, we get the final three base classifiers that will be integrated when all base learners achieve convergence after fine-tuning. With this, the whole BoostingAE model is constructed and trained completely. When it comes to predicting the real samples, three Naïve Bayes combination rules will be respectively used for voting the integrative result of three classifiers to improve the performance.

Table 2 shows that three individual classifiers have better classification results than KNN and SVM because of the introduction of the sparsity-induced layer in SparsityAE. And the BoostingAE with three different fusion rules gets better performance than any individual classifier and achieves 98.37%, 98.43% and 98.87% accuracy rate respectively, which is very close to the result of L_p -norm AE [24]. Moreover, stacked CAE [25] and CASE [26] employ more feature maps obtained by convolutional operations and hidden layers, so our performance is slightly worse than these.

Table 2. Classification results on three datasets.

	MNIST	CIFAR-10	SVHN
KNN	91.32%	84.47%	78.32%
SVM	94.02%	88.45%	83.24%
L_p -norm AE(KNN) [24]	97.44%	/	67.23%
L_p -norm AE(SVM) [24]	98.64%	/	71.19%
Stacked CAE [25]	99.29%	79.20%	/
CSAE [26]	99.39%	/	/
CDSA [27]	/	74.18%	/
Classifier_1	96.73%	91.46%	88.46%
Classifier_2	96.35%	91.83%	88.93%
Classifier_3	95.89%	90.96%	87.69%
BoostingAE(MAX)	98.37%	92.32%	89.94%
BoostingAE(MIN)	98.43%	91.49%	90.35%
BoostingAE(AVG)	98.87%	92.63%	90.87%

4.3 BoostingAE for Classification on CIFAR-10 and SVHN

CIFAR-10 is a dataset contains ten kinds of color images, each category contains 6000 color images. The training set contains 5000 images of each category, the remaining is used for testing. SVHN dataset can be regarded as the upgrade of MNIST and also contains ten kinds of color images. Both are captured from the real life so the background is more complex and the images are difficult to identify. SVHN is divided into training set, testing set and extra set; the validation set is constructed in a random way: the 2/3 of them is derived from the training set (400 samples per class), and the remaining samples come from the extra set (200 samples per class).

Before the experiment, the original images of CIFAR-10 and SVHN should be transformed from RGB space into grey space, and then normalized. To improve the training efficiency, the mini-batch gradient descent algorithm is used when pre-training and fine-tuning. Considering the unsupervised learning mechanism of autoencoder, both CIFAR-10 and SVHN use a certain proportion of unlabeled samples as training set in pre-training; and in the process of fine-tuning, two datasets require ground-truth to implement the classification. Next, the topology of SparsityAE is determined as $1024 \times 500 \times 250 \times 100 \times 10$. The subsequent operations are similar to those performed on MNIST.

We report the results of comparison methods, individual classifiers and the proposed method in Table 2 and get the similar conclusion as MNIST. Our methods achieve the best results among comparison methods. The results illustrated the BoostingAE could capture more sparse representation and utilize multi-layer features, resulting in the improvement of accuracy and diversity of overall.

5 Conclusion

In this work, we first built SparsityAE by adding an extra sparsity-induced layer, which efficiently abstract the sparse feature representations, and then based on SparsityAE and ensemble learning, we further proposed a BoostingAE model to integrate sparse feature learned from multi-layer, so as to improve the performance of individual sparse encoder, which has been successfully applied to image classification.

The main advantage of our approach is that it could abstract more significantly sparse representations that reflect the distribution of original data better and make full use of the features learned from multi-layer to improve the diversity of base learners. What's more, it also promotes the overall performance after integrating multiple weak learners. Additional experiments on three different datasets validate the effectiveness of the proposed algorithm in image classification.

References

1. Goh, H., Thome, N., Cord, M., Lim, J.H.: Learning deep hierarchical visual feature coding. *IEEE Trans. Neural Netw. Learn. Syst.* **25**(12), 2212–2225 (2014)
2. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: *IEEE International Conference on Computer Vision*, pp. 3074–3082. IEEE Computer Society Press, Santiago (2015)
3. Yang, X., Ye, W., Li, X., Lau, R.Y.K., Zhang, X., Huang, X.: Hyperspectral image classification with deep learning models. *IEEE Trans. Geosci. Remote Sens.* **56**(9), 1–16 (2018)
4. Mei, S., Jiang, R., Ji, J., Sun, J., Peng, Y.: Invariant feature extraction for image classification via multi-channel convolutional neural network. In: *International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 491–495. IEEE, Japan (2018)

5. Durand, T., Mordan, T., Thome, N., Cord, M.: WILDCAT: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 642–651. IEEE Computer Society, Hawaii (2017)
6. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: using knowledge graphs for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2673–2681. IEEE Computer Society Press, Hawaii (2017)
7. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: International conference on Neural Information Processing Systems, pp. 153–160. MIT Press, Vancouver (2006)
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006)
9. Wen, W., Wu, C., Wang, Y., Chen, Y., Li, H.: Learning structured sparsity in deep neural networks. In: International conference on Neural Information Processing Systems, pp. 2074–2082. MIT Press, Barcelona (2016)
10. Shahnawazuddin, S., Sinha, R.: Sparse coding over redundant dictionaries for fast adaptation of speech recognition system. *Comput. Speech Lang.* **43**, 1–17 (2017)
11. Srinivas, M., Lin, Y., Liao, H.Y.M.: Learning deep and sparse feature representation for fine-grained object recognition. In: IEEE International Conference on Multimedia and Expo, pp. 1458–1463. IEEE Press, Hong Kong (2017)
12. Ghifary, M., Kleijn, W.B., Zhang, M.: Sparse representations in deep learning for noise-robust digit classification. In: International Conference on Image and Vision Computing New Zealand, pp. 340–345. IEEE Press, Wellington (2013)
13. Zhang, L., Zhou, W.: Sparse ensembles using weighted combination methods based on linear programming. *Pattern Recogn.* **44**(1), 97–106 (2011)
14. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360–3367. IEEE Press, San Francisco (2010)
15. Gao, S., Tsang, I.W., Chia, L., Zhao, P.: Local features are not Lonely–Laplacian sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3555–3561. IEEE Computer Society Press, San Francisco (2010)
16. Ngiam, J., Koh, P.W., Chen, Z., Bhaskar, S., Ng, A.Y.: Sparse filtering. In: International Conference on Neural Information Processing Systems, pp. 1125–1133. MIT Press, Granada (2011)
17. Le, Q.V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Ng, A.Y.: On optimization methods for deep learning. In: 28th International Conference on Machine Learning, pp. 265–272. Omnipress, Bellevue (2011)
18. Kuncheva, L.I.: *Combining Pattern Classifiers: Methods and Algorithms*, 1st edn. Wiley, Hoboken (2004)
19. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.* **15**(12), 3736–3745 (2016)
20. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. Image Process.* **16**(8), 2080–2095 (2007)
21. Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2862–2869. IEEE Computer Society Press, Columbus (2014)
22. Burger, H.C., Schuler, C.J., Harmeling, S.: Image denoising: can plain neural networks compete with BM3D? In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2392–2399. IEEE Computer Society Press, Providence (2012)

23. Chen, Y., Pock, T.: Trainable nonlinear reaction diffusion: a flexible framework for fast and effective image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1256–1272 (2017)
24. Mehta, J., Gupta, K., Gogna, A., Majumdar, A., Anand, S.: Stacked robust autoencoder for classification. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D. (eds.) *ICONIP 2016*. LNCS, vol. 9949, pp. 600–607. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46675-0_66
25. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) *ICANN 2011*. LNCS, vol. 6791, pp. 52–59. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_7
26. Luo, W., Li, J., Yang, J.: Convolutional sparse autoencoders for image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**(99), 1–6 (2017)
27. Chen, S., Liu, H., Zheng, X., Qian, S., Yu, J., Guo, W.: Image classification based on convolutional denoising sparse autoencoder. *Math. Probl. Eng.* **2017**, 1–16 (2017)



Matrix-Instance-Based One-Pass AUC Optimization

Changming Zhu¹(✉) , Chengjiu Mei¹, Hui Jiang², and Rigui Zhou¹

¹ College of Information Engineering, Shanghai Maritime University,
Shanghai 201306, People's Republic of China
{cmzhu,rgzhou}@shmtu.edu.cn, 544306495@qq.com

² School of Mechanical Engineering,
University of Shanghai for Science and Technology,
Shanghai 200093, People's Republic of China
huijiang@hotmail.com

Abstract. Area under the receiver operating characteristic curve, i.e., AUC, is a widely used performance measure. Traditional off-line and some online AUC optimization methods should store the entire or part of dataset in memory which is infeasible to process big data or streaming data applications. So some scholars develop one-pass AUC optimization (OPAUC) which is independent from the data size. While OPAUC cannot process matrix instances. So we propose a matrix-instance-based one-pass AUC optimization model, i.e., MOPAUC, to overcome such an issue. Related experiments on some benchmark datasets including five image datasets validate that MOPAUC can improve the average AUC, cost little running time with matrix-instance cases. Furthermore, some parameters including regularization parameters and weights have less influence on the average AUC while step sizes have strong influence.

Keywords: One-pass · Matrix instance · AUC

1 Introduction

1.1 Background

As we all know, the area under the receiver operating characteristic (ROC) curve (i.e., AUC) is an important performance measure and it has been widely used in many tasks [1–5]. According to [6] said, AUC is measured by the losses defined over pairs of instances from different classes which is different from the classical classification and regression problems where the loss function can be gotten by a single training instance. In present, during the procedure of design, many classifiers demand the AUC be maximization [7–9]. Thus, the optimization of AUC is a hot spot of present research. The traditional AUC optimization methods include semi-supervised learning receiver operating characteristic (SSLROC) algorithms which utilize unlabeled test instances in classifier training to maximize AUC

[10], direct-AUC which is a boosting method to directly optimizes AUC value as a classification performance measure [11], semi-supervised AUC optimization method with generative models (OptAG) which utilizes generative models to assist the incorporation of unlabeled instances in AUC-optimized classifiers [12]. While all those traditional AUC optimization methods exist two defects. One is that those off-line AUC optimization methods [2, 8, 10–12] need to store the entire dataset in memory before an optimization procedure is applied while this is infeasible for applications involving big data or streaming data in which a large volume of data come in a short time period. The other is that for some online AUC optimization methods [1, 7, 9], they find the optimal solution of some performance measures by only scanning the training data once, but these methods still need to store \sqrt{T} instances where T is the number of training instances.

1.2 Proposal

As [6] said, a good AUC optimization method (i.e., one-pass AUC optimization) should be independent from the number of training instances since it is always difficult to expect how many data will be received in the applications. Until now, only few scholars pay attention to one-pass AUC optimization problems. To the best of our knowledge, work [13] is the extended work of [6] which aims to process one-pass AUC optimization and except the scholars of [6] and [13], we have not found any other scholars to pay attention to this field.

Moreover, it is found that in [6] and [13], the used datasets consist vector instances, i.e., each instance $x \in \mathbb{R}^{d \times 1}$ is a d -dimensionality one. This representation can bring a convenience in mathematics. But as we know, in real world applications, more and more instances are represented in matrix form, i.e., a matrix instance $A \in \mathbb{R}^{m \times n}$ and its dimensionality is $m \times n$. Classical matrix datasets include images. Since the model named one-pass AUC (OPAUC) which is developed by [6] and [13] cannot process matrix instances, thus this paper will develop a matrix-instance-based one-pass AUC optimization model, i.e., MOPAUC, so as to process the matrix datasets.

1.3 Difficulty

As we said, MOPAUC can process matrix instances. While the difficulty that extends OPAUC to handle matrix instances is obvious. Once we extend the model of vector-instance-based learning machine to the one of the matrix-instance-based learning machine, we should optimize more parameters due to for a matrix instance, a more classifier weight is needed. How to optimize them is the difficulty which should be conquered. Thus, in our work, in order to solve this difficulty, we adopt gradient descent method and details are given in Sect. 2.

1.4 Contribution and Framework of the Manuscript

The contributions of the MOPAUC are (1) it can process the matrix-instance-based AUC optimization problems; (2) compared with the OPAUC whose

required storage is $O(d^2)$ where $d = m \times n$, the storage requirement is reduced to $O(m^2 + n^2)$; (3) it inherits the advantage of OPAUC which is independent from the number of training instances.

What's more, Sect. 2 shows the framework of the developed MOPAUC. Section 3 gives the experiments. The conclusion is given in Sect. 4.

2 Matrix-Instance-Based One-Pass AUC Optimization

There is a matrix instance $A \in \mathbb{R}^{m \times n}$ and its dimensionality is $m \times n$. The class label of each instance is selected from the set $y = \{+1, -1\}$. Here, the instances form the instance space \mathcal{A} while the labels form the label space \mathcal{Y} . Denote \mathcal{D} by an unknown distribution over the product space $\mathcal{A} \times \mathcal{Y}$. Let $\mathcal{S} = \{(A_1, y_1), (A_2, y_2), \dots, (A_T, y_T)\}$ be a series of instances which arrive continuously and each instance arrives identically and independently from \mathcal{D} . Moreover, we denote $[n] = \{1, 2, \dots, n\}$ where the integer $n > 0$ and $\lfloor \alpha \rfloor$ represents the largest integer which is no more than α where the real $\alpha > 0$. Then we adopt $|\mathcal{A}|$ to denote its cardinality.

Now we let $f : \mathcal{A} \rightarrow \mathbb{R}$ be a real-valued function, and for \mathcal{S} , the AUC of function f is defined as:

$$AUC(f, \mathcal{S}) = \sum_{i=1}^T \sum_{j=1}^T B. \quad (1)$$

where $B = \frac{(\prod [f(A_i) > f(A_j)] + \frac{1}{2} \prod [f(A_i) = f(A_j)]) \prod [y_i > y_j]}{T_s^+ T_s^-}$, $\prod[\star]$ is the indicator function which returns 1 if the argument is true and 0 otherwise, $T_s^+ = |\{(A_i, y_i) \in \mathcal{S} : y_i = +1\}|$ and $T_s^- = |\{(A_i, y_i) \in \mathcal{S} : y_i = -1\}|$.

Then the optimization of AUC can be turned to optimize the pairwise surrogate losses as follows:

$$\begin{aligned} \mathcal{L}(f, \mathcal{S}) &= \sum_{i=1}^T \sum_{j=1}^T \frac{\ell(f(A_i) - f(A_j)) \prod [y_i > y_j]}{T_s^+ T_s^-} \\ &= \sum_{i=1}^T \sum_{j=1}^{i-1} \frac{\ell(y_i(f(A_i) - f(A_j))) \prod [y_i \neq y_j]}{T_s^+ T_s^-}. \end{aligned} \quad (2)$$

where $\ell : \mathbb{R} \rightarrow \mathbb{R}^+$ is a convex function. Then we say in the \mathcal{D} , the loss can be computed as bellow.

$$\begin{aligned} \mathcal{L}(f, \mathcal{D}) &= E_{A_i \sim \mathcal{D}^+, A_j \sim \mathcal{D}^-} [\ell(f(A_i) - f(A_j))] \\ &= E_{(A_i, y_i) \sim \mathcal{D}, (A_j, y_j) \sim \mathcal{D}} [\ell(f(A_i) - f(A_j)) | y_i > y_j]. \end{aligned} \quad (3)$$

In order to optimize the Eq. (3) in convenience, we let $\ell(t) = (1 - t)^2$. Then for \mathcal{S} , its pairwise least square loss is given below.

$$\mathcal{L}(u, v, \mathcal{S}) = \frac{\lambda_1}{2} |u|^2 + \frac{\lambda_2}{2} |v|^2 + \frac{1}{2} \sum_{i=1}^T \sum_{j=1}^{i-1} \frac{(1 - y_i u^T (A_i - A_j)^T v)^2}{T_s^+ T_s^-}. \quad (4)$$

where the weights are $u \in \mathbb{R}^{m \times 1}$ and $v \in \mathbb{R}^{n \times 1}$. λ_1 and λ_2 are regularization parameters that control the model complexity. The constant $\frac{1}{2}$ is introduced for simplicity. Moreover, we define that the pairwise least square loss with respect to distribution \mathcal{D} as

$$\mathcal{L}(u, v, \mathcal{D}) = E_{\mathcal{S}}[\mathcal{L}(u, v, \mathcal{S})] = \frac{1}{T} \sum_{t=1}^T \mathcal{L}_t(u, v). \quad (5)$$

where

$$\mathcal{L}_t(u, v) = \frac{\lambda_1}{2} |u|^2 + \frac{\lambda_2}{2} |v|^2 + \sum_{i=1}^{t-1} \frac{\prod[y_i \neq y_j](1 - y_i u^T (A_i - A_j)^T v)^2}{2|i \in [t-1] : y_i y_t = -1|}. \quad (6)$$

Now we can say the minimization of Eq. (6) is the equivalent problem to optimize the AUC of instances \mathcal{S} . In order to process this problem, we will adopt the gradient descent method and here, we define $\mathcal{L}_t(u, v) = 0$ when $T_t^+ T_t^- = 0$ where T_t^+ and T_t^- denote the cardinalities of positive and negative instances in $\mathcal{S}_t = \{(A_1, y_1), (A_2, y_2), \dots, (A_t, y_t)\}$, respectively.

If $y_i = +1$, the gradient of Eq. (6) is

$$\frac{\partial \mathcal{L}_t(u, v)}{\partial u} = \lambda_1 u - A_t v + c_t^- v + [(A_t - c_t^-) v v^T (A_t - c_t^-)^T] u + S_{t_u}^- u. \quad (7)$$

$$\frac{\partial \mathcal{L}_t(u, v)}{\partial v} = \lambda_2 v - A_t^T u + [c_t^-]^T u + [(A_t - c_t^-)^T u u^T (A_t - c_t^-)] v + S_{t_v}^- v. \quad (8)$$

where

$$c_t^- = \sum_{i:i < t, y_i = -1} \frac{A_i}{T_t^-}. \quad (9)$$

$$S_{t_u}^- = \sum_{i:i < t, y_i = -1} \frac{A_i v (A_i v)^T - (c_t^- v) (c_t^- v)^T}{T_t^-}. \quad (10)$$

$$S_{t_v}^- = \sum_{i:i < t, y_i = -1} \frac{(u^T A_i)^T u^T A_i - (u^T c_t^-)^T u^T c_t^-}{T_t^-}. \quad (11)$$

Otherwise, if $y_i = -1$, the gradient of Eq. (6) is

$$\frac{\partial \mathcal{L}_t(u, v)}{\partial u} = \lambda_1 u + A_t v - c_t^+ v + [(A_t - c_t^+) v v^T (A_t - c_t^+)^T] u + S_{t_u}^+ u. \quad (12)$$

$$\frac{\partial \mathcal{L}_t(u, v)}{\partial v} = \lambda_2 v + A_t^T u - [c_t^+]^T u + [(A_t - c_t^+)^T u u^T (A_t - c_t^+)] v + S_{t_v}^+ v. \quad (13)$$

where

$$c_t^+ = \sum_{i:i < t, y_i = +1} \frac{A_i}{T_t^+}. \quad (14)$$

$$S_{t_u}^+ = \sum_{i:i < t, y_i = +1} \frac{A_i v (A_i v)^T - (c_t^+ v) (c_t^+ v)^T}{T_t^+}. \quad (15)$$

$$S_{t_v}^+ = \sum_{i:i < t, y_i = +1} \frac{(u^T A_i)^T u^T A_i - (u^T c_t^+)^T u^T c_t^+}{T_t^+}. \quad (16)$$

Once we compute the gradient $\frac{\partial \mathcal{L}_t(u, v)}{\partial u}$ and $\frac{\partial \mathcal{L}_t(u, v)}{\partial v}$, we can update the classifier weights by

$$u_t = u_{t-1} - \eta_{ut} \frac{\partial \mathcal{L}_t(u_{t-1}, v_{t-1})}{\partial u_{t-1}}. \quad (17)$$

$$v_t = v_{t-1} - \eta_{vt} \frac{\partial \mathcal{L}_t(u_{t-1}, v_{t-1})}{\partial v_{t-1}}. \quad (18)$$

where $u_t(v_t)$ represents the $u(v)$ under t -th iteration. Here each iteration represents one instance arrives. η_{ut} and η_{vt} are the step sizes in the t -th iteration. During the procedure, we should notice that once a new instance arrives, the T_t^- , c_t^- , $S_{t_u}^-$, $S_{t_v}^-$, T_t^+ , c_t^+ , $S_{t_u}^+$, $S_{t_v}^+$ are also updated. If $y_t = -1$, we have

$$T_t^- = T_{t-1}^- + 1, T_t^+ = T_{t-1}^+, c_t^- = c_{t-1}^- + \frac{1}{T_t^-} (A_t - c_{t-1}^-), \quad (19)$$

$$\begin{aligned} c_t^+ &= c_{t-1}^+, S_{t_u}^+ = S_{t-1_u}^+, S_{t_v}^+ = S_{t-1_v}^+, \\ S_{t_u}^- &= S_{t-1_u}^- + c_{t-1}^- v (c_{t-1}^- v)^T - c_t^- v (c_t^- v)^T + \\ &\quad (A_t v (A_t v)^T - S_{t-1_u}^- - c_{t-1}^- v (c_{t-1}^- v)^T) / T_t^-, \\ S_{t_v}^- &= S_{t-1_v}^- + (u^T c_{t-1}^-)^T u^T c_{t-1}^- - (u^T c_t^-)^T u^T c_t^- + \\ &\quad ((u^T A_t)^T u^T A_t - S_{t-1_v}^- - u^T c_{t-1}^-)^T u^T c_{t-1}^- / T_t^-. \end{aligned}$$

Otherwise, if $y_t = +1$, we have

$$T_t^+ = T_{t-1}^+ + 1, T_t^- = T_{t-1}^-, c_t^+ = c_{t-1}^+ + \frac{1}{T_t^+} (A_t - c_{t-1}^+), \quad (20)$$

$$\begin{aligned} c_t^- &= c_{t-1}^-, S_{t_u}^- = S_{t-1_u}^-, S_{t_v}^- = S_{t-1_v}^-, \\ S_{t_u}^+ &= S_{t-1_u}^+ + c_{t-1}^+ v (c_{t-1}^+ v)^T - c_t^+ v (c_t^+ v)^T + \\ &\quad (A_t v (A_t v)^T - S_{t-1_u}^+ - c_{t-1}^+ v (c_{t-1}^+ v)^T) / T_t^+, \\ S_{t_v}^+ &= S_{t-1_v}^+ + (u^T c_{t-1}^+)^T u^T c_{t-1}^+ - (u^T c_t^+)^T u^T c_t^+ + \\ &\quad ((u^T A_t)^T u^T A_t - S_{t-1_v}^+ - u^T c_{t-1}^+)^T u^T c_{t-1}^+ / T_t^+. \end{aligned}$$

Once we get weights u_T and v_T , we can treat u_T and v_T as the weights of classifier with T instances arrive continuously in a short time. For convenience, we summary the algorithm in Table 1.

3 Experiments

3.1 Experiments on Benchmark Datasets

We conduct the experiments on 27 benchmark datasets [14] which can be found in Table 2. Since our MOPAUC can be used for matrix datasets, so we also adopt some image datasets for experiments, they are Coil-20, Letter-Image, ORL, CIFAR-10, and MNIST. For each dataset, we scale the features to $[-1, 1]$ and each multi-class dataset is transformed into a binary one by randomly partitioning classes into two groups, where each group contains the same or similar number of classes. Then in order to validate the effectiveness of MOPAUC, we can also reshape the vector instance into different matrix forms with the way given in [15] and select a feasible form for experiments. Contrariwise, if we conduct other AUC optimization methods which aim to process vector instances, we can vectorize the matrix instances to vector ones. Moreover, we adopt the following methods for comparison. Since some online and off-line AUC optimization methods has been compared in OPAUC [6, 13] and it has been validate that OPAUC outperforms those methods, especially some online ones including online AUC optimization with a sequential updating method or with a gradient descent updating method [16], online gradient descent algorithm which optimizes the (weighted) univariate exponential loss or optimizes the (weighted) univariate logistic loss or optimizes the (weighted) univariate least square loss [17], thus we compare some new methods including *OPAUC* (one-pass AUC [13]), *KOAUC* (kernel online AUC maximization [1]), *KOIL* (kernel online imbalanced learning with AUC [7]), *SSAUC_{GM}* (semi-supervised AUC optimization method with generative models [18]), *ELMAUC* (off-line binary AUC optimization algorithm [19]), *SVM_{pAUC}* (support vector algorithms for optimizing the partial area under the ROC curve [20]).

Among these methods, OPAUC, KOAUC, and KOIL are online ones and others are new off-line AUC optimization methods. Since *SSAUC_{GM}* is a semi-supervised method, so for the experiments about *SSAUC_{GM}*, each dataset is divided into two parts. We choose 30% instances in random as labeled instances and the rest is treated as the unlabeled part. Of course, as we know, more labeled instances brings a better classification performance. But according to our all experimental results which include those not written in this manuscript, even though we adopt 100% labeled instances, the performance of *SSAUC_{GM}* is still worse than the proposed MOPAUC in average. Thus, we only show the results when 30% instances are chosen in random as the labeled instances here.

Experimental environment is given below. All the computations are performed on a node of compute cluster with 16 CPUs (Intel Core Due 3.0 GHz) running RedHat Linux Enterprise 5 with 48 GB main memory which is similar with the one used in OPAUC. The coding environment is MATLAB 2016.

For each dataset, we choose 80% for training and the rest is used for test. Since some datasets maybe have many instances and limited to our memory, so we select 10000 training instances at random (without replacement) over the whole training data for batch algorithms if training size exceeds 10000. For all online methods, we go through the entire training data only once. Then in terms of the parameter settings for the compared methods, we can refer to each related reference. For our MOPAUC, the parameter setting is similar with the one in OPAUC for fair comparison. Namely, η_{ut} and η_{vt} are selected from the set $2^{[-12:10]}$, the regularization parameters λ_1 and λ_2 are selected from the set $2^{[-10:2]}$, weights u and v are initialized from the set $10^{[-3:3]}$. In order to get the optimal parameters, for each compared method, we carry out 10-fold cross-validation and repeat for ten times so as to get the average optimal results. In other words, the results in the following tables are from ten runs.

Table 3 shows the average testing AUC results for all compared methods on the benchmark datasets after we carry out the experiments for ten runs. From this table, it is found that in terms of testing AUC, our proposed MOPAUC is better than other compared online and off-line AUC optimization methods in average. Moreover, for the used five image datasets, MOPAUC performs best which validates that MOPAUC is feasible for the matrix-instance-based AUC optimization problems. Furthermore, the win/tie/loss counts show that MOPAUC is clearly superior to these online methods, as it wins for most times and never loses. For the other off-line AUC optimization methods, they performs better than MOPAUC sometimes. The reason is that these off-line methods can store the whole dataset so that they have potential for better performances.

What's more, it is found that the proposed MOPAUC outperforms OPAUC on vector instances. The reason for such a result can refer to the relationship between some vector-instance-based learning machines (for example, MHKS, i.e., modification of Ho-Kashyap algorithm with squared approximation of the misclassification errors [21]) and their corresponding matrixized versions (for example, MatMHKS, i.e., matrix-instance-based MHKS [15]). As we know, MHKS is a learning machine to process vector instances directly and MatMHKS which is developed on the base of MHKS is a one to process matrix instances directly. In MHKS, ωx^T is used to label a vector instance x while in MatMHKS, $uA^T v^T$ is used for labeling. Here, ω , u , and v are classifier weights and A is the matrix version of x . As [15] and [22] said, with $uA^T v^T$ used, MatMHKS is treated as MHKS imposed with Kronecker product decomposability constraint and MatMHKS has more constraints than MHKS since MatMHKS should optimize more weights. More constraints bring more prior information such as structural or local contextual information and the information bring a better performance. For that, MatMHKS outperforms MHKS even though they process vector instances. According to the same reason, in terms of the forms of models, the relationship between OPAUC and MOPAUC is same as the one between MHKS and MatMHKS, thus MOPAUC has more constraints than OPAUC, and then MOPAUC has more useful information to design a feasible classifier. That's why our developed MOPAUC outperforms OPAUC on vector instances.

Table 1. Algorithm: MOPAUC

Input: Regularization parameters $\lambda_1 > 0$, $\lambda_2 > 0$, step sizes $\{\eta_{ut}\}_{t=1}^T$, $\{\eta_{vt}\}_{t=1}^T$

Initialize: Set $T_0^+ = T_0^- = 0$, $c_0^+ = c_0^- = [0]_{m \times n}$, $S_{0u}^+ = [0]_{m \times m}$, $S_{0v}^- = [0]_{n \times n}$

$S_{0u}^+ = [0]_{n \times n}$, $S_{0u}^- = [0]_{m \times m}$, $S_{0v}^- = [0]_{n \times n}$

1. for $t=1, 2, \dots, T$ do
2. Arrive a training instance (A_t, y_t)
3. if $y_t = +1$ then
4. $T_t^+ = T_{t-1}^+ + 1$ and $T_t^- = T_{t-1}^-$
5. $c_t^+ = c_{t-1}^+ + \frac{1}{T_t^+}(A_t - c_{t-1}^+)$ and $c_t^- = c_{t-1}^-$
6. Update $S_{t_u}^+$, $S_{t_v}^+$, $S_{t_u}^-$, $S_{t_v}^-$ with Eq. (20)
7. Calculate the gradient of $\mathcal{L}_t^+(u, v)$
8. else
9. $T_t^- = T_{t-1}^- + 1$ and $T_t^+ = T_{t-1}^+$
10. $c_t^- = c_{t-1}^- + \frac{1}{T_t^-}(A_t - c_{t-1}^-)$ and $c_t^+ = c_{t-1}^+$
11. Update $S_{t_u}^+$, $S_{t_v}^+$, $S_{t_u}^-$, $S_{t_v}^-$ with Eq. (19)
12. Calculate the gradient of $\mathcal{L}_t^-(u, v)$
13. end if
14. Update u_t and v_t with Eqs. (17) and (18)
15. end for

Output: weights u_T and v_T

Table 2. Benchmark datasets

datasets	No. instances	No. features	datasets	No. instances	No. features	datasets	No. instances	No. features
AuC	690	14	PID	768	8	BA	1372	4
BCW	699	9	Satellite Image	6435	36	TSE	5820	32
GeD	1000	24	Shuttle	58000	9	UKM	403	5
Glass	214	9	Sonar	208	60	QSAR	1055	41
Heart	270	13	Thyroid	7200	21	Coil-20	1440	32 × 32
Iris	150	4	Vowel	990	10	Letter-Image	500	24 × 18
Letter	20000	16	Waveform	5000	21	ORL	400	32 × 20
Liver	345	6	Waveform-noise	5000	40	CIFAR-10	60000	32 × 32
Pendigits	7494	16	Wine	178	13	MNIST	60000	28 × 28

Table 3. Testing AUC (mean \pm std.) of MOPAUC with compared methods on benchmark datasets. ●/○ indicates that MOPAUC is significantly better/worse than the corresponding method (pairwise t-tests at 95% significance level). The best average AUC for each dataset is shown in bold.

datasets	MOPAUC	OPAUC	KOAUC	KOIL	SSAUC _{GM}	ELMAUC	SVM _{PAUC}
AuC	78.07 \pm 1.00	77.50 \pm 1.77	77.77 \pm 0.54 ●	76.23 \pm 1.41	76.44 \pm 1.03	74.37 \pm 1.12	73.40 \pm 1.33
BCW	89.85 \pm 1.46	89.44 \pm 1.99 ●	88.59 \pm 2.76	88.56 \pm 1.27 ●	88.58 \pm 0.48 ●	88.09 \pm 0.30 ○	87.22 \pm 1.69 ●
GeD	71.77 \pm 0.07	79.78 \pm 1.46 ●	71.74 \pm 1.36 ●	70.17 \pm 2.10 ●	70.57 \pm 0.71 ●	68.27 \pm 2.13 ○	67.58 \pm 0.26 ●
Glass	84.58 \pm 1.95	83.03 \pm 0.44 ●	83.21 \pm 1.25	83.15 \pm 1.89	83.46 \pm 0.77	82.48 \pm 2.48	81.74 \pm 2.00
Heart	79.34 \pm 0.59	79.79 \pm 0.73	80.20 \pm 0.18 ●	77.17 \pm 1.17 ●	77.98 \pm 0.95 ○	73.79 \pm 2.83	72.96 \pm 2.14 ○
Iris	91.37 \pm 0.99	89.70 \pm 0.20 ●	89.01 \pm 1.75	88.81 \pm 2.16	89.22 \pm 2.71	88.91 \pm 0.93 ○	88.09 \pm 0.55
Letter	86.02 \pm 2.76	81.14 \pm 0.64	85.09 \pm 0.65 ●	79.65 \pm 1.65 ●	81.06 \pm 0.21 ●	74.41 \pm 2.66 ●	72.84 \pm 1.56 ●
Liver	64.43 \pm 2.99	63.23 \pm 0.95 ●	63.61 \pm 1.73	63.57 \pm 0.84	63.91 \pm 2.11	63.28 \pm 0.31	62.63 \pm 2.30 ●
Pendigits	91.58 \pm 2.70	90.75 \pm 1.71 ●	90.54 \pm 1.21 ●	89.40 \pm 0.51	89.65 \pm 0.70	88.12 \pm 0.37 ●	87.30 \pm 2.67 ●
PID	66.30 \pm 0.38	64.84 \pm 0.77 ●	64.49 \pm 0.30	64.40 \pm 2.26 ●	64.33 \pm 0.78 ●	63.41 \pm 2.42	63.03 \pm 2.22
Satellite Image	77.96 \pm 1.46	75.99 \pm 0.62	75.88 \pm 0.33 ●	75.30 \pm 0.92	75.33 \pm 0.58	74.41 \pm 1.72 ●	73.79 \pm 1.86 ●
Shuttle	86.38 \pm 1.82	84.54 \pm 1.79 ●	83.25 \pm 1.63	83.32 \pm 2.11	82.36 \pm 1.80	81.28 \pm 2.91	80.48 \pm 1.54
Sonar	70.94 \pm 2.25	69.41 \pm 0.94	69.19 \pm 0.36 ●	67.71 \pm 1.53	68.22 \pm 2.59 ○	65.90 \pm 2.80	65.18 \pm 0.21 ○
Thyroid	84.47 \pm 0.71	83.16 \pm 0.09 ●	83.79 \pm 1.09	84.08 \pm 0.84	84.44 \pm 0.66 ●	84.22 \pm 2.82	83.84 \pm 0.99 ●
Vowel	54.73 \pm 2.64	53.75 \pm 0.78	53.37 \pm 1.46	50.54 \pm 2.21	51.45 \pm 2.21	48.35 \pm 0.05	47.31 \pm 1.10
Waveform	72.18 \pm 0.55	71.60 \pm 0.33 ●	71.05 \pm 1.10 ●	70.18 \pm 1.41 ●	70.91 \pm 1.82 ●	70.09 \pm 0.40	69.26 \pm 0.73
Waveform-noise	77.68 \pm 0.48	76.95 \pm 0.56 ●	77.42 \pm 1.63	76.70 \pm 1.39	77.24 \pm 0.02	75.13 \pm 0.80	74.65 \pm 2.41 ●
Wine	82.41 \pm 0.27	80.83 \pm 1.62 ●	80.61 \pm 1.12 ●	81.00 \pm 1.43 ●	80.67 \pm 2.43 ○	80.26 \pm 2.71	79.93 \pm 1.88 ●
BA	85.38 \pm 1.82	85.86 \pm 0.99 ●	83.33 \pm 1.95 ●	88.42 \pm 1.67	88.45 \pm 0.57	88.80 \pm 1.29 ○	88.02 \pm 2.68 ●
TSE	85.29 \pm 1.58	83.85 \pm 2.84 ●	84.46 \pm 0.30	85.11 \pm 1.57 ●	85.28 \pm 0.92 ●	85.04 \pm 2.04	84.69 \pm 0.34
UKM	81.71 \pm 1.25	80.69 \pm 2.62	79.21 \pm 0.43	78.57 \pm 0.25	78.91 \pm 2.81 ○	77.68 \pm 0.79 ○	76.68 \pm 2.29 ○
QSAR	92.51 \pm 0.39	90.33 \pm 0.23	89.69 \pm 0.41 ●	84.31 \pm 2.94	86.53 \pm 2.76	79.50 \pm 1.51	78.05 \pm 2.54
Coil-20	79.64 \pm 1.47	77.27 \pm 1.02 ●	76.84 \pm 2.10 ●	76.44 \pm 2.99 ●	76.77 \pm 0.20 ●	76.10 \pm 2.94	75.89 \pm 1.98 ●
Letter-Image	79.64 \pm 2.11	78.31 \pm 2.35 ●	78.34 \pm 2.48 ●	77.11 \pm 2.77 ●	77.37 \pm 1.51 ●	75.90 \pm 1.19	75.35 \pm 0.02 ●
ORL	83.11 \pm 0.54	82.84 \pm 2.04 ●	82.81 \pm 2.52 ●	79.56 \pm 2.26 ●	81.29 \pm 2.94	77.25 \pm 0.89	76.00 \pm 2.42
CIFAR-10	94.03 \pm 0.31	93.88 \pm 0.14 ●	93.01 \pm 2.58 ●	89.45 \pm 1.63 ●	78.32 \pm 2.65 ●	74.09 \pm 0.67 ●	72.47 \pm 2.07 ●
MNIST	93.56 \pm 0.36	92.42 \pm 0.21 ●	92.34 \pm 0.57 ●	84.32 \pm 1.07 ●	78.41 \pm 1.83 ●	81.55 \pm 0.70 ●	73.30 \pm 0.15 ●
win/tie/loss		18 / 9 / 0	16 / 11 / 0	15 / 12 / 0	13 / 10 / 4	11 / 11 / 5	14 / 9 / 4

Moreover, we also compare the average running time of MOPAUC and the other three online AUC optimization methods after we carry out the experiments for ten runs. Table 4 shows the comparison of the running time (in seconds) of MOPAUC and the compared online methods on the used datasets. From this table, it is found that for the datasets except the five image datasets, KOAUC and KOIL can cost least running time in average. The reason is that KOAUC and KOIL optimize on single instance loss, whereas MOPAUC and OPAUC optimize on pairwise loss. Moreover, compared with OPAUC, MOPAUC costs less running time. Especially, for the image datasets, our proposed MOPAUC costs least running time which validate the effectiveness of our method. Indeed, as OPAUC said, its required storage is $O(d^2)$ where $d = m \times n$ while for our MOPAUC, the storage requirement is reduced to $O(m^2 + n^2)$. Furthermore, as [23] said, compared with vector leaning machine, matrix learning machine can reduce the computational complexity and improve the classification performance. The reduction of computational complexity always brings less running time.

Table 4. Comparison of the running time (in seconds) on datasets for the online AUC optimization methods.

datasets	MOPAUC	OPAUC	KOAUC	KOIL	datasets	MOPAUC	OPAUC	KOAUC	KOIL
AuC	0.10	0.37	0.02	0.01	Thyroid	0.38	2.86	0.74	0.09
BCW	0.03	0.15	0.03	0.01	Vowel	0.19	0.65	0.07	0.16
GeD	0.10	1.08	0.11	0.05	Waveform	0.38	2.86	0.15	0.22
Glass	0.03	0.15	0.01	0.02	Waveform-noise	0.58	10.37	1.09	0.59
Heart	0.32	0.32	0.02	0.09	Wine	1.10	1.10	0.17	0.14
Iris	0.01	0.03	0.01	0.01	BA	0.05	0.10	0.02	0.02
Letter	0.21	1.66	0.07	0.07	TSE	0.52	6.64	1.19	1.75
Liver	0.08	0.23	0.05	0.02	UKM	0.13	0.16	0.02	0.04
Pendigits	0.21	1.66	0.55	0.02	QSAR	10.81	10.90	2.30	0.96
PID	0.13	0.41	0.03	0.06	Coil-20	13.28	6797.67	735.72	1519.31
Satellite Image	0.63	8.40	0.83	1.59	Letter-Image	5.83	1209.84	55.07	81.79
Shuttle	0.12	0.53	0.11	0.10	ORL	9.23	2655.34	58.46	344.08
Sonar	0.88	23.34	4.96	3.15	CIFAR-10	13.28	6797.67	283.49	463.37
MNIST	10.16	3984.67	188.00	856.00					

3.2 Experiments About Parameter Influence

In our proposed MOPAUC, it consists many adjustable parameters including regularization parameters λ_1 , λ_2 , step sizes η_{ut} , η_{vt} , weights u and v . So here, we discuss the influence of them. Since $\eta_{ut} \in 2^{[-12:10]}$, $\eta_{vt} \in 2^{[-12:10]}$, $\lambda_1 \in 2^{[-10:2]}$, $\lambda_2 \in 2^{[-10:2]}$, $u \in 10^{[-3:3]}$, and $v \in 10^{[-3:3]}$, so we use the following three figures to show the influence. For the convenience of elaboration, we only select four datasets, they are GeD, Letter, CIFAR-10, MNIST. Figure 1 shows the influence of the regularization parameters; Fig. 2 shows the influence of the step sizes; Fig. 3 shows the one of the weights. Each sub-figure in each figure, 2^x and 2^y just represent the power operation. Namely, the parameter is 2^{-10} , 2^6 and so on. For 10^x and 10^y , the meaning is same. According to these three figures, it is found that the regularization parameters and weights have less influence on the average AUC. While the step sizes should not be set to values bigger than 1, whereas there is a relatively big range between $[2^{-12}, 2^{-4}]$ where MOPAUC achieves good results. This conclusion is similar with one given in OPAUC [13].

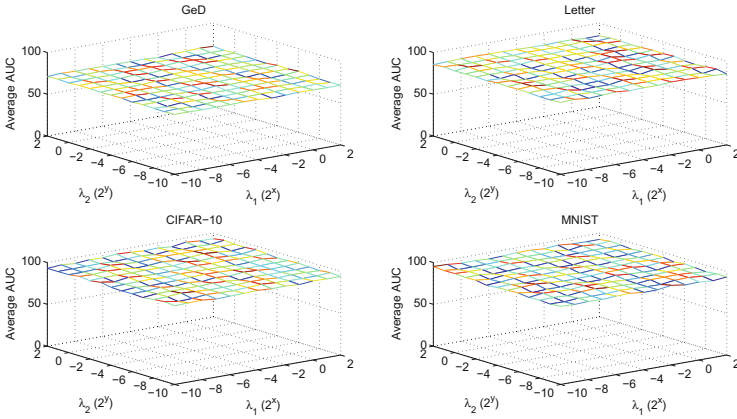


Fig. 1. Influence of regularization parameters.

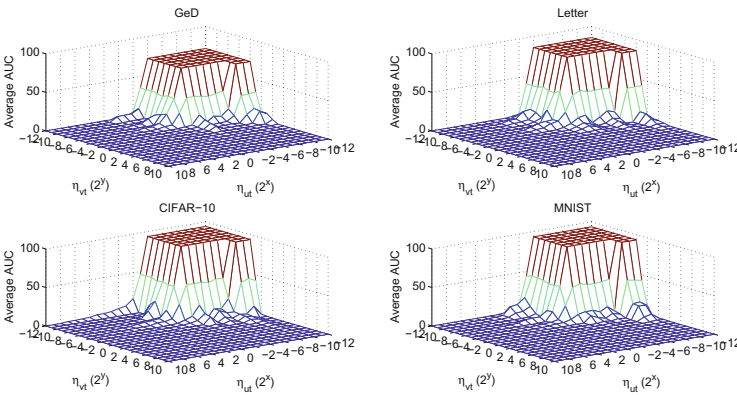


Fig. 2. Influence of step sizes.

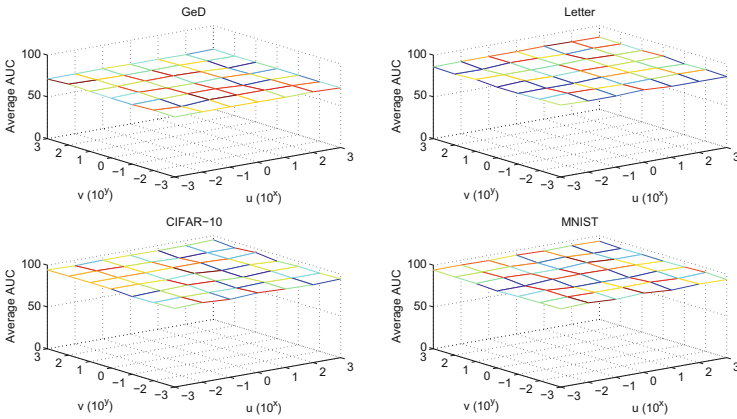


Fig. 3. Influence of weights.

4 Conclusion

AUC is an important performance measure and it is always measured by the losses defined over pairs of instances from different classes. Conducting traditional off-line AUC optimization methods should store the entire dataset in memory which is infeasible for big data or streaming data applications. Online AUC optimization methods need not store the entire dataset, but the present mostly online AUC optimization methods still need to store \sqrt{T} instances yet where T is the number of the entire training dataset. One-pass AUC optimization (OPAUC) is a new online one and it is independent from the number of training instances. While OPAUC is infeasible for matrix datasets including images. So this paper extends the model of OPAUC and develops a matrix-instance-based one-pass AUC optimization model, i.e., MOPAUC, so as to process the matrix datasets. Related experiments on some datasets including the vector ones and matrix ones validate that (1) MOPAUC has a best average testing AUC compared with the online and off-line AUC optimization methods; (2) MOPAUC is superior to some online methods from the statistical view; (3) MOPAUC can cost less running time compared with other online methods for processing image datasets; (4) regularization parameters and weights have less influence on the average AUC for MOPAUC while step sizes have strong influence. If the values of step sizes range from $[2^{-12}, 2^{-4}]$, MOPAUC can achieve good results. If the values are bigger than 1, the average AUC will decrease to be 0. In general, our proposed MOPAUC can process matrix-instance-based AUC optimization problems without storing the dataset and only scanning the training data once.

Acknowledgment. This work is supported by (1) Natural Science Foundation of Shanghai under grant number 16ZR1414500 (2) National Natural Science Foundation of China under grant number 61602296 and the authors would like to thank their supports.

References

1. Ding, Y., Liu, C.H., Zhao, P.L., Hoi, S.C.H.: Large scale kernel methods for online AUC maximization. In: 2017 IEEE International Conference on Data Mining, pp. 91–100 (2017)
2. Liu, R.H., Hall, L.O., Bowyer, K.W., Goldgof, D.B., Gatenby, R., Ahmed, K.B.: Synthetic minority image over-sampling technique: How to improve AUC for glioblastoma patient survival prediction. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1357–1362 (2017)
3. Jiang, H., Yi, J., Chen, S., Zhu, X.: A multi-objective algorithm for task scheduling and resource allocation in cloud-based disassembly. *J. Manuf. Syst.* **41**, 239–255 (2016)
4. Fan, Q.Q., Wang, W.L., Yan, X.F.: Multi-objective differential evolution with performance-metric-based self-adaptive mutation operator for chemical and biochemical dynamic optimization problems. *Appl. Soft Comput.* **59**, 33–44 (2017)
5. Fan, P., Zhou, R.G., Jing, N., Li, H.S.: Geometric transformations of multidimensional color images based on NASS. *Inf. Sci.* **340**, 191–208 (2016)

6. Gao, W., Jin, R., Zhu, S.H., Zhou, Z.H.: One-pass AUC optimization. In: Proceedings of the 30th International Conference on Machine Learning, pp. 906–914 (2013)
7. Hu, J.J., Yang, H.Q., Lyu, M.R., King, I., So, A.M.C.: Online nonlinear AUC maximization for imbalanced data sets. *IEEE Trans. Neural Netw. Learn. Syst.* **99**, 1–14 (2017)
8. Khajavi, N.T., Kuh, A.: The covariance selection quality for graphs with junction trees through AUC bounds. In: 54th Annual Allerton Conference on Communication Control and Computing (Allerton), pp. 1252–1258 (2016)
9. Kim, Y.S., Toh, K.A., Teoh, A.B.J., Eng, H.L., Yau, W.Y.: An online AUC formulation for binary classification. *Pattern Recogn.* **45**(6), 2266–2279 (2012)
10. Wang, S.J., Li, D., Petrick, N., Sahiner, B., Linguraru, M.G., Summers, R.M.: Optimizing area under the ROC curve using semi-supervised learning. *Pattern Recogn.* **48**(1), 276–287 (2015)
11. Li, Z.L., Zhai, S.D., Xia, T., Wang S.J.: A boosting method for direct AUC optimization. In: 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), pp. 797–801 (2015)
12. Fujino, A., Ueda N.: A semi-supervised AUC optimization method with generative models. In: 2016 IEEE 16th International Conference on Data Mining, pp. 883–888 (2016)
13. Gao, W., Wang, L., Jin, R., Zhu, S.H., Zhou, Z.H.: One-pass AUC optimization. *Artif. Intell.* **236**, 1–29 (2016)
14. Blake, C.L., Newman, D.J., Hettich, S., Merz C.J.: UCI repository of machine learning databases (2012)
15. Chen, S.C., Wang, Z., Tian, Y.J.: Matrix-pattern-oriented Ho-Kashyap classifier with regularization learning. *Pattern Recogn.* **40**(5), 1533–1543 (2007)
16. Zhao, P., Hoi, S.C.H., Jin, R., Yang T.: Online AUC maximization. In: Proceedings of the 25th International Conference on Machine Learning, pp. 233–240 (2011)
17. Kotlowski, W., Dembczynski, K., Hüllermeier, E.: Bipartite ranking through minimization of univariate loss. In: Proceedings of the 28th International Conference on Machine Learning, pp. 1113–1120 (2011)
18. Fujino, A., Ueda, N.: A semi-supervised AUC optimization method with generative models. In: IEEE International Conference on Data Mining, pp. 883–888 (2017)
19. Yang, Z.Y., Zhang, T.H., Lu, J.C., Zhang, D.Z., Kalui, D.: Optimizing area under the ROC curve via extreme learning machines. *Knowl.-Based Syst.* **130**, 74–89 (2017)
20. Narasimhan, H., Agarwal, S.: Support vector algorithms for optimizing the partial area under the ROC curve. *Neural Comput.* **29**(7), 1919–1963 (2017)
21. Leski, J.: Ho-Kashyap classifier with generalization control. *Pattern Recogn. Lett.* **24**(14), 2281–2290 (2003)
22. Zhu, C.M., Wang, Z., Gao, D.Q.: New design goal of a classifier: global and local structural risk minimization. *Knowl.-Based Syst.* **100**, 25–49 (2016)
23. Zhu, C.M.: Double-fold localized multiple matrix learning machine with Univer-sum. *Pattern Anal. Appl.* **20**(4), 1091–1118 (2017)



Piecewise Harmonic Image Restoration with High Order Variational Model

Bibo Lu^{1(✉)}, Zhenzhen Huang¹, and Rui Huang²

¹ Henan Polytechnic University, Jiaozuo 454003, Henan, China
lubibojz@gmail.com

² South China Normal University, Guangzhou 510631, Guangdong, China

Abstract. Image denoising is a fundamental problem in image processing and computer vision. A main challenge is to remove noise while preserving features and developing piecewise smoothing image. Piecewise constant and linear image recovery has been focused in the past decades. In this paper, we propose a model recover a class more smoothing image with complex geometrical structure. We first give definition of piecewise harmonic image, which covers a wide range piecewise smoothing image. Then a multiplicative framework for high order variational construction is introduced. Within this framework, we present a geometrical weighted Laplace (GWL) high order model. The proposed model is discussed and compared to some typical related methods. Experimental results on test images show the performance of the proposed method.

Keywords: Image denoising · Piecewise smoothing image
High order · Harmonic function

1 Introduction

In a standard problem of gray scale image denoising problem, the noisy image u_0 corrupted by additive white Gaussian noise is modeled as

$$u_0(x, y) = u(x, y) + \sigma(x, y), \quad (1)$$

where u is the unknown noisy free image and σ is assumed as known noise level: $\int_{\Omega} (u - u_0)^2 dx dy = \sigma^2$. The goal of image restoration is to remove noise while preserving the important structure features from the observed noisy image u_0 [1]. An usual regularization approach to remove noise by minimizing the following functional:

$$E(u, \lambda) = E(u) + \frac{\lambda}{2} \int_{\Omega} (u - u_0)^2 dx dy, \quad (2)$$

where $E(u)$ is the regularization term to measure the variation of the noise intensity and $\lambda \geq 0$ is the Lagrange multiplier. The first regularization term on the

Supported by NSFC (U1404103) and Guangdong Engineering Research Center for Data Science.

right-hand side of Eq. (2) is to measure the oscillations using weighted Laplace operator. The second fitting term is to measure the identification between u and u_0 . In seminar total variational (TV) method [2], the regularization functional is defined as

$$E_{\text{TV}}(u) = \int_{\Omega} |\nabla u| dx dy, \quad (3)$$

which produces a piecewise constant image while removing noise. However, TV suffers from staircase effect in smoothing transition region [3]. A more smoothing image is also expected in varying image processing fields, including computer photography [4], medical image processing [5], image registration [6], Retinex problem [7]. Some operations have been used to construct high order models, such as Laplace operation based YK model [3] and LLT model [5], the Frobenius norm of the Hessian based affine TV model [8], curvature based elastic model [9] mean curvature based model [10] and Gaussian curvature based model [11]. A variable exponent high order variational model was proposed in [12], where the Gaussian convolution was used for detecting edges.

Low order model and high order operators are combined to construct new methods: one part to produce flat image and the other part to generate smoothing transition. In [15], Papafitsoros and Schönlieb considered a general additive high order functional and proved its existence and uniqueness. A popular high order model, total generalized variation (TGV), involves high order derivatives and automatically balances the first to k th derivatives [13]. The second order TGV is defined as following:

$$E(u)_{\text{TGV}} = \text{TGV}_{\alpha}^2 = \alpha_1 \int_{\Omega} |\nabla u - v| dx dy + \alpha_2 \int_{\Omega} |\varepsilon(v)| dx dy, \quad (4)$$

where the minimum is taken over the vector fields v and $\varepsilon(v) = \frac{1}{2}(\nabla v + \nabla v^{\text{T}})$ denotes the symmetrized derivative. TGV reduces the staircase effect and leads to piecewise polynomial intensities [14]. The connections between some typical additive high order models are detailed in [15]. Typical non-variational methods includes bilateral filter [16], nonlocal means filter [17,18], guided filter [19] and BM3D [20].

In this paper, we will introduce piecewise harmonic image, which is more smoother beyond the classical piecewise constant image and piecewise linear image. It allow a weak edge between different regions and it is difficult to recovery it. We will present a a new model to address this problem. The rest of this paper is organized as follows. Since our aim is to recover a more smoothing image, Sect. 2 introduce the definition of harmonic image and a new multiplicative framework for model construction. A new high order model is presented and its features are discussed in Sect. 3, Experimental results are shown in Sect. 4 and a brief conclusion is given in Sect. 5.

2 Framework for Piecewise Smoothing Image Recovery

2.1 Piecewise Smoothing Image: From Constant to Harmonic

Let $\Omega_i, i = 1, 2, \dots, n$, be a partition of Ω . A common piecewise image is defined as

$$u(x, y) = \sum_{i=0}^m u_i(x, y), \tag{5}$$

where

$$u_i(x, y) = \begin{cases} \text{smoothing image} & (x, y) \in \Omega_i, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

We require that the smoothing image in (6) is continuous and differentiable in every partition Ω_i . An usual way is to use homogeneous polynomial to represent the smoothing function. Therefore, image I is named as a piecewise constant image when $u(i) = c_i$ and a piecewise linear or affine image when $u(i) = a_i x + b_i y + c_i$. TV can recover piecewise constant image successfully. Several high order models have been proposed to recover piecewise linear image.

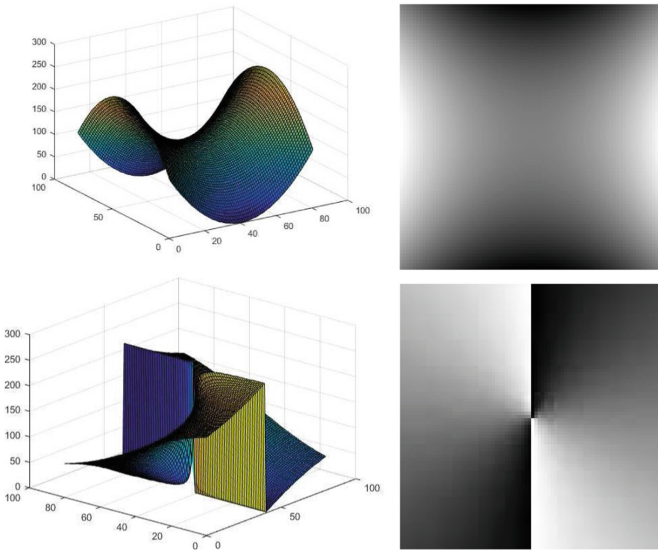


Fig. 1. Two harmonic functions and their corresponding images. The left-up is the shape of the harmonic function $f(x, y) = \frac{x^2}{a^2} - \frac{y^2}{b^2}$ and right-up is its corresponding image. The right-down is the shape of the harmonic function $f(x, y) = \frac{y}{x}$ and right-up is its corresponding image.

In this paper, for the first time, we consider the recovery of a class more smoothing image: piecewise harmonic image. In mathematic, a function f is said to be harmonic if it satisfies the following Laplace equation:

$$\Delta f(x, y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0. \tag{7}$$

Except the traditional constant function and the linear ones, many more smoothing function are harmonic. A quadratic one is a special case when $a = b$ for the hyperbolic paraboloid in geometry:

$$f(x, y) = \frac{x^2}{a^2} - \frac{y^2}{b^2}, \tag{8}$$

which describes the shape for a doubly ruled surface in 3D space. Another example is $f(x, y) = \arctan(\frac{y}{x})$, which has a non-vanishing derivatives to infinity. Figure 1 illustrates the the profiles of two harmonic functions and their corresponding images. These two functions and images has a complex and smoothing geometrical structure. Therefore, an image u is said to be a piecewise harmonic if $u(i)$ meets Eq. (7) in partition Ω_i . It permits more wild range smoothing functions beyond polynomial, though it is an extension to the traditional piecewise constant image and linear image. The sharp edges between the piecewise constant are easy to preserved and it is difficult to preserve the edges between the different harmonic regions, as its gradient may be small.

2.2 Multiplicative High Order Variational Framework

Based on the decisions above, we may infer that it is a challenge to recover piecewise harmonic image as it permits more smoothing structures beyond constant region and affine region. Before constructing a feasible variation model to this problem, we should consider two issues. The first is how to judge where is the boundaries of different smoothing transition regions, which is helpful for a reasonable piecewise. The second is how to choose a proper way to describe the smoothing function, which is responsible for smoothing control. The answers to the two problems need to be integrated into the variational model. To improve the smoothing degree of the restored image, one need to incorporate high order operator to describe the smoothing requirement. Therefore, we proposed the following general high order framework for piecewise smoothing image recovery:

$$E(u) = \int_{\Omega} f_p(u, u_i, u_{ij}) f_s(u, u_i, u_{ij}) dx dy, \tag{9}$$

where f_p provides the clues for judging the boundaries between piecewise regions and f_s conveys the smoothing control respectively. Contrary to the traditional additive high order variational model framework, the multiplicative model is easy to extend to other imaging tasks.

3 Proposed Weighted Laplacian Model

By consider a gray scale image $u(x, y)$ as a surface $S = (x, y, u(x, y))$, we propose the the following geometrical weighted Laplace (GWL) energy functional:

$$E_{GWL}(u) = \int_{\Omega} \frac{|\Delta u|}{\sqrt{1 + |\nabla u|^2}} dx dy. \tag{10}$$

The kernel in the the energy (10) is a product of two functions and it can be seen as a special case for (9) when choosing $f_p = \frac{1}{\sqrt{1+|\nabla u|^2}}$ and $f_s = |u_{xx} + u_{yy}| = |\Delta u|$.

The key of recovery of the piecewise harmonic image is the interaction between two functions.

1. Piecewise. The piecewise effect in a certain partition is guaranteed by edge boundary detector $g = \sqrt{1 + |\nabla u|^2}$, which has a remarkable geometrical interpretation:

$$r = \frac{1}{g} = \frac{1}{\sqrt{1 + u_x^2 + u_y^2}} = \frac{dxdy}{gdx dy} = \frac{A^{\text{domain}}}{A^{\text{surface}}}, \tag{11}$$

where A^{domain} is the area of the infinitesimal surface in the image domain (x, y) , and A^{surface} is its corresponding area on the image surface $(x, y, u(x, y))$. Therefore, r conveys the height variation on the surface as well the intensity variation on the image data [21]. r is equal 1 for flat surface and its Laplacian is zero too, such structure will be preserved. r is equal 0 near edges, which is helpful to preserve edges.

2. Harmonic. The smoothing harmonic constrain is mainly performed by Laplacian operator Δu . As $g = \sqrt{1 + |\nabla u|^2} > 1$, zero Laplacian means the kernel function will be zero too and functional reaches the minimizer in this region. Therefore, smoothing structures will be kept if they can be represented as any harmonic function.
3. Edge preserving. For an ideal typical sharp edge, its Laplace has a famous zero crossing property: near the midpoint of the edge, its second order derivative would cross zero. The kernel function will be 0 as $\Delta u = 0$ and $r = 0$ for a true sharp edge, which will be recognized and well preserved.

Therefore, the proposed model permits discontinuous while preserving piecewise smoothing regions.

Adding an artificial time to the Euler-Lagrange equation derived to (10), we can obtain the following an anisotropic high order nonlinear diffusion equation:

$$u_t = -\Delta \left(\frac{\Delta u}{g|\Delta u|} \right) + \text{div} \left(\frac{|\Delta u|}{g^3} \nabla u \right) - \lambda(u - u_0). \tag{12}$$

The initial condition is $u(x, 0) = u_0$ and its boundary condition is

$$(u_x, u_y) \cdot \boldsymbol{\mu} = 0, \quad (\gamma_1, \gamma_2) \cdot \boldsymbol{\mu} = 0, \tag{13}$$

where $\boldsymbol{\mu}$ is the unit outward normal direction to $\partial\Omega$ and γ_1 and γ_2 are defined as

$$\gamma_1 = \left(\frac{\Delta u}{g|\Delta u|} \right)_x + \frac{|\Delta u|u_x}{g^3}, \tag{14}$$

$$\gamma_2 = \left(\frac{\Delta u}{g|\Delta u|} \right)_y + \frac{|\Delta u|u_y}{g^3}. \tag{15}$$

The diffusion of Eq. (12) is decided by the interaction of the first order edge detector g and the second order information $\frac{\Delta u}{|\Delta u|}$. Noting $\frac{\Delta u}{|\Delta u|} = \text{sign} \Delta u$, only three values, $-1, 0, 1$ are permitted. When it equals 0, the diffusion stop automatically. It means that the local structure described by the harmonic function maybe preserved. When it equals 1 or -1 , the diffusion now depends on the magnitude the boundary detector. The diffusion speed will slow down as the sign of the Laplace operator is scaled by the inverse of a large gradient magnitude. A fast diffusion will be performed in flat region as the image gradient is small and the boundary detector $g \simeq 1$.

As the evolution equation is nonlinear highly, we now consider to solve it by an explicit finite difference method. For time discretion, forward difference is used and the space grid size is set as $h = 1$. Table 1 lists the scheme for time and spatial operators in high order nonlinear Eq. (12).

Table 1. The discrete scheme for operators in high order nonlinear Eq. (12).

Continuous variable	Discrete variable	Discrete scheme
t	Δt	Time space
u	$u_{i,j}^0$	Initial image
Δu	$\Delta(u_{i,j})$	$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}$
u_x in fourth term	$D_x(u_{i,j})$	$D_x(u_{i,j}) = \frac{u_{i+1,j} - u_{i-1,j}}{2}$
u_y in fourth term	$D_y(u_{i,j})$	$D_y(u_{i,j}) = \frac{u_{i,j+1} - u_{i,j-1}}{2}$
u_x in third term	$D_x^\mp u_{i,j}$	$\mp(u_{i\mp 1,j} - u_{i,j})$
u_y in third term	$D_y^\mp u_{i,j}$	$\mp(u_{i,j\mp 1} - u_{i,j})$

4 Experimental Results

In this section, we conduct several experiments to demonstrate the performance of the high order GWL model. We make comparisons with three related methods. The first one is second order TV method, which is famous for its edge preserving ability. The second method is TGV method, which is implemented by a primal-dual splitting method in [22]. The code is also available: <http://www.gipsa-lab.fr/~laurent.condat/software.html>. The third one is state of art BM3D method. To do a quantitative comparison, peak signal-to-noise-ratio (PSNR) is used for quantitative comparison. For the proposed method, we set time space $\Delta t = 10^{-2}$ and $\lambda = 0.01$.

The first experimental results on a synthesized piecewise quadratic image are shown in Fig. 2. The test image is composed by two constant functions (one for left side and another for right side), a linea function (up middle and down middle) and a selected quadratic harmonic function for $u(x, y) = \frac{x^2}{16} - \frac{y^2}{16}$ in (8) (middle). The noise level is 10 and the denosing results for the noisy image

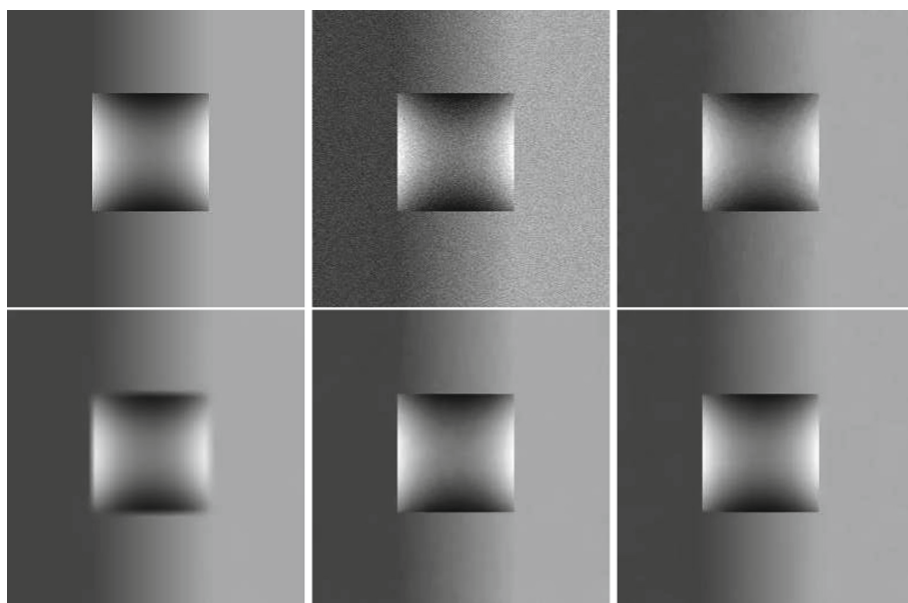


Fig. 2. Piecewise quadratic denoised images. The image is composed by two constant functions (one for left side and another for right side), a linea function (up middle and down middle) and a selected quadratic harmonic function for $u(x, y) = \frac{x^2}{16} - \frac{y^2}{16}$. From the left to right, the first row: clean image, noisy image (PSNR = 28.1376), TV result (PSNR = 43.3636). From the left to right, the second row, from the left to right, TGV result(PSNR = 32.3433), BM3D result (PSNR = 47.0606), GWL result (PSNR = 49.3684).

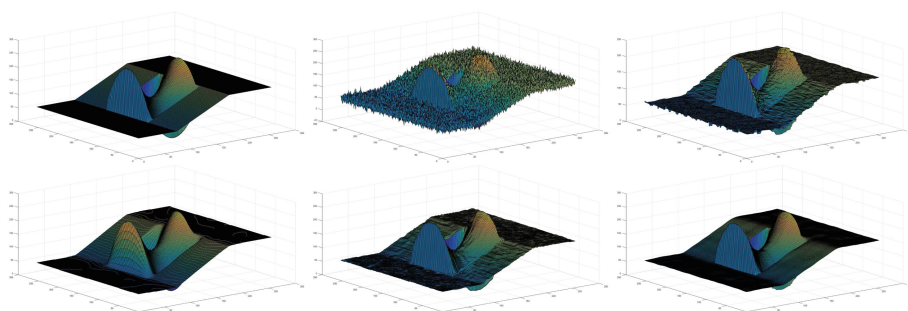


Fig. 3. The induced surfaces of piecewise quadratic denoised images. The order is the same as Fig. 2.

(PSNR = 28.1376) by four methods are shown in 2. The staircase effect is obvious in quadratic region for TV denoised image (PSNR = 43.3636). The TGV denoised image (PSNR = 32.3433) shows a good smoothing ability but blurs

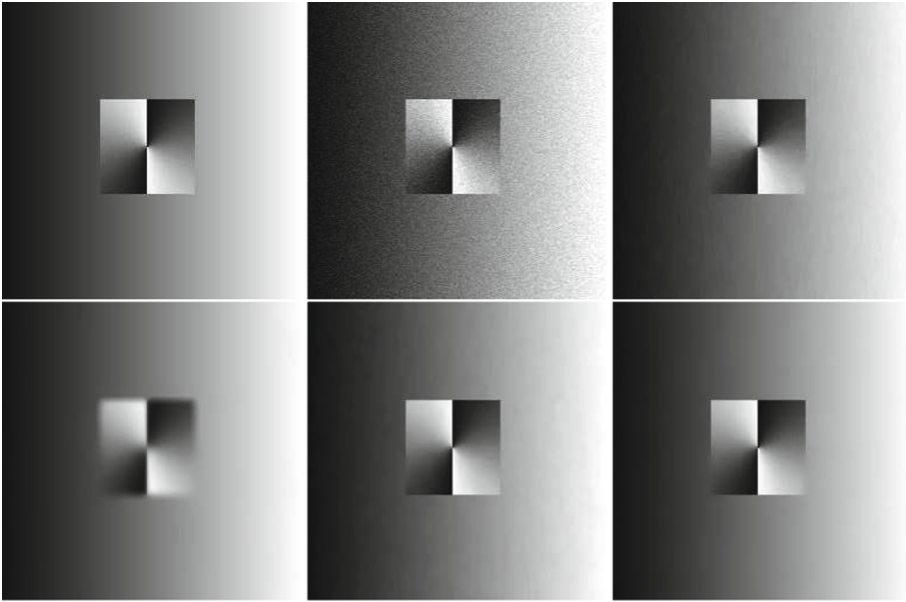


Fig. 4. Piecewise smoothing denoised images beyond quadratic. It is composed of a linear function and a smoothing function $u(x, y) = \arctan(\frac{y}{x})$, whose infinite derivatives are non-vanishing. From the left to right, the first row: clean image, noisy image (PSNR = 28.1221), TV result (PSNR = 43.2176). From the left to right, the second row, from the left to right, TGV result (PSNR = 31.8597), BM3D result (PSNR = 45.4445), GWL result (PSNR = 49.6065).

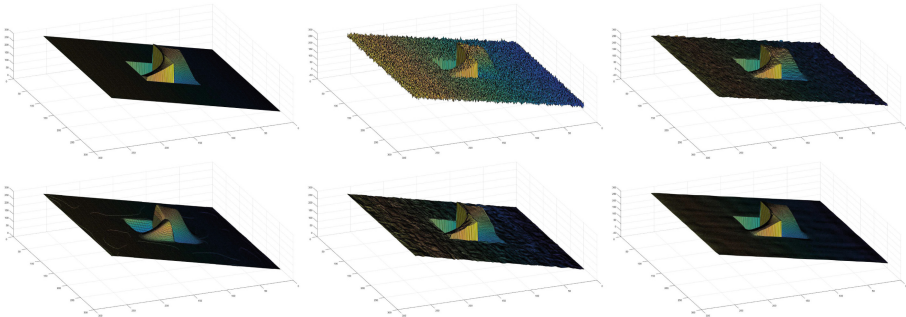


Fig. 5. The induced surfaces of piecewise smoothing denoised images in Fig. 4.

the edges seriously. The staircase effect in linear regions and quadratic regions is unpleasant in visual for BM3D denoised image (PSNR = 47.0606). The proposed GWL method provide an almost perfect denoised image visually and quantitatively (PSNR = 49.3684). The corresponding induced surfaces are displayed in

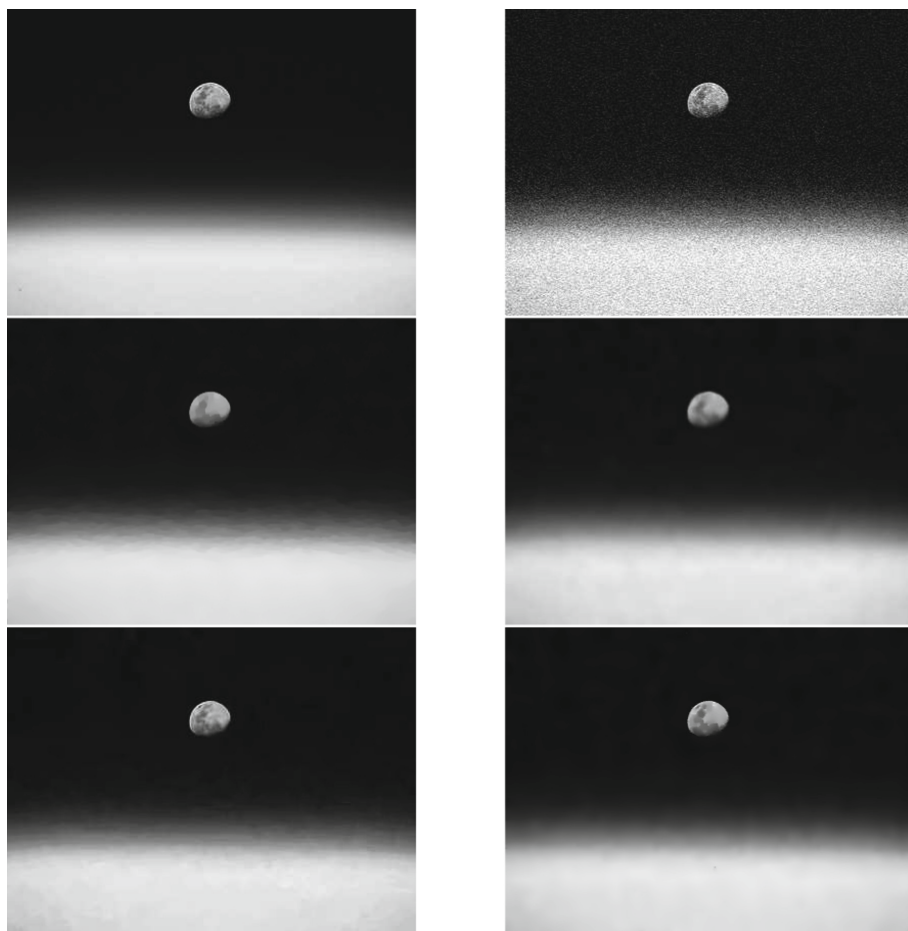


Fig. 6. Nasa denoised image. From the left to right, the first row: clean image, noisy image (PSNR = 22.1151), TV result (PSNR = 40.1939). From the left to right, the second row, from the left to right, TGV result (PSNR = 33.5203), BM3D result, GWL result (PSNR = 42.2764).

Fig. 3. It can be observed that TGV and GWL shows a better smoothing effect than TV and BM3D.

The second test synthesized image has a more complex structures: it is composed of a linear function and a smoothing function $u(x, y) = \arctan \frac{y}{x}$, whose infinite derivatives are non-vanishing. The noise level is 10 and the denoising results for the noisy image (PSNR = 28.1221) by four methods are shown in 4. TV result shows a serious staircase effect for the tangent function region (PSNR = 43.2176). The TGV denoised image (PSNR = 31.8597) blurs the edges heavily again. BM3D performs better than TV and TGV but staircase effect is visual for linear region (PSNR = 45.4445). The proposed GWL method yields a

best result among four methods visually and quantitatively (PSNR = 49.6065). The corresponding induced surfaces are displayed in Fig. 5.

The third test image is a picture of moon rise captured from the space station by NASA astronaut Randy Bresnik on August 3, 2017. The noise level is 20 and the denosing results for the noisy image (PSNR = 22.1151) by four methods are shown in Fig. 6. Four methods remove noise in white and black background. The differences between them lie in the moon surface and the smoothing transition regions in the middle of the image. BM3D provides the best detail preservation ability for moon surface (PSNR = 41.3395) while GWL produces a good transition effect between the white region and black region (PSNR = 42.2764). TV still suffers from the staircase (PSNR = 40.1939) and TGV blurs edges (PSNR = 33.5203).

5 Conclusions

We present a high order variational method to recover a class more smoothing piecewise image beyond quadratic, which we call piecewise harmonic image. Piecewise harmonic image covers the popular piecewise constant and piecewise linear images and beyond them, even including some certain function with infinite order non-vanishing derivatives. We construct the new model within a multiplicative variational framework and its kernel is based on a geometrical weighted Laplacian operation. The research in this paper shows that we can restore piecewise harmonic image perfectly. Its major limitation is the fact that the natural image do not always contain standard piecewise quadratic geometrical structures. Therefore, improvement on its adaptability to more image is part of our future work. Another important work is to devise an efficient speeding up algorithms for GWL model.

References

1. Chan, T., Shen, J.: *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*. SIAM Publisher, Philadelphia (2005)
2. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* **60**, 258–268 (1992)
3. You, Y.L., Kaveh, M.: Fourth-order partial differential equation for noise removal. *IEEE Trans. Image Process.* **9**(10), 1723–1730 (2000)
4. Tumblin, J., Turk, G.: LCIS: a boundary hierarchy for detail-preserving contrast reduction. In: *Proceedings of the SIGGRAPH 1999 Annual Conference on Computer Graphics*, Los Angeles, CA, USA, 83–90 (1999)
5. Lysaker, M., Lundervold, A., Tai, X.C.: Noise removal using fourth order partial differential equation with applications to medical magnetic resonance images in space and time. *IEEE Trans. Image Process.* **12**(12), 1579–1590 (2003)
6. Jewprasert, S., Chumchob, N., Chantrapornchai, C.: A fourth-order compact finite difference scheme for higher-order PDE-based image registration. *East Asian J. Appl. Math.* **5**(4), 361–386 (2015)

7. Liang, J., Zhang, X.: Retinex by higher order total variation L^1 decomposition. *J. Math. Imaging Vis.* **52**(3), 345–355 (2015)
8. Yuan, J., Schnörr, C., Steidl, G.: Total-variation based piecewise affine regularization. In: Tai, X.-C., Mörken, K., Lysaker, M., Lie, K.-A. (eds.) *SSVM 2009*. LNCS, vol. 5567, pp. 552–564. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02256-2_46
9. Tai, X.C., Hahn, J., Chung, G.J.: A fast algorithm for Euler’s elastica model using augmented Lagrangian method. *SIAM J. Imaging Sci.* **4**(1), 313–344 (2010)
10. Zhu, W., Chan, T.: Image denoising using mean curvature of image surface. *SIAM J. Imaging Sci.* **51**, 1–32 (2012)
11. Brito-Loeza, C., Chen, K., Uc-Cetina, V.: Image denoising using the Gaussian curvature of the image surface. *Numer. Methods Partial. Differ. Equ.* **32**(3), 1066–1089 (2016)
12. Bibo, L., Jianlong, W., Zhang, Q.: A variable exponent high-order variational model for noise removal. *J. Comput. Inf. Syst.* **11**(13), 4605–4614 (2015)
13. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. *SIAM J. Imaging Sci.* **3**(3), 492–526 (2010)
14. Wu, Y., Feng, X.: Speckle noise reduction via nonconvex high total variation approach. *Math. Probl. Eng.* **20**(15), 11 (2015)
15. Papafitsoros, K., Schönlieb, C.B.: A combined first and second order variational approach for image reconstruction. *J. Math. Imaging Vis.* **48**, 308–333 (2014)
16. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: *Proceeding of International Conference on Computer Vision*, pp. 839–846 (1998)
17. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms with a new one. *SIAM J. Multi-Scale Model. Simul.* **4**(2), 490–530 (2005)
18. Jin, Q., Grama, I., Kervrann, C., et al.: Nonlocal means and optimal weights for noise removal. *SIAM J. Imaging Sci.* **10**(4), 1878–1920 (2017)
19. Kaiming, H., Jian, S., Xiaoou, T.: Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(6), 1397–1409 (2013)
20. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. Image Process.* **16**(8), 2080–2095 (2007)
21. Sochen, N., Kimmel, R., Malladi, R.: A general framework for low level vision. *IEEE Trans. Image Process.* **7**, 310–318 (1998)
22. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.* **158**(2), 460–479 (2013)



Dynamic Delay Based Cyclic Gradient Update Method for Distributed Training

Wenhui Hu^(✉), Peng Wang, Qigang Wang, Zhengdong Zhou, Hui Xiang, Mei Li, and Zhongchao Shi

Artificial Intelligence Lab, Lenovo Research, Beijing 100085, China
{huwh1,wangpeng31,wangqg1,zhouzd2,xianghui1,limeis,shizc2}@lenovo.com

Abstract. Distributed training performance is constrained by two factors. One is the communication overhead between parameter servers and workers. The other is the unbalanced computing powers across workers. We propose a dynamic delay based cyclic gradient update method, which allows workers to push gradients to parameter servers in a round-robin order with dynamic delays. Stale gradient information is accumulated locally in each worker. When a worker obtains the token to update gradients, the accumulated gradients are pushed to parameter servers. Experiments show that, compared with the previous synchronous and cyclic gradient update methods, the dynamic delay cyclic method converges to the same accuracy at a faster speed.

Keywords: Distributed training · Deep learning
Cyclic delayed method · Stochastic optimization

1 Introduction

Deep learning trains deep neural networks with huge volumes of data. The training process is compute-intensive. It can take weeks or months with one modern GPU. Many researchers employ distributed training to accelerate the training process with a server cluster [1].

Model parallelism and data parallelism are two commonly adopted paradigms for distributed training. Model parallelism splits the model into different parts and allocates each part to one GPU [2]. Although model parallelism can speed up the training process with parallel computing, it has two drawbacks which limit its application. The first drawback is scalability, which means it is hard to create a generic model parallelism solution which splits arbitrary model into balanced parts, allocates to adequate GPUs and achieves sublinear scaling ratio. The second drawback is that model parallelism has high communication-to-computation ratio and the communication overhead may counteract the performance gain. Data parallelism is more widely adopted for its simplicity and generality. The training dataset is usually large and easy to split into sub-datasets. Each GPU hosts a replica of the model and trains it with its sub-dataset concurrently.

Various architectures have been proposed for data parallelism, e.g. parameter server (PS) [3], peer-to-peer, ring-based structure [4]. PS architecture has been proved to be effective and are widely adopted [5,6]. There are two entities defined in the PS architecture: parameter servers and workers. Parameter servers are responsible for collecting gradient updates from workers and calculating new model parameters with received gradients. Workers pull latest parameters from parameter servers, train their model replicas with their sub-datasets, calculate gradients, and push gradients to parameter servers. The gradient update method between parameter servers and workers can be roughly classified into synchronous method and asynchronous method. For synchronous method, all workers push gradients to parameter servers in every training iteration. This method is robust, fast and has been proved to be equivalent to the standard stochastic gradient descent (SGD) in single GPU training. But the synchronous method has two issues. One is traffic burst when all workers push gradients at roughly the same time. The other one is that if workers are not homogeneous, the slowest one will slow down the overall training process. Asynchronous methods have been proposed to overcome these issues [7]. However, asynchronous methods may suffer from slower convergence or divergence issues due to stale gradients [8].

In this paper, we propose a method to delay the gradient updates between parameter servers and workers dynamically. Experimental results show that our method increases the distributed training throughput, reduces the network bandwidth requirement, and achieves almost the same accuracy as the synchronous method.

2 Related Works

Many previous works target at reducing the communication overhead in distributed training. Chen et al. [9] propose a double buffering technique which shows the delayed update works well. Seide et al. [10] and Strom et al. [11] use an 1-bit SGD method which adds delay to gradient updates. Lin et al. [12] propose a gradient threshold algorithm, which throttles small gradient updates and accumulates them locally. These gradient sparsification technologies can reduce the communication volume and they are validated by experiments. But the convergence of these implicit delayed methods are not proved in theory. Agarwal et al. [7] propose explicitly delayed gradient update methods to reduce the communication frequency. For convex optimization problem, it has been theoretically proved that the delayed gradient update can be asymptotically negligible and the convergence rate scales as $\mathcal{O}\left(1/\sqrt{nT}\right)$ for n -node cluster after T iterations. However, this cyclic delayed method suffers from the unbalanced worker computing power issues. In the next section, we introduce a dynamic delay based algorithm to overcome these problems and improve the performance.

3 Dynamic Delay Based Cyclic Gradient Update Method

We propose a dynamic delay based cyclic gradient update method, which extends the previous cyclic delayed method. A dynamic delay is applied to the gradient update of each worker. The delay is calculated from the real-time global gradient updating status. This method decouples the cyclic period and actual delays of workers.

The conventional cyclic delayed architecture computes the stochastic gradients in parallel and updates the model parameters in sequence. The worker i computes the gradient $g_i(t - \tau) = \nabla F[x(t - \tau)]$ from the stale parameters $x(t - \tau)$ of τ updates before. The central parameter server obtains $g_i(t - \tau)$ from worker i , computes the updated model $x(t + 1)$ and pushes it back only to worker i . Meanwhile, other workers do their computations on the stale parameters other than the latest $x(t + 1)$. The delay τ comes from the sequential updating of the parameters among the workers, where $\tau = n - 1$ for a n worker cluster in the simplest case. The errors coming from τ is a second order effect, which makes the penalty of delay asymptotically negligible [7].

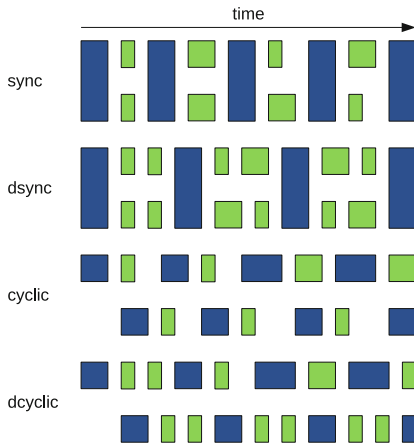


Fig. 1. Runtime illustration with two workers (Color figure online)

The behavior of each worker can be roughly classified into two phases: a communication phase for gradient synchronization, and a computation phase for gradient calculation and accumulation. Although it is possible to overlap part of the backward computation phase with the communication phase for a single worker, the cyclic method focuses more on overlapping the computation phase of a worker with the communication phases of other workers.

In our proposed method, an additional delay is introduced to improve the throughput performance. Each worker maintains an independent local training pool. When a worker is in computation phase, it keeps doing local training and

Algorithm 1.1. Dynamic delay cyclic method

```

1: Initialize  $x_0, t \leftarrow 0$  and  $x_{-1} \leftarrow x_0$ 
2: for all  $n < N$  do
3:    $\tilde{G} \leftarrow 0$ 
4:   while  $t < T$  do
5:     while  $\text{mod}(t, N) \neq n$  do
6:       Wait
7:     end while
8:     Push  $x_t \leftarrow x_{t-1} - \tilde{\eta}_t \tilde{G}$ 
9:     Pull  $\tilde{x} \leftarrow x_t$ 
10:     $t \leftarrow t + 1, \tilde{G} \leftarrow 0$ 
11:    for  $\tilde{d} < D$  do
12:      Compute  $\tilde{g} = \nabla f(\tilde{x})$ 
13:      Accumulate  $\tilde{G} \leftarrow \tilde{G} + \tilde{g}, \tilde{x} \leftarrow \tilde{x} - \tilde{\eta}_t \tilde{g}$ 
14:      if  $\text{mod}(t, N) = n$  then
15:        Break
16:      end if
17:    end for
18:  end while
19: end for

```

gradient accumulation. Mini-batches of the dataset are fetched continuously to train the local model replica. The following communication phase is dynamically postponed until the worker obtains the token. The delay is adaptive, which helps to maintain a good load balance. A powerful worker does more training (and hence processes more training data examples) in its computation phase and a weak worker does less.

Figure 1 shows the runtime illustration of different gradient update methods. Blue blocks denote communication phases, and green blocks denote training operations. The width denotes the duration of different operations in the runtime, which is variational because of the imbalance of workloads. In the synchronous method (sync), a strong worker has to wait for a weak worker in every communication phase. The delayed synchronous method (dsync) postpones the synchronizations with a fixed amount of local computations. This additional delay alleviates the load imbalance and reduces the communication volume. In the cyclic method, the round-robin communication phases prevents the network traffic burst. However the computation phases are not fully utilized if the workers are heterogeneous. Additional computations of the strong workers are introduced in our dynamic delay cyclic method (dcyclic), where the computation phase is prolonged to overlap the communication phase of other workers. The dynamic delay makes full use of every worker’s computation power while minimizing the network traffic.

The dynamic delay cyclic method is described in Algorithm 1.1, where the local variables on the workers are decorated with a tilde. N denotes the number of workers, T denotes the maximum global step, and D denotes the maximum amount of accumulations, i.e. the limitation on the delay of the communication

phase in every training iteration. x denotes the weights, g denotes the gradients and G denotes the accumulations. The global step t serves as the token for the synchronization communication and is maintained by the PS. The subscript -1 of x is introduced for convenience, which is unnecessary in the implementation. Each worker implements two operations. One is the communication operation (remote push-to/pull-from the PS), the other one is the local computation operation (computing/accumulating of gradients).

The communication operation is based on the cyclic delayed method [7]. All workers cooperate in a round-robin order. The worker obtaining the communication token performs the communication operation, including the pushing of gradients and the pulling of updated weights. Then the global step t increases by one, in which case the token is relayed to the next worker.

The dynamic delay occurs in the computation phase of the worker. Compared to the single gradient computation in the conventional cyclic architecture, our dynamic cyclic method enables additional gradient computations and accumulations before the worker obtains the token. In the meanwhile, the amount of accumulations is adaptive in runtime, which is limited by the predefined largest delay D . When $D = 1$, this method falls back to the conventional cyclic delay method [7]. When $D > 1$, local updates and gradient accumulations are activated. In the computation phase when the worker processes new mini-batches, it keeps monitoring the global step t . As soon as it obtains the communication token, the worker aborts the remnant local operations in order to do the communication operation at the earliest.

The dynamic delay cyclic method brings two benefits. One is the optimized throughput (e.g. in examples/second) due to gradient accumulations. By doing as many training as possible in the computation phase, device utilization is improved. As a result, the total processing time for the same quantity of examples is decreased. The other benefit is the convergence conservation. Being able to abort the computation helps to suppress the actual delay and the staleness of gradients, even when the predefined D is large. This helps to achieve the convergence state.

4 Experimental Results

In this section, the dynamic delay cyclic method is evaluated with two large-scale datasets.

4.1 Datasets and Experiment Setup

Two datasets are selected for the evaluations. One is the ILSVRC2012 [13] dataset, which focuses on the image object classification. The training set contains 1.2 million images and the validation set contains 150 thousand images. Both of them are labeled with the presence or absence of 1000 object categories. The ResNet-V2-50 [14] model is adopted for the classification task. The other dataset is the union of the 10^9 Word Parallel Corpus for training and the updated

development set of the News Crawl for validation from the WMT'15 [15]. The training corpus consists of over 22 million sentences, and the validation corpus consists of 3 thousand sentences. Both focus on the recurring translation task on the French–English pair. The Seq2Seq model [16] is adopted for the translation task.

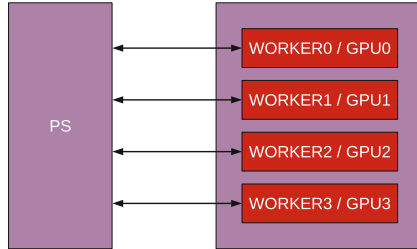


Fig. 2. Experiment setup. Workers are bound to different GPUs inside one node. All workers connect directly to the PS on the other node. The traffic goes over the network in the same manner as a distributed cluster

Two computing nodes are utilized for all experiments. Both nodes are equipped with dual Intel Xeon E5-2600v4 CPUs, 512 GB memory and a Mellanox 40 Gbit/s network adapter. One node has 4 NVIDIA Tesla P100 GPUs, and the other node has no GPU. The distributed computing environment is simulated with these two nodes by making use of the GPU affinity as illustrated in Fig. 2. The worker procedures are bound to different GPUs, in the meanwhile the PS procedure is launched on the other node. PS and workers communicate through the network adapter, in the same way as a real distributed cluster.

4.2 Algorithm and Implementation

We compare our method with the cyclic and the delayed synchronous methods, and take the vanilla synchronous method as the baseline. Workers in the cyclic method update the parameters in a round-robin order [7]. In the vanilla synchronous method, the weights on the PS are updated by gradients received from all workers at around the same time. In the delayed synchronous method, the gradients are accumulated and applied to the local model replicas first. And then the gradient update to the PS works similarly with the vanilla synchronous method.

We implement these four methods with the PS architecture [17], where the server maintains the parameters and the workers do the computations. The data manipulation is automatically managed by TensorFlow [6] from the implicit insertion of nodes to the computation graph.

The ResNet-V2-50 model is trained with the Nesterov accelerated gradient (NAG) method [18, 19] with a batch size of 32, a momentum of 0.9 and a learning rate of 0.005 in 80 epochs. The learning rate is exponentially decayed with a

factor of 0.1 every 20 epochs. The learning rate warmup [20] is implemented in the synchronous methods in order to accelerate the convergence. The vanilla SGD is used to train the Seq2Seq model in 1 epoch with a batch size of 64. The learning rate starts at 0.02 and decays every 0.01 epoch with a decay factor of 0.99. The learning rate warmup is not utilized in the training of the Seq2Seq model.

4.3 Results

We first investigate the performance of different methods. The convergence rates of train (dashed) and validation (solid) are plotted in Fig. 3. The columns from left to right show the synchronous (blue), the delayed synchronous (green), the cyclic (red) and the dynamic delay cyclic (cyan) methods. The top-5 error of the ResNet model is on the top, and the perplexity of the Seq2Seq model is at the bottom. The actual amount of gradient accumulations are tuned to be the same during the training of each model.

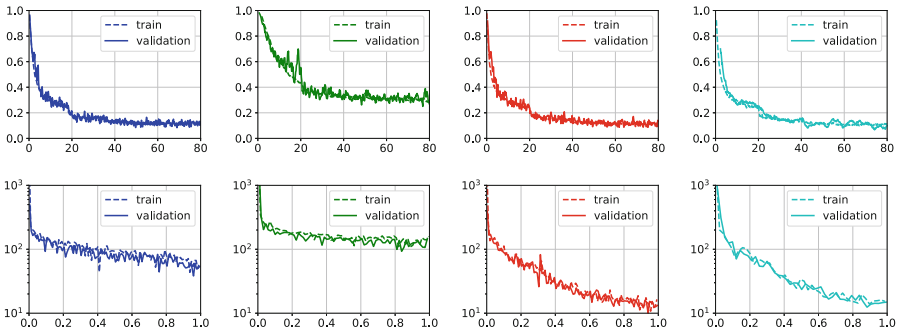


Fig. 3. (Color online) The convergence after definite epochs. The top row presents the top-5 error of the ResNet model, and the bottom shows the perplexity of the Seq2Seq model. The columns indicate the synchronous (blue), the delayed synchronous (green), the cyclic (red) and the dynamic delay cyclic (cyan) methods from left to right. The dashed lines denotes the training and the solid lines denotes the validation. (Color figure online)

In the ResNet model, the cyclic methods achieve the same performance with the synchronous method. The rate of convergence is not impacted by the inherent gradient staleness from the round-robin order. The additional gradient accumulations limit the rate of convergence in the delayed synchronous method. Nevertheless, it takes little effect on the dynamic-cyclic method.

In the Seq2Seq model, the cyclic methods perform better than the synchronous methods, where the perplexity converges quickly to a lower value in the limited number of epochs. The additional accumulations impacts the convergence rate negatively in the delayed synchronous method. The result shows

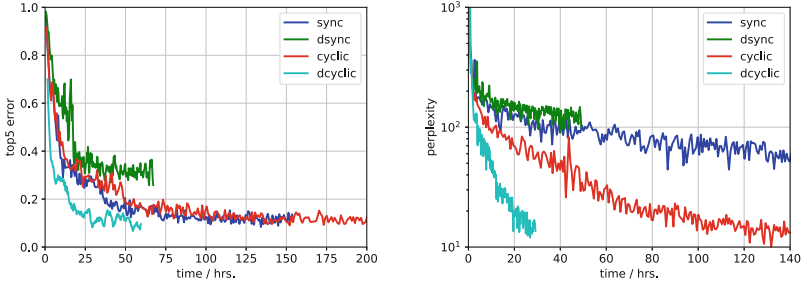


Fig. 4. (Color online) The wall-clock time of different methods. On the left shows the top-5 error of ResNet, and on the right shows the perplexity of Seq2Seq. The vanilla synchronous SGD (blue) is taken as the baseline. The cyclic method (red) finishes after a long time due to the low device utilization. The delayed synchronous SGD (green) obtains slow convergence in the limited number of epochs. The dynamic delay cyclic method (cyan) converges faster in less wall-clock time because of its high throughput. (Color figure online)

that the dynamic delay method is more robust to the staleness of the gradient information than the delayed synchronous method.

The gradient accumulations improve the throughput performance significantly. In the delayed methods, the synchronizations are postponed by the local operations on the workers. This delay reduces the communication-to-computation ratio and increases the utilization of the computing device, which leads to a higher throughput as illustrated in Fig. 4. Large datasets are trained to the same convergence rate at a faster speed with the dynamic delay cyclic method.

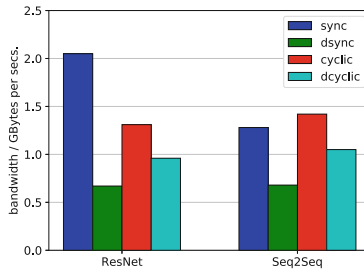


Fig. 5. (Color online) The actual network bandwidth consumption under different update methods. To achieve the same state of convergence, the dynamic delay cyclic method requires less network traffic than the vanilla synchronous and cyclic methods. (Color figure online)

The network traffic is reduced with the dynamic delay cyclic method. The delay reduces the communication frequency and the total communication volume. In the cyclic methods, the PS responds to only one worker at a time.

The rolling of the communication token prevents the traffic burst issue in the synchronous methods and reduces the network requirements. In our experiments, the delay and cyclic methods significantly reduce the network traffic as shown in Fig. 5. The dynamic delay cyclic method preserves the convergence and requires less network traffic than the synchronous and the cyclic methods.

5 Conclusions and Discussions

We propose a dynamic delay based cyclic gradient update method, which benefits from the cyclic gradient update architecture and the local gradient accumulations. The network traffic burst is relieved from the round-robin updating order, and the communication volume and frequency is suppressed by the explicit delay of gradient updates. This method keeps the rate of convergence from the restricted duration between synchronizations, and improves the throughput performance by the dynamic extension of the actual delay. The wall-clock time is reduced in the training of large datasets.

The cyclic methods take full use of the gradients computed from every mini-batch of examples. The gradients are not only employed to update local model replicas, but also accumulated to update the global model on the PS. A fixed (perhaps with decay) learning rate is more applicable for these aggressive methods.

The actual delay is bounded to prevent the convergence problem rising from the gradient staleness. The PS cycles the refresh of the local replicas among all workers in the cluster. The duration of the computation phase scales linearly with the number of workers, which comes from the round-robin nature of the cyclic methods. An oversize delay may limit the convergence rate because of the gradient staleness. An optimized delay restriction should be selected to accelerate the training and preserve the convergence simultaneously.

Acknowledgements. We would like to acknowledge the computation power support from the appliance group in the laboratory. We would also like to thank Mr. Zhenhua Liu from the computer vision group for the fruitful discussion.

References

1. Dean, J., Corrado, G., Monga, R., et al.: Large scale distributed deep networks. In: NIPS (2012)
2. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks. [arXiv:1404.5997](https://arxiv.org/abs/1404.5997) (2014)
3. Li, M., et al.: Scaling distributed machine learning with the parameter server. In: OSDI 2014, pp. 583–598 (2014)
4. Zhang, H., et al.: Poseidon: an efficient communication architecture for distributed deep learning on GPU clusters. [arXiv:1706.03292](https://arxiv.org/abs/1706.03292) (2017)
5. Chen, T., et al.: MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. [arXiv:1512.01274](https://arxiv.org/abs/1512.01274) (2015)

6. Abadi, M., Barham, P., Chen, J.M., et al.: TensorFlow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016), pp. 265–283 (2016)
7. Agarwal, A., Duchi, J.C.: Distributed delayed stochastic optimization. In: NIPS 2011, 4247 (2011)
8. Ho, Q., et al.: More effective distributed ML via a stale synchronous parallel parameter server. In: NIPS 2012, pp. 2141–2149 (2012)
9. Chen, X., Eversole, A., Li, G., Yu, D., Seide, F.: Pipelined back-propagation for context-dependent deep neural networks. In: Interspeech 2012 (2012)
10. Seide, F., Fu, H., Droppo, J., Li, G., Yu, D.: 1-bit stochastic gradient descent and application to data-parallel distributed training of speech DNNs. In: Interspeech 2014 (2014)
11. Strom, N.: Scalable distributed DNN training using commodity GPU cloud computing. In: Interspeech 2015 (2015)
12. Lin, Y., Han, S., Mao, H., Wang, Y., Dally, W.J.: Deep gradient compression: reducing the communication bandwidth for distributed training. In: ICLR 2018 (2018)
13. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015)
14. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034 (2015)
15. Bojar, O., Buck, C., Federmann, C., et al.: Findings of the 2015 workshop on statistical machine translation. In: Tenth Workshop on Statistical Machine Translation (2015). <http://www.statmt.org/wmt15>
16. Vinyals, O., Kaiser, L., Koo, T., et al.: Grammar as a foreign language. In: NIPS 2015, pp. 2773–2781 (2015)
17. Li, M., Andersen, D.G., Smola, A.J., Yu, K.: Communication efficient distributed machine learning with the parameter server. In: NIPS 2014, pp. 19–27 (2014)
18. Qian, N.: On the momentum term in gradient descent learning algorithms. *Neural Netw.* **12**(1), 145–151 (1999)
19. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Soviet Math. Doklady* **27**(2), 372–376 (1983)
20. Goyal, P., Dollár, P., Girshick, R., et al.: Accurate, large minibatch SGD: training ImageNet in 1 hour. [arXiv:1706.02677](https://arxiv.org/abs/1706.02677) (2017)



Semi-supervised Dictionary Active Learning for Pattern Classification

Qin Zhong^{1,2}, Meng Yang^{1(✉)}, and Tiancheng Zhang²

¹ School of Data and Computer Science, Sun Yat-sen University,
Guangzhou, China

zhongqin0820@163.com, yangm6@mail.sysu.edu.cn

² School of Computer Science and Engineering, Northeastern University,
Shenyang, China

tczhang@mail.neu.edu.cn

Abstract. Gathering labeled data is one of the most time-consuming and expensive tasks in supervised machine learning. In practical applications, there are usually quite limited labeled training samples but abundant unlabeled data that is easy to collect. Semi-supervised learning and active learning are two important techniques for learning a discriminative classification model when labeled data is scarce. However, unlabeled data with significant noises and outliers cannot be well exploited and usually worsen the performance of semi-supervised learning and the performance of active learning also needs a powerful initial classifier learned from the quite limited labeled training data. In order to solve the above issues, in this paper we proposed a novel model of semi-supervised dictionary active learning (SSDAL), which aims to integrate semi-supervised learning and active learning to effectively use all the training data. In particular, two criterions based on estimated class possibility are designed to select the unlabeled data with confident class estimation for semi-supervised learning and the informative unlabeled data for active learning, respectively. Extensive experiments are conducted to show the superior performance of our method in classification applications, e.g., handwritten digit recognition, face recognition and large-scale image classification.

Keywords: Semi-supervised learning · Dictionary learning · Active learning
Pattern classification

1 Introduction

Considering the explosion of digital images in the real world, it is necessary to collect, classify and organize them in a simple, fast and efficient way. In order to use these increasing images as labeled data, automatic image annotation [28] is proposed by establishing statistical models, which can significantly reduce the labor cost of manually annotating images. However, statistical models, which need a large amount of labeled training samples, are not applicable for the case with a quite limited labeled data. How to build an accurate classification model with limited labeled samples for multi-class classification is still an open question.

Semi-Supervised learning (SSL) [1–4, 12] are potential solutions to the problem with a quite limited labeled data. SSL utilizes unlabeled samples to enhance the generalization ability of supervised learning. Classical SSL algorithms include Co-Training [2], graph-based semi-supervised learning [3], semi-supervised support vector machines (S3VM) [4] and semi-supervised dictionary learning (SSDL) [5–10, 12]. Recently promising performance has been achieved by jointly learning a dictionary based classifier and the class estimation of unlabeled data. However, it has been pointed by [11] that directly using unlabeled samples may significantly reduce classification performance when there are large amounts of noisy samples and outliers in the unlabeled data.

In order to effectively adopt the unlabeled training samples, which disturb semi-supervised learning methods due to their noise and variations, active learning (AL) [11, 29, 30] methods attract much attention recently. AL trains the model in an interactive way, which is capable of selecting the representative data based on the classification model learned in different iterations. However, the performance of AL quite depends on the effectiveness of the initial classifier.

Semi-supervised learning and active learning are not perfect alone but complementary to each other together. The classifier obtained by SSL, which takes both the labeled and unlabeled samples into account, can act as a good initial classifier; the introduction of AL can eliminate the problem of the model performance reduction caused due to the presence of a large number of noise samples and outliers in the unlabeled samples. Meanwhile, the introduction of AL can also gradually get labeled samples from the unlabeled data set for training without the need to prepare the required large-scale labeled datasets at the beginning. Several methods have been developed to study how to effectively combine SSL and AL. Song et al. [13] proposed an active learning method based on co-training in video annotation. Jiang et al. [14] developed a graph-based SSL method for video concept detection and used active learning to select data-concept pairs for human annotation. Although these combinations have improved the performance, the recently developed powerful semi-supervised dictionary learning (SSDL) models are not well exploited and how to jointly integrate SSDL and AL is still an open question.

In order to solve above issues, in this paper we proposed a novel framework of semi-supervised dictionary active learning (SSDAL) to effectively integrate semi-supervised dictionary learning (SSDL) and active learning (AL). Initially, we use a handful of labeled samples and abundant unlabeled samples to train a SSDL model. Based on that, we introduce AL algorithm to select the informative samples to boost the training. Compared to the original SSDL model, it is not necessary to prepare all the labeled samples at the beginning. Compared with the simple AL algorithm, it has a great advantage in learning from less labeled data and more unlabeled data. The experimental results on the benchmark datasets clearly show the superior performance of the proposed

To summarize, the main contributions of our work are as follows:

- A novel semi-supervised dictionary active learning (SSDAL) framework is proposed to integrate the advantages of SSDL and AL for the first time.

- The representative unlabeled samples selected by AL and the unlabeled samples with confident class estimation are complementary to each other.
- Experiments on the benchmark datasets are conducted, with remarkable performance reported.

The rest of the paper is organized as follows. Section 2 presents a brief review of related work. Section 3 overviews the pipeline of our framework, followed by a discussion of model formulation and optimization in Sect. 4. The experimental results are presented in Sect. 5. Section 6 concludes the paper.

2 Related Work

2.1 Semi-supervised Dictionary Learning

Owing to the impressive performance of sparse representation and dictionary learning [16, 17, 31–34], semi-supervised dictionary learning (SSDL) algorithms [5–10, 12] have been proposed recently.

Most of SSDL methods aim to learn a shared dictionary. Pham et al. [5] incorporated the reconstruction error of both the labeled and unlabeled data with sparsity constraint into a joint objective function. Zhang et al. [6] proposed an online semi-supervised dictionary learning model, in which the reconstruction error of both labeled data and unlabeled data, label consistency and the classification error were integrated into a joint model. Wang et al. [9] proposed a robust dictionary learning method by exploiting the global structure of all labeled and unlabeled data. In these semi-supervised dictionary methods mentioned above, the unlabeled training data is only used to learn a shared dictionary, ignoring to explore the discrimination hidden in the unlabeled data.

In order to utilize the class information of unlabeled data, Shrivastava et al. [7] learnt a class-specific semi-supervised dictionary with estimating the class possibility of unlabeled data. Wang et al. [10] proposed an adaptively unified semi-supervised dictionary learning model which integrated the reconstruction error of both the labeled data and unlabeled data, and classifier learning into a unified framework. Vu et al. [27] proposed a shared dictionary learning by grouping the unlabeled samples via using the coefficient-based relationship between the labeled and unlabeled samples. The methods above try to exploit the discrimination hidden in the unlabeled data. However, the class probability of unlabeled training samples is artificially designed but not derived from the objective function. And the powerful class specific representation ability cannot be used in the shared dictionary learning model.

Recently, Yang et al. [12] proposed a discriminative semi-supervised dictionary learning (DSSDL) method, which achieves superior performance by introducing a regularization of entropy and using an extended dictionary to explore the discrimination embedded in the unlabeled data. However, there are some representative samples (e.g., nearby the border of different classes), which cannot be correctly estimated by DSSDL, preventing the further improvement of DSSDL.

2.2 Active Learning

Active learning (AL) has been widely studied in [11, 29, 30] for its ability to reducing human labor. In the view of sampling strategy, active learning can be roughly divided into three categories [28]: (i) membership query synthesis, (ii) stream-based selective sampling, and (iii) pool-based sampling.

Membership query synthesis assumes that the system can interact with the surrounding environment, e.g., the annotator can be asked to determine the category of some samples and learn the unknown concepts. But the disadvantage of this method is that all unlabeled samples are labeled by the annotator without considering the actual distribution of samples. To solve this issue with a large scale of unlabeled data, stream-based selective sampling introduced. Although the stream-based selective strategy can solve the problems caused by direct query methods to some extent, it often needs to set a fixed threshold to measure the information content of the sample, thus lack the universality of different tasks. Moreover, because of the way it compares, the actual distribution of unlabeled data sets and the difference between the unlabeled data can not be obtained [28].

Pool-based sampling active learning is proposed to overcome the drawbacks above. Lewis et al. [29] solved this by proposing pool-based sampling, which compares the information of unlabeled samples, and then selects the sample with the highest amount of information to ask the annotator. Since the pool-based sampling strategy has inherited the previous two methods and overcome the shortcomings of the above two methods, it has become the most widely studied and used sampling strategy [29, 30]. It has also pointed out by Lin et al. [30] that the sample selection criterion is the another key in AL algorithm, and there exists many sample selection criteria including risk reduction, uncertainty, diversity and so on [28]. The criteria is typically defined according to the classification uncertainty of samples. Specifically, the samples of low classification confidence, together with other informative criteria like diversity, are generally treated as the candidates for model retraining. The accuracy of progressively selecting uncertain unlabeled sample depends on the recognition ability of the desired classifier, which needs to perform well in the case with limited labeled training data.

3 Semi-supervised Dictionary Active Learning

We propose a novel SSDL-based active learning framework which is composed of a SSDL model and an active learning algorithm. Figure 1 illustrates the overall framework. Initially, the training set includes a limited labeled samples and abundant unlabeled samples. Next, we use semi-supervised dictionary learning to train a dictionary, which is supposed to have a good representative ability with a small within-class variation but a bad interclass representative ability. Then we select the most informative sample through active learning technique to retrain the proposed model. For the most informative sample, we introduce a user to annotate it and add it into labeled data set for the next dictionary training until the model converges.

3.1 Model of SSDAL

As many prevailing semi-supervised dictionary learning models [5–10, 12], we focus on the case that the identity of unlabeled training data lies in the training set. In order to overcome the drawbacks of the prevailing semi-supervised learning (e.g., its performance will be worsened by the unlabeled noisy samples and outliers) and active learning (e.g., a powerful initial classifier is needed), we proposed a novel model of semi-supervised dictionary active learning to fully exploit the benefits of both of semi-supervised dictionary learning [12] and active learning.

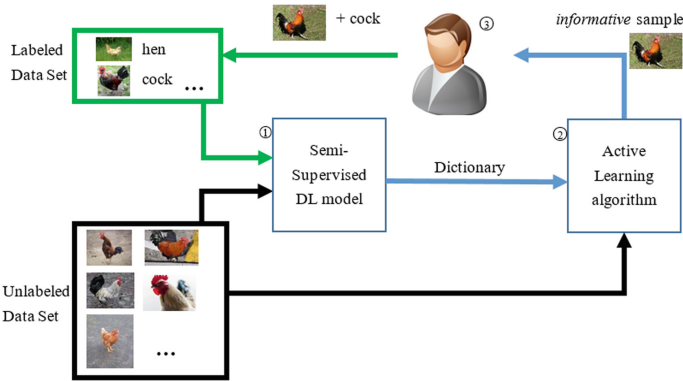


Fig. 1. Illustration of our proposed SSDAL framework. Firstly, the SSDL model is learned with quite limited labeled samples and all of the unlabeled samples. Secondly, we use AL algorithm to select the most *informative* samples iteratively from the unlabeled data set. Thirdly, we introduce a user to label those *informative* samples and add them into labeled data set to update the model with the new labeled samples and the rest of unlabeled data.

Given data points set $A = [A_1, \dots, A_i, \dots, A_C, B]$ where A_i denotes the i^{th} -class training data and each column of A_i is a training sample while the remaining $B = [b_1, \dots, b_i, \dots, b_N]$ is the N unlabeled training samples from class 1 to C . Let $D = [D_1, \dots, D_i, \dots, D_C]$ denote the supervised dictionary initialized by A , while $E = [E_1, \dots, E_i, \dots, E_C]$ is an extended dictionary that mainly explore the discrimination of unlabeled training data. Both D_i and E_i are associated to class i , and they are required to well represent i^{th} -class data but with a bad representation ability for all the other classes. As $P_{i,j}$ indicates the probabilistic relationship between the j^{th} -unlabeled training sample and i^{th} -class. The model of our proposed SSDAL framework is:

$$\begin{aligned}
 & \min_{\hat{\mathbf{D}}, \mathbf{P}, \mathbf{X}} \sum_{i=1}^C \left(\|\mathbf{A}_i - \hat{\mathbf{D}}_i \mathbf{X}_i^i\|_F^2 + \gamma \|\mathbf{X}_i^i\|_1 + \lambda \|\mathbf{X}_i^i - \mathbf{M}_i\|_F^2 \right) \\
 & + \sum_{j=1}^{N-L} \left(\sum_{i=1}^C \mathbf{P}_{i,j} \|\mathbf{b}_j - \hat{\mathbf{D}}_i \mathbf{y}_j^i\|_F^2 + \gamma \|\mathbf{y}_j^i\|_1 \right) \\
 & - \beta \left(- \sum_{j=1}^{N-L} \sum_{i=1}^C \mathbf{P}_{i,j} \log \mathbf{P}_{i,j} \right)
 \end{aligned} \tag{1}$$

s.t. semi supervised learning for confident estimation
active learning for unconfident class estimation

where \mathbf{X}_i^i and \mathbf{y}_j^i are the coding coefficient matrix of \mathbf{A}_i and unlabeled data \mathbf{b}_j on the class-specific dictionary $\hat{\mathbf{D}}_i = [\mathbf{D}_i \mathbf{E}_i]$, respectively.

The confidence of the estimated class possibility can be measured by the entropy

$$H(\mathbf{b}_i) = - \sum_{i=1}^C \mathbf{P}_{i,j} \log \mathbf{P}_{i,j} \tag{2}$$

The entropy value of Eq. (2) indicates the uncertainty of the class estimation. For instance, if the unlabeled data is definitely assigned to some class (e.g., $\mathbf{P}_{i,j} = 1$ for some j when the sample is assigned to the i^{th} class, and $\mathbf{P}_{i,j} = 0$ for $j \neq i$), the entropy value will be zero.

3.2 Semi-supervised Dictionary Learning

When the class estimation is confident, the proposed SSDAL model changes to

$$\begin{aligned}
 & \min_{\hat{\mathbf{D}}, \mathbf{P}, \mathbf{X}} \sum_{i=1}^C \left(\|\mathbf{A}_i - \hat{\mathbf{D}}_i \mathbf{X}_i^i\|_F^2 + \gamma \|\mathbf{X}_i^i\|_1 + \lambda \|\mathbf{X}_i^i - \mathbf{M}_i\|_F^2 \right) \\
 & + \sum_{j=1}^{N-L} \left(\sum_{i=1}^C \mathbf{P}_{i,j} \|\mathbf{b}_j - \hat{\mathbf{D}}_i \mathbf{y}_j^i\|_F^2 + \gamma \|\mathbf{y}_j^i\|_1 \right) \\
 & - \beta \left(- \sum_{j=1}^{N-L} \sum_{i=1}^C \mathbf{P}_{i,j} \log \mathbf{P}_{i,j} \right)
 \end{aligned} \tag{3}$$

$$\text{s.t. } H(\mathbf{b}_j) < T$$

where T is a threshold, which is usually set as 0.5. In the dictionary learning, we only use the unlabeled data whose entropy is smaller than the threshold, i.e., their class estimation is relatively confident.

3.3 Active Learning

Considering the combination of active learning, Let $\hat{\mathbf{D}}$ denote the output of $[\mathbf{DE}]$ in Eq. (1). Set L as the number of labeled samples for active learning. In model's iteration, we can get the probabilistic outputs \mathbf{P} for all the unlabeled samples and a class-specific dictionary $\hat{\mathbf{D}} = [\mathbf{DE}]$. If we want to boost the performance of our model by acquiring

some labeled examples, the main issue is how to select the most valuable examples to query the user for labels. Considering that the SSDL model can naturally provide the probabilistic outputs, which is convenient to measure the uncertainty of all unlabeled samples, we adopt the uncertainty measurement to select the most uncertain samples.

For the unlabeled data, there are C candidate classes. Therefore, the semi-supervised dictionary learning provides C classifiers. When multiple learners exist, a widely applied strategy is to select the samples that have the maximum disagreement amongst them. Here the disagreement of multiple learners can also be regarded as an uncertainty measure, and this strategy is categorized into the uncertainty criterion as well. Inspired by [15], we use the uncertainty estimation method that considers the posterior probabilities of the best and the second best predictions, that is,

$$\text{Uncertainty}(\mathbf{x}) = P(\mathbf{c}_1|\mathbf{x}) - P(\mathbf{c}_2|\mathbf{x}) \quad (4)$$

where \mathbf{c}_1 and \mathbf{c}_2 are the classes with the largest and second largest posterior class probabilities, respectively. If their margin is small, it means that the model is more confused on the sample and thus it is with high uncertainty. We use Eq. (3) as the final sample selection strategy in the active learning.

3.4 Classification Model

We utilize different coding models when dealing with the testing sample, e.g., collaborative representation of Eq. (5) for face recognition and the large scale image classification, while local representation of Eq. (6) is used in digit recognition [12].

$$\text{Code_Classify}(\mathbf{b}_j, \hat{\mathbf{D}}) = \underset{y_j}{\text{argmin}}_y \|\mathbf{b}_j - \hat{\mathbf{D}}\mathbf{y}_j\|_F^2 + \gamma \|\mathbf{y}_j\|_1 \quad (5)$$

$$\text{Code_Classify}(\mathbf{b}_j, \hat{\mathbf{D}}) = \underset{y_j^i}{\text{argmin}}_{y_j^i} \|\mathbf{b}_j - \hat{\mathbf{D}}_i \mathbf{y}_j^i\|_F^2 + \gamma \|\mathbf{y}_j^i\|_1 \quad \forall i \quad (6)$$

where $\mathbf{y}_j = [\mathbf{y}_j^1, \dots, \mathbf{y}_j^i, \dots, \mathbf{y}_j^c]$ is the coding vector on the whole dictionary, $\hat{\mathbf{D}} = [\mathbf{D}\mathbf{E}]$ is the learned structured dictionary associated with class i , and \mathbf{y}_j^i is the coding vector associated to i^{th} class of the j^{th} unlabeled data. Then the final classification is conducted by

$$\text{identity}(\mathbf{b}) = \underset{i}{\text{arg min}} \{e_i\} \quad (7)$$

where $e_i = \|\mathbf{b} - \hat{\mathbf{D}}_i \mathbf{y}_j^i\|_2^2$.

4 Optimization of SSDAL

The optimization of SSDAL is an alternative solving procedure, which includes the selection of unlabeled data and the semi-supervised dictionary learning of Eq. (3). And the semi-supervised dictionary learning can further be divided into two sub-problems by doing class estimation of unlabeled data and discriminative dictionary learning

alternatively: updating P by fixing D, E and X , while updating D, E and X alternatively by fixing P [12]. These processes enable the model to converge.

Selection of Unlabeled Data. With the class estimation of all unlabeled data, the ones with confident class estimation will be integrated into the model of discriminative semi-supervised dictionary learning.

For the unlabeled data with unconfident class estimation, we select the most informative samples from the rest of unlabeled data set iteratively via Eq. (4). Then, we introduce a user to label those informative samples and then add them into the annotated dataset.

Update P . By fixing the class-specific dictionary and the corresponding coding coefficient (e.g., D, E, X and y), and let $\mathbf{e}_j^i = \|\mathbf{b}_j - \hat{D}\mathbf{y}_j^i\|$. The class probability of \mathbf{j}^{th} unlabeled training sample is

$$P_{i,j} = \exp\{-\mathbf{e}_j^i/\beta\} / \sum_{i=1}^C \exp\{-\mathbf{e}_j^i/\beta\} \quad (8)$$

Update D, E and X . The unlabeled data, which are not included into the active learning or don't have a confident estimation, their probability of class will be set as zero, i.e., $P_{i,j} = 0$. Then the proposed SSDAL changes to

$$\begin{aligned} \min_{\hat{D}, X} \sum_{i=1}^C \left(\|\mathbf{A}_i - \hat{D}\mathbf{X}_i\|_F^2 + \gamma \|\mathbf{X}_i\|_1 + \lambda \|\mathbf{X}_i - \mathbf{M}_i\|_F^2 \right) \\ + \sum_{j=1}^{N-L} \left(\sum_{i=1}^C P_{i,j} \|\mathbf{b}_j - \hat{D}\mathbf{y}_j^i\|_F^2 + \gamma \|\mathbf{y}_j^i\|_1 \right) \end{aligned} \quad (9)$$

which can efficiently solved by using the method in Yang et al. [12].

5 Experiments

In this section, extensive experiments were conducted over on the benchmark datasets, such as LFW [24], Web Vision 1.0 [25], USPS [22] and MNIST [23] to demonstrate the effectiveness of our proposed semi-supervised dictionary active learning (SSDAL). The competing methods include several representative supervised dictionary learning methods: SRC [18], FDDL [19], DKSVD [20], LCKSVD [26] and semi-supervised dictionary learning methods: JDL [5], OSSDL [6], S2D2 [7], SSRD [9], SSP-DL [21] and recently proposed DSSDL [12] algorithm. Here we don't include deep learning related models because our base classifier is a dictionary learning related model and the number of labeled samples is too limited to train a good enough deep learning model. The coding of unlabeled training data and testing data in our proposed framework adopts the same coding representation.

The SSDL model used in our framework has three super parameters, λ, γ and β . We set them as $\lambda = 0.01, \gamma = 0.001, \beta = 0.01$ in all experiments as same as [12].

We evaluate the performance of our proposed SSDAL in the classification accuracy with the same amount of user annotation totally. The classification accuracy is defined

as the top one rate for digit recognition and face recognition, with an extra top-5 rate in Web Vision large-scale image classification task.

5.1 Datasets and Results

Face Identification. Following the same experimental setting in [10], we estimate our proposed framework in the LFW database [24], which is a large-scale database consists of 4,174 face images of 143 individuals taken under varying pose, expression, illumination, misalignment and occlusion conditions. Each individual has no less than 11 images and we select the first 10 samples for training data with the remaining samples for testing. We randomly select 2 samples from each class as the initial labeled data, then we set 5 times of user-query iteration, which makes the final amount of labeled data as same as other methods. As shown in Fig. 2, the data is divided into 3 parts, the data not used, the training data, and the test data.



Fig. 2. Illustration of how the data is divided. In this experiment, firstly, the data is randomly divided into 3 parts during the whole training process. Secondly, for training data, we randomly select 2 of them as the initial labeled data (i.e., orange frame) and the rest as unlabeled data. Then, we gradually add the labeled data (i.e., green frame) from the rest of the unlabeled data (i.e., red frame) via AL algorithm to boost our model. After all, we use testing data to test our model. (Color figure online)

We use the same feature in [12] which reduces the feature vectors to 500 dimension. Table 1 lists the identification results of the LFW database, which show clearly that our proposed method achieves the highest recognition rates with the same amount of labeled data among the competing schemes. Compare to DSSDL, the improvement of the performance stems from the integration of active learning algorithm, which can select the most informative samples and no need to get all the labeled data ready.

Digit Recognition. Use the same experimental setting in [12], we evaluate the performance on both the USPS dataset [22] and MNIST dataset [23]. In the USPS dataset, there are 9,298 digital images consisting of 10 classes. We randomly select 110 images

Table 1. The recognition rates (%) on LFW database.

Methods	LFW
SRC	62.2 \pm 2.7
DKSVD	56.7 \pm 1.8
LC-KSVD	58.6 \pm 1.3
FDDL	66.1 \pm 1.5
JDL	64.8 \pm 2.1
S2D2	65.4 \pm 2.1
DSSDL	67.5 \pm 1.2
SSDAL	72.0 \pm 0.7

from each class and then randomly select 2 images as the labeled samples for the initial dictionary training, 58 images as the unlabeled samples and the left as the testing samples. For MNIST dataset, there are 10 classes and 70,000 handwritten digital images totally, 60,000 for training and 10,000 for testing respectively. But we randomly select 200 samples from each class then we randomly select 2 images each class as the labeled samples for the initial dictionary training, 98 images as the unlabeled, and 100 images as the testing samples. The feature we used is the whole image, which was normalized to have unit l_2 -norm. We set 18 times user-query iteration, which with 10 labels updated in each iteration. This makes the final labeled data amount as same as other methods, which use 20 labeled images per class for training.

All relevant results for ten independent tests are listed in Table 2, which calculates the mean accuracy and standard deviation. It can be seen that the proposed SSDAL is able to find the informative samples from the unlabeled dataset for next round training and can then utilize information of the selected unlabeled data to improve the classification accuracy. Compare to all the competing methods, our proposed SSDAL achieves the best performance.

Table 2. The recognition rates (%) on USPS and MNIST

Methods	USPS	MNIST
SRC	68.6 \pm 2.7	72.9 \pm 2.3
DKSVD	67.5 \pm 1.8	71.4 \pm 1.7
FDDL	85.2 \pm 1.2	82.5 \pm 1.3
LC-KSVD	76.9 \pm 1.3	73.0 \pm 1.3
OSSDL	80.8 \pm 2.8	73.2 \pm 1.8
S2D2	86.6 \pm 1.6	77.6 \pm 0.8
SSR-D	87.2 \pm 0.5	83.8 \pm 1.2
SSP-DL	87.8 \pm 1.1	85.8 \pm 1.2
DSSDL	90.2 \pm 0.9	88.3 \pm 1.5
SSDAL	90.9 \pm 0.9	88.8 \pm 1.1

Web Vision Database 1.0. Web Vision database 1.0 [25] is larger than all the database we evaluated. We use a subset with the same number of classes (i.e., 1,000 classes) as the dataset, which contains 50 samples in each class. For each class, we randomly set 30 samples for train and 20 samples for test. From the training set, we select the first 5 samples as the initial labeled data. Next we set 8 times of user-query iteration. This makes it 13 labeled samples for each class finally.

We extract feature as same as [25] then we reduced it to 300 dimension. The top-1 result and top-5 result of the proposed SSDAL and two most competing methods, such as the supervised LCKSVD and the semi-supervised DSSDL. The results of all methods are listed in Table 3, from which we can observed that the improvements of SSDAL over DSSDL are 1.3% in Top-1 accuracy and 2.7% in Top-5 accuracy. Compared to LCKSVD, the advantages of SSDAL is larger.

Table 3. The recognition rates (%) on web Vision sub-database.

Methods	Top-1	Top-5
LCKSVD	52.6 ± 1.4	71.4 ± 2.9
DSSDL	56.0 ± 1.1	78.0 ± 1.1
SSDAL	57.3 ± 0.9	80.7 ± 1.2

6 Conclusions

In this paper, we proposed a new model of semi-supervised dictionary active learning (SSDAL), which integrates the state-of-the-art semi-supervised dictionary learning and active learning for the first time. Based on the proposed criterion which based on the estimated class possibility, the unlabeled data with confident class estimation and the representative information are returned into the training of SSDAL. Extensive experiments have shown the superior performance of our proposed framework.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China (Grant no. 61772568), the Guangzhou Science and Technology Program (Grant no. 201804010288), the Fundamental 535 Research Funds for the Central Universities (Grant no. 18lgzd15), the Shenzhen Scientific Research and Development Funding Program (Grant no. JCYJ20170302153827712).

References

1. Zhu, X.: Semi-supervised learning literature survey. Technical report 1530, Wisconsin-Madison (2005)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT (1998)
3. Zhu, X.: Semi-supervised learning with graphs. In Proceedings of IJCNLP (2005)
4. Sindhvani, V., Keerthi, S.S.: Large scale semi-supervised linear SVMs. In: ACM SIGIR (2006)

5. Pham, D.-S., Svetha, V.: Joint learning and dictionary construction for pattern recognition. In: Proceedings of CVPR (2008)
6. Zhang, G., Jiang, Z., Davis, L.S.: Online semi-supervised discriminative dictionary learning for sparse representation. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 259–273. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_20
7. Shrivastava, A., Pillai, J.K., Patel, V.M., Chellappa, R.: Learning discriminative dictionaries with partially labeled data. In: Proceedings of ICIIP (2012)
8. Babagholami-Mohamadabadi, B., Zarghami, A., Zolfaghari, M., Baghshah, M.S.: PSSDL: probabilistic semi-supervised dictionary learning. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS (LNAI), vol. 8190, pp. 192–207. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40994-3_13
9. Wang, H., Nie, F., Cai, W., Huang, H.: Semi-supervised robust dictionary learning via efficient $l_{2,0}$ -norms minimization. In: Proceedings of ICCV (2013)
10. Wang, X., Guo, X., Li, S.: Adaptively unified semisupervised dictionary learning with active points. In: Proceeding of the ICCV (2015)
11. Li, Y.-F., Zhou, Z.-H.: Towards making unlabeled data never hurt. IEEE Trans. Pattern Anal. Mach. Intell. **37**(1), 175–188 (2015)
12. Yang, M., Chen, L.: Discriminative semi-supervised dictionary learning with entropy regularization for pattern classification. In: AAAI (2017)
13. Song, Y., Hua, X.-S., Dai, L.-R., Wang, M.: Semi-automatic video annotation based on active learning with multiple complementary predictors. In: MIR, pp. 97–104 (2005)
14. Jiang, W., Loui, A.: Laplacian adaptive context-based SVM for video concept detection. In: ACMSIGMM Workshop, pp. 15–20 (2011)
15. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009. IEEE, pp. 2372–2379 (2009)
16. Yang, M., Dai, D., Shen, L., Gool, L.V.: Latent dictionary learning for sparse representation based classification. In: Proceedings of CVPR (2014)
17. Yang, M., Zhang, L., Yang, J., Zhang, D.: Metaface learning for sparse representation based face recognition. In: Proceedings of ICIIP (2010)
18. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE TPAMI **31**(2), 210–227 (2009)
19. Yang, M., Zhang, L., Feng, X.: Fisher discrimination dictionary learning for sparse representation. In: Proceedings of ICCV (2011)
20. Zhang, Q., Li, B.: Discriminative K-SVD for dictionary learning in face recognition. In: Proceedings of CVPR (2010)
21. Wang, D., Zhang, X., Fan, M., Ye, X.: Semi-supervised dictionary learning via structural sparse preserving. In: Proceedings of AAAI (2016)
22. Hull, J.: A database for handwritten text recognition research. IEEE TPAMI **16**(5), 550–554 (1994)
23. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradientbased learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
24. Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009. LNCS, vol. 5995, pp. 88–97. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12304-7_9
25. Li, W., Wang, L., Li, W., et al.: WebVision database: visual learning and understanding from web data (2017)
26. Jiang, Z., Lin, Z., Davis, L.S.: Label consistent K-SVD: learning a discriminative dictionary for recognition. IEEE TPAMI **35**(11), 2651–2664 (2013)

27. Vu, T.H., Monga, V.: Learning a low-rank shared dictionary for object classification. In: International Conference on Image Processing (ICIP) (2016)
28. Settles, B.: Active learning literature survey. Computer Sciences Technical report 1648, University of Wisconsin–Madison (2009)
29. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR 1994, pp. 3–12. Springer, London (1994). https://doi.org/10.1007/978-1-4471-2099-5_1. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval
30. Lin, L., Wang, K., Meng, D., et al.: Active self-paced learning for cost-effective and progressive face identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**(99), 7–19 (2017)
31. Jing, X.Y., Zhang, D.: A face and palmprint recognition approach based on discriminant DCT feature extraction. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **34**(6), 2405 (2004)
32. Jing, X.Y., Zhu, X., Wu, F., et al.: Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. *IEEE Trans. Image Process.* **26**(3), 1363–1378 (2017)
33. Zhu, X., Jing, X.Y., You, X., et al.: Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix. *IEEE Trans. Inf. Forensics Secur.* **PP**(99), 1 (2017)
34. Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image. In: AAAI 2017, pp. 4341–4348 (2017)



Multi-feature Shared and Specific Representation for Pattern Classification

Kangyin Ke and Meng Yang^(✉)

School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China
keky@mail2.sysu.edu.cn, yangm6@mail.sysu.edu.cn

Abstract. Sparse representation has been widely applied to pattern classification, where the input is coded as a sparse linear combination of training samples and classified to a category with the minimum reconstruction error. In the recent years, multi-feature representation based classification has attracted widespread attention and most of these methods have showed the superiorities compared to the classification model with single feature. One key issue in multi-feature representation is how to effectively exploit the similarity and distinctiveness of different feature, which is still an open question. In this paper, we present a novel multi-feature shared and specific representation (MFSSR) model, which not only keeps the distinctiveness of different features, but further exploits their similarity with a shared representation coefficient. In addition, different features are weighted differently to reflect their discriminative abilities. Several representative experiments have shown the effectiveness and simplicity of the proposed MFSSR.

Keywords: Multi-feature representation
Shared and specific representation · Pattern classification

1 Introduction

Over the past decade, sparse representation has achieved great success [14] and has been widely applied to various applications, such as face recognition [21, 22, 24], image classification [25], signal classification [18], and image restoration [12]. The main idea of sparse representation is to approximate a testing sample by a linear combination of training samples and the representation coefficients should be sparse to some extent. Both l_0 -norm (i.e., the number of non-zero elements) and l_1 -norm (i.e., the count of absolute values of elements) minimizations can be applied to implement sparsity coding. However, considering l_0 -norm minimization is an NP-hard problem, most of sparse representation methods employ l_1 -norm minimization. And the simplest and standard sparse representation can be regarded as the following regularized linear problem.

$$\min_{\alpha} \|\mathbf{y} - \mathbf{D}\alpha\|_2^2 + \lambda \|\alpha\|_1 \quad (1)$$

where \mathbf{y} is the feature vector of testing sample, $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_i, \dots, \mathbf{d}_n]$ is the dictionary in which \mathbf{d}_i is the feature vector of the i^{th} training sample, and λ is a positive scalar as sparsity penalty parameter.

Based on Eq. (1), Wright et al. [22] proposed the sparse representation based classification (SRC) method for robust face recognition, which not only achieves high classification accuracy, but is also robust to face occlusion. In SRC, the query face image is approximated by a linear combination of all the training samples, and then classified to the category i with the lowest reconstruction error:

$$i = \min_i \|\mathbf{y} - \mathbf{D}_i \hat{\alpha}_i\|_2^2 \quad (2)$$

where \mathbf{D}_i is the sub-dictionary of i^{th} class, $\hat{\alpha}_i$ is the coefficient associated with the sub-dictionary \mathbf{D}_i .

Compared to single feature, multiple features obviously can provide more effective information [6] to recognize the class of testing sample. Amount of works [4, 15] have shown the benefits of complementary information provided by different features. Hence a mass of methods have sprung up to deal with multi-feature representation based classification problems, which can be divided into two categories, namely classifier fusion [17] and feature fusion [2].

Multi-feature sparse representation based methods have emerged in recently years. In [26], a multi-task joint sparse representation based classification method (MTJSRC) is proposed, which adopts a mixed norm regularization on the representation coefficients to enforce the similarity of different features and the sparsity of classes. Considering that multiple features may have different contributions for the representation and classification, [23] proposed a relaxed collaborative representation (RCR) model with a weighted within-class regularization on the coding coefficient. Although promising performance of RCR is reported, the within-class regularization has no direct connection with the final classifier. Very recently, in order to keep the distinctiveness of each feature, a joint similar and specific learning (JSSL) model [9] is proposed, which divides representation coefficients into two parts to balance the similarity and distinctiveness among different features. However, the model of JSSL is a little unnecessary complex due to its double flexibility on the coding coefficients.

Fortunately, all the above issues can be solved by designing a suitable multi-feature representation model. Lately, many effective models have been proposed. For example, Luo *et al.* [11] proposed the consistent and specific multi-view subspace clustering, and Lan *et al.* [8] learned common and feature-specific patterns for multiple features. In this paper, we propose a multi-feature shared and specific representation (MFSSR) model. In the proposed model, different features have a shared representation for their commonality and different specific representations for their specificity. Moreover, a weighted representation term, which has direct connection with the final classifier, is designed to handle some features with outliers. An efficient solving algorithm is also proposed for the proposed MFSSR. Extensive experiments have been conducted to show the advantages in accuracy and running time.

The rest of this paper is organized as the following. Section 2 reviews some related works. Section 3 introduces the proposed model and its optimization. Section 4 illustrates the experimental results in several representative databases, and Sect. 5 concludes the paper.

2 Brief Review of Related Works

MTJSRC

With multiple types features for joint sparse representation and recognition, the multi-task joint sparse representation based classification (MTJSRC) has been proposed in [26]

$$\min_{\alpha_k} \sum_{k=1}^K \|\mathbf{y}_k - \mathbf{D}_k \alpha_k\|_2^2 + \lambda \sum_{j=1}^C \|\alpha_j\|_2 \quad (3)$$

where K is the number of different features, C is the number of classes, \mathbf{y}_k denotes the k^{th} feature vector of query samples, $\mathbf{D}_k = [\mathbf{D}_{k,1}, \dots, \mathbf{D}_{k,j}, \dots, \mathbf{D}_{k,C}]$ represents the k^{th} -feature dictionary, and $\alpha_j = [\alpha_{1,j}, \dots, \alpha_{k,j}, \dots, \alpha_{K,j}]$ is the coefficient associated to class j of all features, where $\alpha_{k,j}$ is associated to the k^{th} feature and the class j . It can be seen clearly that, by using a mixed-norm regularization, the representation coefficients of different features can be similar and sparse in terms of classes.

RCR

Different from the mixed-norm regularization of MTJSRC, the relaxed collaborative representation (RCR) model in [23] utilizes a weighted within-class regularization term and a l_2 -norm for representation coefficients, assuming that the coding vectors from different features have a small variance. Besides, in order to exploit the discrimination of different features, the weight for each feature can be learned in the stage of coding process. The whole formulation of RCR is as the following.

$$\min_{\alpha_k} \sum_{k=1}^K (\|\mathbf{y}_k - \mathbf{D}_k \alpha_k\|_2^2 + \lambda \|\alpha_k\|_2^2 + \tau \omega_k \|\alpha_k - \bar{\alpha}\|_2^2) \quad (4)$$

where $\bar{\alpha}$ is the mean of all α_k . τ and λ are positive scalar constants and ω_k indicates the discrimination of k^{th} features. It can be observed that the similarity between different features is exploited by reducing the variance representation coefficients.

JSSL

Although minimizing the distance between coefficients can exploit the similarity of different features, it is too restrictive since there is also distinctiveness among the coefficients. In order to keep the distinctiveness of them, [9] proposed a joint similar and specific learning (JSSL) model to address the problem. In the model of JSSL, representation coefficients are divided into two parts, namely similar part and specific part. On the one hand, similar part exploits similarity

of different features. On the other hand, specific part keeps their distinctiveness. The model of JSSL can be written as the following.

$$\min \sum_{k=1}^K \|\mathbf{y}_k - \mathbf{D}_k(\alpha_k^c + \alpha_k^s)\|_2^2 + \tau \|\alpha_k^c - \bar{\alpha}^c\|_2^2 + \sum_{k=1}^K \lambda (\|\alpha_k^c\|_1 + \|\alpha_k^s\|_1) \quad (5)$$

where α_k^c and α_k^s are similar part and specific part, respectively. It can be seen that α_k^c achieves similarity between different features by minimizing the distance of their coefficients. Besides, the specific part α_k^s extracted from α_k can keep the distinctiveness of k^{th} feature, which makes their representation more flexible.

MTJSRC assumes that multiple features have the same contributions for the representation and classification, which may not be correct in practice. For RCR, the within-class coding coefficient term has no direct connection with the final classifier, lacking a meaningful illustration on the discriminative dictionary learning model. Although promising performance has been reported by JSSL, there are still several issues. It is not necessary to introduce double flexilities, e.g., $\alpha_k^c - \bar{\alpha}^c$ and α_k^s , because the introduction of α_k^s has represent the speciality. The model of JSSL is complex due to unnecessary unknown variable, e.g., $\bar{\alpha}$. Another drawback of JSSL doesn't consider the features with outliers although the specific representation can tolerate noises to some extent.

3 Multi-feature Shared and Specific Representation

3.1 Multi-feature Shared and Specific Representation Model

Similarity and distinctiveness of multiple features are always the main problem considered in multi-feature representation based classification. Both of them are important. On the one hand, the similarity of different features means the same information they keep, which should be exploited to make the classification stable. On the other hand, different features may have extra valuable information, which may improve the recognition performance.

In order to solve the issues presented in Sect. 2, we proposed a novel multi-feature shared and specific representation (MFSSR) model

$$\min_{\alpha^c, \alpha^s, \omega} \sum_{k=1}^K (\|\mathbf{y}_k - \mathbf{D}_k(\alpha^c + \alpha_k^s)\|_2^2 + \tau \omega_k \|\mathbf{y}_k - \mathbf{D}_k \alpha^c\|_2^2 + \lambda_2 \|\alpha_k^s\|_1) + \lambda_1 \|\alpha^c\|_1 \quad (6)$$

where K is the number of different features and τ , λ_1 and λ_2 are positive scalar constants. $\mathbf{y}_k = [y_k^1; y_k^2; \dots; y_k^n] \in \mathbb{R}^n$ denotes the k^{th} feature vector of query samples. $\mathbf{D}_k = [\mathbf{d}_k^1, \mathbf{d}_k^2, \dots, \mathbf{d}_k^m] \in \mathbb{R}^{n \times m}$ represents the k^{th} -feature dictionary. $\alpha^c = [\alpha^{c,1}; \alpha^{c,2}; \dots; \alpha^{c,m}] \in \mathbb{R}^m$ is the shared coefficient vector of each feature vector \mathbf{y}_k over dictionary \mathbf{D}_k . $\alpha_k^s = [\alpha_k^{s,1}; \alpha_k^{s,2}; \dots; \alpha_k^{s,m}] \in \mathbb{R}^m$ is the specific coefficient vector of the k^{th} feature vector \mathbf{y}_k over the dictionary \mathbf{D}_k . ω_k is the weight assigned to the k^{th} feature.

Inspired by JSSL [9], in order to exploit the similarity and distinctiveness of multiple features, we also divide the representation coefficients into two parts. The primary coefficient α_k can be written as the following form

$$\alpha_k = \alpha^c + \alpha_k^s \quad (7)$$

where α^c is the shared part of all features. Different from JSSL, we required the shared coding coefficient be same for different features, i.e., α^c is the shared coding vector for all \mathbf{y}_k . The reason is that the specific representation part, i.e., α_k^s has introduced enough flexibility, and it is more effective and simpler than that of JSSL.

As the related works mentioned above, RCR introduces a weighted within-class regularization to minimize the distance of coefficients between different features under the assumption that their representation coefficients should be close to some extent. However, the regularization of RCR is too restrictive to keep enough distinctiveness of various features. Compared to RCR, which uses weighted within-class variance of coding vectors to handle bad features, the proposed MFSSR directly weights the class-specific dictionary representation, which is also the criterion of final classification. The benefit of Eq. (6) is that the training phase and testing phase are consistent. In the proposed model, we can learn the weights of different features as the following term.

$$\sum_{k=1}^K \omega_k \|\mathbf{y}_k - \mathbf{D}_k \alpha^c\|_2^2 \quad (8)$$

It is obvious that the weight ω_k should be big when \mathbf{y}_k can be well reconstructed by using only the shared coefficient, indicating the k^{th} modality is more distinctive. In order to keep our model more stable, some regularization constraints on ω_k can be adopted. For example, both Karush-Kuhn-Tucker condition and maximum entropy principle can deal with it. In this paper, we use maximum entropy principle to regularize the prior ω_k :

$$-\sum_{k=1}^K \omega_k \ln \omega_k > \delta \quad (9)$$

where δ is a positive scalar constant.

3.2 Optimization Algorithm

The objective function (6) can be minimized by alternately updating the shared coefficient α^c , the specific one α_k^s and the weight ω_k until the function converges to local minimum.

Initialization:

We initialize the proposed model by simply setting the shared coefficient and the specific coefficients as zero vectors, and the weights as one.

$$\alpha^c = \mathbf{0}, \alpha_k^s = \mathbf{0}, \omega_k = 1 \quad (10)$$

Updating the Shared Representation:

If we fix the specific coefficient α_k^s and the weight ω_k , the objective function (6) is reduced to Eq. (11).

$$\min_{\alpha^c} \sum_{k=1}^K (\|(\mathbf{y}_k - \mathbf{D}_k \alpha_k^s) - \mathbf{D}_k \alpha^c\|_2^2 + \tau \omega_k \|\mathbf{y}_k - \mathbf{D}_k \alpha^c\|_2^2) + \lambda_1 \|\alpha^c\|_1 \quad (11)$$

Obviously, we can combine all the K sub-functions because they share the same coefficient α^c . The minimization of Eq. (11) with respect to α^c can be rewritten as the following function.

$$\min_{\alpha^c} (\| \begin{bmatrix} \mathbf{y}_1 - \mathbf{D}_1 \alpha_1^s \\ \mathbf{y}_2 - \mathbf{D}_2 \alpha_2^s \\ \vdots \\ \mathbf{y}_k - \mathbf{D}_k \alpha_k^s \end{bmatrix} - \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_k \end{bmatrix} \alpha^c \|_2^2 + \tau \| \begin{bmatrix} \sqrt{\omega_1} \mathbf{y}_1 \\ \sqrt{\omega_2} \mathbf{y}_2 \\ \vdots \\ \sqrt{\omega_k} \mathbf{y}_k \end{bmatrix} - \begin{bmatrix} \sqrt{\omega_1} \mathbf{D}_1 \\ \sqrt{\omega_2} \mathbf{D}_2 \\ \vdots \\ \sqrt{\omega_k} \mathbf{D}_k \end{bmatrix} \alpha^c \|_2^2 + \lambda_1 \|\alpha^c\|_1) \quad (12)$$

Here all the left two terms in Eq. (12) are differentiable. For convenience, We rewrite Eq. (12) as the following.

$$\min_{\alpha^c} (\mathbf{F}(\alpha^c) + \lambda_1 \|\alpha^c\|_1) \quad (13)$$

where $\mathbf{F}(\alpha^c)$ represents the left two terms of the objective function (12). Since $\mathbf{F}(\alpha^c)$ is differentiable, The Iterative Projection Method (IPM) [16] can be applied to minimize Eq. (11), as described in Algorithm 1.

Updating the Specific Representation:

We fix the shared coefficient α^c and the weight ω , the specific coefficient α_k^s can be updated by reducing the objective function (6) to Eq. (14)

$$\min_{\alpha_k^s} (\|(\mathbf{y}_k - \mathbf{D}_k \alpha^c) - \mathbf{D}_k \alpha_k^s\|_2^2 + \lambda_2 \|\alpha_k^s\|_1) \quad (14)$$

It is clear that the objective function is similar to Eq. (13), which can be also optimized by The Iterative Projection Method [16].

Updating the Weight:

If the coefficient α^c and α^s are known, under the condition of maximum entropy principle (9), the objective function (6) becomes Eq. (15).

$$\min_{\omega_k} (\tau \omega_k \|(\mathbf{y}_k - \mathbf{D}_k \alpha^c)\|_2^2 + \gamma \omega_k \ln \omega_k) \quad (15)$$

The weight ω_k can be derived:

$$\omega_k = \exp(-\tau \|\mathbf{y}_k - \mathbf{D}_k \alpha^c\|_2^2 / \gamma) \quad (16)$$

In all, our optimization algorithm alternately updates the shared representation coefficient α^c , the specific coefficients α_k^s and the weights ω_k , until Eq. (6) converges. The summary of the algorithm is described in Algorithm 2. Since each sub-problem in the optimization of Eq. (6) will reduce the objective, Algorithm 2 will converge to a local optimal solution.

Algorithm 1. The coding algorithm of shared coefficient

- 1: **Input:** $\sigma, \lambda_1 > 0$
 - 2: **Initialization:** $\tilde{\alpha}^{c(1)} = \mathbf{0}$ and $h=1$.
 - 3: **while** convergence and maximal iteration number are not reached **do**
 - $h = h + 1$
 - $\tilde{\alpha}^{c(h)} = \mathbf{S}_{\lambda_1/\sigma}(\tilde{\alpha}^{c(h-1)} - \frac{1}{2\sigma}\nabla\mathbf{F}(\tilde{\alpha}^{c(h-1)}))$

where $\nabla\mathbf{F}(\tilde{\alpha}^{c(h-1)})$ is the derivative of $\mathbf{F}(\alpha^c)$ w.r.t. $\tilde{\alpha}^{c(h-1)}$, and $\mathbf{S}_{\lambda_1/\sigma}$ is a soft threshold operator defined in [16].
 - 4: **Return** $\alpha^c = \tilde{\alpha}^{c(h)}$.
-

Algorithm 2. Multi-Feature Shared and Specific Representation (MFSSR)

- 1: **Input:** $\lambda_1, \lambda_2, \tau, \mathbf{y}_k, \mathbf{D}_k, k = 1, 2, \dots, K$
 - 2: **Initialization:** $\alpha^c = \mathbf{0}, \alpha_k^s = \mathbf{0}, \omega_k = 1, k = 1, 2, \dots, K$
 - 3: **while** not converged **do**
 - update coefficients** α^c **following Eq. (12)**
 - update coefficients** α_k^s **following Eq. (14)**
 - update weights** ω_k **following Eq. (15)**
 - 4: **Return** α^c and $\alpha_k^s, k = 1, 2, \dots, K$
-

3.3 Classification

When the coding coefficients and the weights are obtained, its label is decided based on the lowest reconstruction error over all K vector:

$$identity = \arg \min_j \sum_{k=1}^K \omega_k \|\mathbf{y}_k - \mathbf{D}_{k,j}(\alpha_j^c + \alpha_{k,j}^s)\|_2^2 \quad (17)$$

where $\mathbf{D}_{k,j}$ is the elements of the dictionary \mathbf{D}_k of class j , and α_j^c and $\alpha_{k,j}^s$ are the shared and specific coefficients associated to the sub-dictionary $\mathbf{D}_{k,j}$.

4 Experiments

In this section, we conduct three face recognition (FR) experiments on two benchmark face databases, including the AR database [13] and the Labeled Faces in the Wild (LFW) database [20], to verify the effectiveness of the proposed model. In the experiments, the methods nearest neighbor (NN) [5] and SVM [7] are used as the baseline. In order to further evaluate the effectiveness of MFSSR, several multi-feature representation based classification methods, including MTJSRC [26], RCR [23] and JSSL [9], are compared with the proposed model.

For AR database, the three parameters $\lambda_1, \lambda_2, \gamma$ (the Lagrange multiplier of the entropy constraint) are set as 0.0005, 0.0005, and 0.02, respectively, which are the same for FR without occlusion and FR with disguise. For the experiments of LFW databases, a challenging task in uncontrolled environment, the parameters, such as λ_1, λ_2 and the weights ω are learned from the validation set.

This section is organized as the following. First, we give the experiment of AR with occlusion in Sect. 4.1, which evaluates MFSSR is robust to face occlusion. In Sect. 4.2, we evaluate the performance of MFSSR on LFW database with multiple features as input. Then, in Sect. 4.3, we evaluate the performance of MFSSR on experiment of AR without occlusion. In Sect. 4.4, we focus on the comparison of time complexity and running time with JSSL.

4.1 AR with Occlusion

In this subsection, we perform face recognition based on AR with occlusion. The AR database contains two-session data of 50 male and 50 female subjects. In each session, every person has 7 images with only illumination and expression variations, and 6 with real face occlusion (sunglass or scarf disguise).

In these experiments, 800 images (8 samples per person with only expression variations from two sessions) serve as the training set, while another 200 images with sunglass (or scarf) disguise are used for testing, as shown in Fig. 1(a). Following the experimental setting of RCR [23], we also resize all images to 83×64 and partitioned them into 4×2 blocks with the size of 20×30 , as shown in Fig. 1(b). Then, each block is resized to a 600-dim vector, which can be regarded as a feature vector.

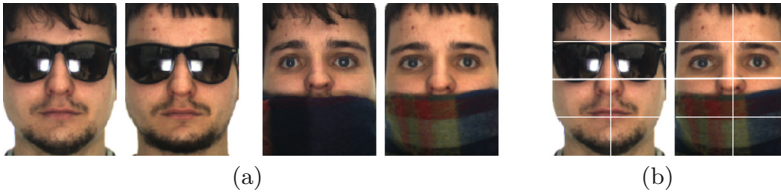


Fig. 1. (a) The testing samples with sunglasses and scarves in AR database; (b) partitioned testing samples.

The experimental results of these methods are listed in Table 1. Our proposed MFSSR achieves the best performance. SVM and NN, which are not designed for dealing with multi-features, get the worse results compared to other methods. As a multi-feature representation based classification, MTJSRC fails to deal with the problem that there is occlusion variation in some blocks, since it treats each block equally. Inversely, RCR learns weights for different blocks, and hence, the occluded parts are set very small weights, which can reduce the interference of occluded blocks in the stage of classification. JSSL, which adopts the similar within-class regularization, achieves similar results. Compared to RCR and JSSL, MFSSR not only automatically learns the weights for different blocks, but also exploits the similarity of different features by shared representation coefficient. What's more, it keep their distinctiveness by the specific part. Therefore, the proposed MFSSR achieves the best performance.

Table 1. Face recognition rates on AR database with disguise.

Method	Sunglass	Scarf
SVM	53.5%	10.5%
NN	63%	12%
MTJSRC	80.5%	90.5%
RCR	97%	94%
JSSL	96%	94%
MFSSR	100%	97.5%

4.2 LFW Face Recognition

Labeled Faces in the Wild (LFW) [20], a large-scale database of human face images, is designed for unconstrained FR in uncontrolled environment with variations of pose, illumination, expression, misalignment and occlusion, etc. (shown in Fig. 2). We use one subset of aligned LFW in our experiments. It contains 143 subjects with at least 11 samples per subject. In our experiments, the first 10 samples serve as training data and the rest as testing data. There are four representative features extracted from each sample, including intensity values, low-frequency Fourier feature [19], Gabor magnitude [10] and LBP [1]. In the feature extraction stage, similar the processing method of LDA [3], we first partition each image into 2×2 blocks, then extract the enhanced discriminative feature in each block. At last, the features of all blocks are concatenated as the final feature.

The comparison of MFSSR with other methods are presented in Table 2. With no obvious occlusion in all images, there is little difference between the recognition rates of all methods. However, SVM and NN are still worse than other multi-feature representation based classification methods, for the reason that multi-feature methods have the powerful ability to mitigate the correlation between different features, which can greatly improve the performance of classification. And among these methods, the proposed MFSSR gets the best performance, with at least 0.6% improvement over other methods. Compared to JSSL, which also divides representation coefficients into two parts, MFSSR further enforce the similarity among different features.

**Fig. 2.** (a) and (b) are samples in training and testing sets of LFW.

Table 2. Face recognition accuracy on LFW.

SVM	NN	MTJSRC	RCR	JSSL	MFSSR
68.3%	70%	77.4%	79.5%	79.3%	80.1%

4.3 AR Without Occlusion

As in RCR [23], the images with only illumination and expression variations are selected, where 700 images (7 samples per person) from Session 1 serve as the training data, while another 700 images (7 samples per person) from Session 2 for the testing data. With no obvious occlusion in these images, we simply divide them into 1×4 blocks. Then each block is resized to a vector, as the same of the experiment of AR with occlusion.

The comparison of proposed model with other competing methods is shown in Table 3. It can be observed that MFSSR has about 2% improvement compared with MTJSRC and RCR, though it is slightly worse than JSSL.

Table 3. Face recognition rates on the AR database without occlusion.

SVM	NN	MTJSRC	RCR	JSSL	MFSSR
87.1%	74.7%	95.8%	95.9%	97.8 %	97.7%

4.4 Time Complexity and Running Time

We verify the efficiency of MFSSR by comparing with JSSL in time complexity and running time.

Suppose that the size of all \mathbf{D}_k are $n \times m$, and the number of testing data is t . All the testing vectors are organized as a matrix. For JSSL updating a_k^c once by Augment Lagrangian Method (ALM), the time complexity of coding is $\mathcal{O}((3K+c)m^3 + Km^2n + 3Kmnt)$. First, Computing Q has complexity $\mathcal{O}(cm^3)$, where c is a positive constant. Second, Computing P_k has complexity $2Km^3$. Third, the time complexity P_kQ is Km^3 . Fourth, the computation complexity of $a_{0,k}^c$ is $\mathcal{O}(K(m^2n + 2mnt))$. At last, soft threshold operation needs $\mathcal{O}(Kmnt)$. Besides, when updating a_k^s one iteration, the time complexity $\mathcal{O}(3Kmnt)$. In all, the time complexity of JSSL is $\mathcal{O}(q((3K+c)m^3 + Km^2n + 6Kmnt))$, where q is the iteration number.

However, in MFSSR, both a^c and a_k^s are updated by IPM, which is very timesaving. First, updating a^c has complexity of $\mathcal{O}(6Kmnt)$. Then, the time complexity of updating a_k^s is $\mathcal{O}(3Kmnt)$ like JSSL. SO, the coding complexity of MFSSR is $\mathcal{O}(q(9Kmnt))$, where q is the iteration number. It is obvious in the proposed MFSSR is more effective than JSSL.

We conduct the running time experiments by the desktop of 3.5 GHz CPU with a 8 GB RAM. As is seen in Table 4, the proposed MFSSR is more efficient than JSSL in all experiments. For instance, MFSSR is two times faster than JSSL on the LFW database.

Table 4. Average computational time (seconds) coding and classifying one testing sample.

Experiment	AR(sunglass)	AR(scarf)	LFW	AR block
JSSL	0.91	0.92	3.96	0.71
MFSSR	0.63	0.78	1.87	0.45

5 Conclusion

In this paper, we propose a multi-feature shared and specific representation model (MFSSR) for pattern recognition, which further exploit the similarity and distinctiveness of different features for coding and classification. By dividing the coefficients into the shared part and the specific part, the discrimination embedded in multiple features is enhanced through the shared representation, while the distinctiveness of different features is tolerated by the specific representation. An adaptively weighted representation term is also proposed, with excellent performance to image recognition with occlusions. The experimental results on several representative databases demonstrated the advantages of our proposed model in accuracy and efficiency.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China (Grant no. 61772568), the Guangzhou Science and Technology Program (Grant no. 201804010288), the Fundamental 535 Research Funds for the Central Universities (Grant no. 18lgzd15), the Shenzhen Scientific Research and Development Funding Program (Grant no. JCYJ20170302153827712).

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24670-1_36
2. Ross, A.A., Govindarajan, R.: Feature level fusion of hand and face biometrics (2005)
3. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 711–720 (1997)
4. Caruana, R.: Multitask learning. *Mach. Learn.* **28**(1), 41–75 (1997)
5. Cover, T., Hart, P.: Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13**(1), 21–27 (1967)

6. Hall, D.L., Llinas, J.: An introduction to multisensor data fusion. *Proceed. IEEE* **85**(1), 6–23 (1997)
7. Heisele, B., Ho, P., Poggio, T.: Face recognition with support vector machines: global versus component-based approach. In: *Proceedings Eighth IEEE International Conference on Computer Vision, ICCV 2001*, vol. 2, pp. 688–694 (2001)
8. Lan, X., Zhang, S., Yuen, P.C., Chellappa, R.: Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. *IEEE Trans. Image Process.* **27**(4), 2022–2037 (2018)
9. Li, J., Zhang, D., Li, Y., Wu, J., Zhang, B.: Joint similar and specific learning for diabetes mellitus and impaired glucose regulation detection. *Inform. Sci.* **384**, 191–204 (2017)
10. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Process.* **11**(4), 467–476 (2002)
11. Luo, S., Zhang, C., Zhang, W., Cao, X.: Consistent and specific multi-view subspace clustering (2018)
12. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: *2009 IEEE 12th International Conference on Computer Vision*, pp. 2272–2279, September 2009
13. Martínez, A., Benavente, R.: The AR face database. *CVC Technical report 24* (1998)
14. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**(6583), 607–9 (1996)
15. Ozawa, S., Roy, A., Roussinov, D.: A multitask learning model for online pattern recognition. *IEEE Trans. Neural Netw.* **20**(3), 430 (2009)
16. Rosasco, L., Verri, A., Santoro, M., Mosci, S., Villa, S.: Iterative projection methods for structured sparsity regularization. *Computation* (2009)
17. Ruta, D., Gabrys, B.: An overview of classifier fusion methods. *Comput. Inform. Syst.* **7**, 1–10 (2000)
18. Schölkopf, B., Platt, J., Hofmann, T.: Sparse representation for signal classification. In: *Proceedings of the Twentieth Conference on Neural Information Processing Systems Advances in Neural Information Processing Systems 19*, Vancouver, British Columbia, Canada, December, pp. 609–616 (2006)
19. Su, Y., Shan, S., Chen, X., Gao, W.: Hierarchical ensemble of global and local classifiers for face recognition. *IEEE Trans. Image Process.* **18**(8), 1885–1896 (2009)
20. Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: Zha, H., Taniguchi, R., Maybank, S. (eds.) *ACCV 2009*. LNCS, vol. 5995, pp. 88–97. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12304-7_9
21. Wright, J., Ma, Y.: Dense error correction via ell^1 -minimization. *IEEE Trans. Inform. Theory* **56**(7), 3540–3560 (2010)
22. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(2), 210–227 (2009)
23. Yang, M., Zhang, L., Zhang, D., Wang, S.: Relaxed collaborative representation for pattern classification. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2224–2231, June 2012
24. Yang, M., Zhang, L.: Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010*. LNCS, vol. 6316, pp. 448–461. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15567-3_33

25. Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In: *Computer Vision and Pattern Recognition*, pp. 625–632 (2011)
26. Yuan, X.T., Liu, X., Yan, S.: Visual classification with multitask joint sparse representation. *IEEE Trans. Image Process.* **21**(10), 4349–4360 (2012)



Distillation of Random Projection Filter Bank for Time Series Classification

Yufei Lin¹, Sen Li¹, and Qianli Ma^{1,2}(✉)

¹ School of Computer Science and Engineering,
South China University of Technology, Guangzhou, China
yufeilincs@foxmail.com, awslee@foxmail.com

² Guangdong Key Laboratory of Big Data Analysis and Processing,
Guangzhou, China
qianlima@scut.edu.cn

Abstract. Time series is widely found in various fields such as geo-science, medicine, finance, and social sciences. How to effectively extract the features of time series remains a challenge due to its potentially complex non-linear dynamics. Recently, Random Projection Filter Bank (RPFb) [5] is proposed as a generic and simple approach to extract features from time series data. It generates the features by randomly generating numerous autoregressive filters that are convolved with input time series. Such numerous random filters inevitably have redundancy and lead to the increased computational cost of the classifier. In this paper, we propose a distillation method of RPFb, named D-RPFb, to not only maintain the high level of quantity of the filters, but also reduce the redundancy of the filters while improving precision. We demonstrate the efficacy of the features extracted by D-RPFb via extensive experimental evaluation in three different areas of time series data with three traditional classifiers (i.e., Logistic Regression (LR) [2], Support Vector Machine (SVM) [14] and Random Forest (RF) [8]).

Keywords: Random projection · Filter bank · Time series
Feature extraction

1 Introduction

Time series data are ubiquitous in many practical applications ranging from health care [3], action recognition [10], financial markets [15] to urban traffic control [16]. How to extract the features of time series effectively is a popular research topic [4, 5, 7, 9, 13]. However, time series extraction remains a challenging task due to the potentially complex non-linear dynamic system behind the time series.

Recently, Random Projection Filter Bank (RPFb) [5] is proposed as a generic and simple approach to extract features from time series data. RPFb is a set of randomly generated stable autoregressive filters that are convolved with the input time series to generate the features. These features can be used by any

conventional machine learning algorithm for solving tasks such as time series prediction, classification with time series data, etc. Different filters in RPFb extract different aspects of the time series, and together they provide a reasonably good summary of the time series.

However, numerous random filters inevitably have redundancy and lead to the increased computational cost of classifier. Moreover, in some cases, redundant features will make the performance of classifier worse. How to reduce redundant features (i.e., estimate the quality of the filter) is an important issue. In this paper, with an aim of reducing the number of redundant filters, we propose a way to distil the filters of RPFb, named D-RPFb, which uses a set of specific rules to filter the filters that are most capable of guiding the classifier to get better performance. D-RPFb can reduce the number of redundant and even potentially mislead filters, thus improving the quality of the features provided to the classifier which directly improves the learning ability of the classifier and obtains a better performance.

2 Preliminaries

There is a crucial process for the distillation of RPFb, which is designed to measure the quality of a specific filter. To do that, we introduce entropy [6]. Considering that entropy is not very common in time series analysis, we first introduce the concept of entropy briefly before proposing our D-RPFb formally.

Entropy [6] is often used in information theory and probability statistics to measure the uncertainty of a variable. Entropy is always a real number larger than 0 but smaller than 1. Its value indicates the degree of uncertainty of random variables. When the entropy is equal to 0, the random variable is completely certain without any randomness. When entropy is equal to 1, the uncertainty of the random variable peaks. This property of entropy makes it possible to use the entropy to measure the classification quality of the classification subset when a classifier uses a single feature extracted by certain filter to classify an instance. The smaller the entropy of a subset, the more the feature extracted by the filter can make the classifier better complete the clustering, and vice versa, the greater the entropy value indicates that the feature extracted by the filter may lead to the confusion of the classification results.

3 Proposed Methods D-RPFb

3.1 Brief Review of Random Projection Filter Bank

The idea behind RPFb is to randomly generate many simple dynamical systems (i.e., $\frac{1}{1-Z'_n z^{-1}}$ denotes a certain simple dynamical system with a given pole Z'_n and z^{-1} denotes the inverse of z-transform [11]) that can approximate optimal dynamical systems with a high accuracy.

In order to do this, what we should do first is to determine the number of filters in the filter bank. After that, given the certain number of filters N , we draw

N random real numbers or the imaginary numbers Z'_1, \dots, Z'_n from the unit circle to construct a filter bank defined by filter $\phi(z^{-1}) = (\frac{1}{1-Z'_1 z^{-1}}, \dots, \frac{1}{1-Z'_n z^{-1}})$ which contains N random projection filters. Then, we pass each input time series through every filter in RPFb to do convolution and generate N features corresponding to each time series at each time step. For example, assuming the length of the each input time series is T , we will get $N * T$ features after passing it through RPFb. Finally, we can input the obtained features into different classifiers for conducting time series classification.

3.2 The Distillation of Random Projection Filter Bank

Introduce the Entropy into Time Series. The entropy is used in the traditional decision tree ID3 algorithm [12] for feature selection. That motivates us to use entropy to evaluate the quality of a certain filter. However, in the traditional decision tree ID3 algorithm [12], the entropy is only applicable to a discrete variable. To solve this issue, we use an extra classifier to introduce the entropy into time series and achieve the purpose of evaluating the quality of a certain filter. In general, assuming the length of the each input time series is T , we will get T features through time after passing it through a certain filter. Then, we input the T features into a certain classifier to get the classification result. In this way, for each time series example, we get a classification result which makes a certain filter become a discrete variable. And, we propose evaluation method combined with entropy and classification result to evaluate the quality of a certain filter.

Computation of Subset Uncertainty and Evaluation of Filters. After using RPFb to generate filter, each filter will be executed with the proposed evaluation algorithms to get their evaluation value. The overall algorithm flow is shown in Algorithm 1. First, in the training data set, randomly select the same number of instances in each category to form data set D_m for avoiding unbalanced sample. For each filter in RPFb, randomly select the half number of instances in D_m as training data D_t , the other half as validation data D_v and then pass the train and valid data into the filter, extracting the corresponding features (denoted by F_t and F_v). Then, fitting the classifier with the F_t . When the remaining features F_v are classified by the classifier, each category (totally M category) will produce a corresponding subset D'_m . Each subset D'_m may contain the instances that belong to the subset or contains instances that do not belong to the subset. Thirdly, we can calculate the uncertainty of each subsets D'_m by entropy. If the uncertainty of the subset D'_m is small it means that D'_m contains many instances of the same category, which means that the feature extracted by the filter can guide the classifier to complete the clustering of the time series. However, only clustering results cannot evaluate whether a filter is really efficient because if a subset D'_m contains many instances of the same category that do not belong to D'_m , the feature extracted by the filter is quite bad which misleads the classifier. Therefore, we have to consider the classification accuracy as the second characteristics of each subset D'_m . In this way, the two important measurements,

the clustering effect and the classification accuracy are both considered. Both of them are equally important for evaluating the quality of the feature extracted by a filter. Therefore, D-RPFB proposes a method for calculating the evaluation value of a certain filter as follow:

Algorithm 1. The distillation of random projection filter bank

Input: Dataset = $(X_{i,1}, Y_{i,1}), \dots, (X_{i,T_i}, Y_{i,T_i})_{i=1}^m$

Output: Classifier \hat{f} and new filter bank ϕ_{new}

- 1 $l : Y \times Y \rightarrow \mathbb{R}$: Loss function;
 - 2 \mathcal{F} : Function space;
 - 3 n : The number of filters in random projection filter banks;
 - 4 ρ : The percentage of remaining filters after the screening filter;
 - 5 Draw Z'_1, \dots, Z'_n uniformly random within the unit circle.
 - 6 Define filter $\phi(z^{-1}) = (\frac{1}{1-Z'_1 z^{-1}}, \dots, \frac{1}{1-Z'_n z^{-1}})$.
 - 7 In the training data set, randomly select the same number of instances in each category to form data set D_m .
 - 8 **foreach** $\phi_i(z^{-1})$ **in** $\phi(z^{-1})$ **do**
 - 9 Pass each time series in D_m through filter $\phi_i(z^{-1})$.
 - 10 Randomly select the half number of instances in D_m as training data D_t , the other half as valid data D_v .
 - 11 Input the corresponding features F_t generated by training data D_t in step 9 into the classifier to fit the model. (The type of classifier used here is the same as the f in line 19.)
 - 12 Input the corresponding features F_v generated by valid data D_v into the fitted model got by step 11 to get the classification subsets D'_m .
 - 13 Use Equation (3) to calculate the evaluation $E_{\phi(z^{-1})}$ of the filter $\phi_i(z^{-1})$.
 - 14 **end**
 - 15 Sort all filters according to their evaluation value $E_{\phi(z^{-1})}$.
 - 16 Select the corresponding number of filters based on ρ with higher evaluation values to form new filter banks ϕ_{new} .
 - 17 Pass each time series in training set through every filter in the new filter bank ϕ_{new} .
 - 18 Use the new extracted features $(X'_{i,1:T_i})$ generated by new filter bank ϕ_{new} to construct the estimator, we use regularized empirical risk minimization to solve it and $J(f)$ controls the complexity of the function space:
 - 19 $\hat{f} \leftarrow \arg \min_{f \in \mathcal{F}} \sum_{i=1}^m \sum_{t=1}^{T_i} l(f(X'_{i,t}, Y_{i,t})) + \lambda J(f)$
 where l denotes the cross entropy cost function, J can be lasso or ridge regression regularization.
 - 20 Return \hat{f} and ϕ_{new}
-

$$H(D'_m) = - \sum_{m=1}^M p_m \log p_m \quad (1)$$

$$Recall_{D'_m} = \frac{TP}{TP + FN} \quad (2)$$

$$E_{\phi(z^{-1})} = \sum_{m=1}^M (1 - H(D'_m)) \times (Recall_{D'_m}) \quad (3)$$

where $H(D'_m)$ is the entropy of a classification subset of the filter i , M is the total number of category, p_m is proportion of an instance of M category in the classification subset D'_m , TP is the number of the samples classified correctly in this category, $TP + FN$ is the number of the total samples in this category, $Recall_{D'_m}$ is the recall of classification subsets D'_m and $E_{\phi(z^{-1})}$ is the total evaluation value E of the i filter.

4 Experiment

In order to verify that the proposed D-RPFB can reduce the redundancy of the numerous filters while also keeping or even improving the performance of classification, we evaluate it in three different areas of time series data with three traditional classifiers (i.e., LR, SVM and RF) compared with RPFB. First, we investigate the effect of the proposed evaluation method for measuring the quality of a specific filter. Then, we show the experimental results on other two time series. Finally, we give an analysis of the screening percentage of the filters to empirically decide how many filters should be retained.

4.1 Analyzing the Effect of the Proposed Evaluation on Star Curve Data Set

The proposed evaluation method in the Eq. (3) for measuring the quality of a certain filter plays an important role in our D-RPFB. We first investigate the effect of the proposed evaluation method on the Star curve data set [1]. We assess the effect of the Eq. (3) by answering the question: Can we use the Eq. (3) to get three group filter banks that correspond to an excellent, inferior, and average property and get the corresponding performance on the test set? If this happens, then the proposed evaluation method is considered to be effective.

Our experimental scheme is as follows. Firstly, a sufficient number of filters are generated to form an initial filter group. Then, we input a part of the training data and the initial filter bank into the filter method to get the evaluation of all the filters by Eq. (3). Third, sorting the filter by the respective evaluation value of E , we divide the filter into four intervals according to the evaluation value of E (i.e., $0 < E < 0.25$ for worst, $0.25 < E < 0.5$ for worse, $0.5 < E < 0.75$ for better, $0.75 < E < 1$ for best). Finally, we construct three group filter banks with 200 filters in each that corresponds to excellent, inferior, and average distribution by randomly selecting a specific number of filters in a specific interval to meet the scheme we need. The corresponding distribution is shown in Fig. 1.

Figure 2 shows clearly the ability of the evaluation method to distinguish high quality filters from inferior filters. Generally speaking, the classification error of the inferior distribution is far higher than the classification error rate of the average distribution and the classification errors of the excellent distribution

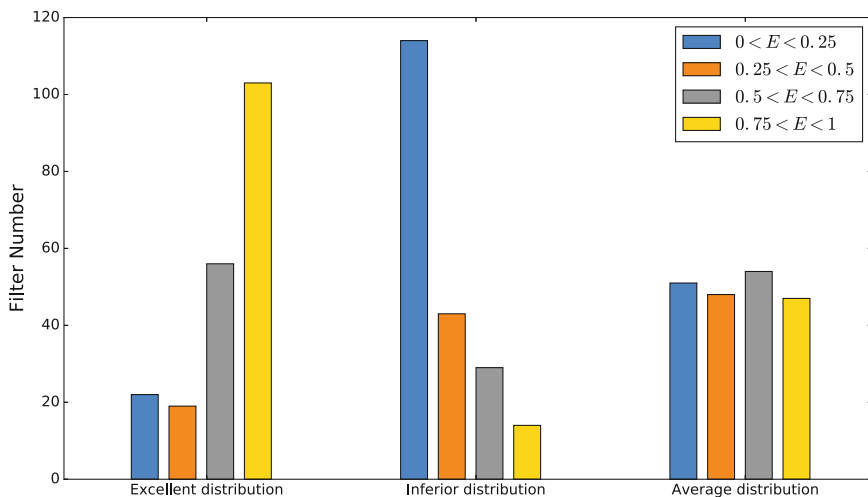


Fig. 1. The number of filters with different evaluation values in the three group of filter banks.

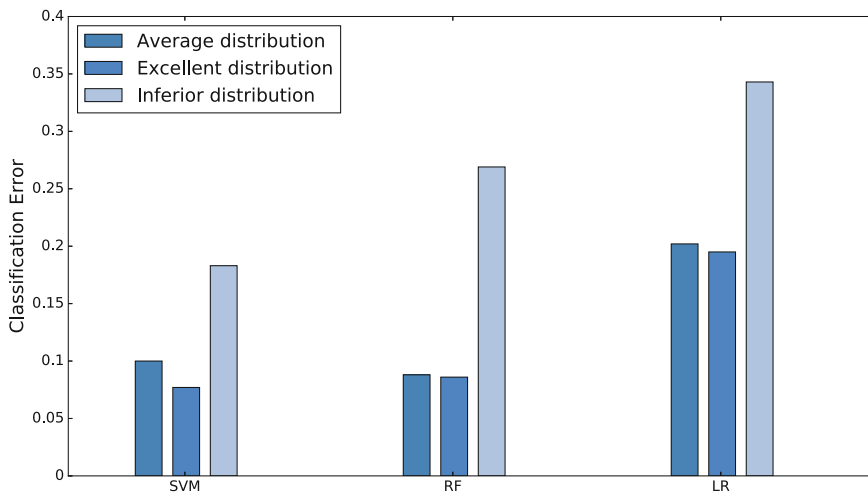


Fig. 2. The performance of classification comparison among three filter banks distribution with three classifiers.

are lower than the average distribution on the three classifiers, which shows that the proposed evaluation method can effectively distinguish high quality and low quality filters.

4.2 Detection of Bearing Defects

To compare D-RPFB and RPFB, we employ the bearing defect detection data set [5] used by the RPFB. We extract 40 time series of length 3333 in each class time series for filtering screening and testing. First, we select 15 time series (3 categories in total 45) in each category to screen the filter. Next, we generate a set of filter banks, each of which will be used in the D-RPFB and RPFB respectively. In RPFB, the filter group will maintain the number of the filters and participate in the classification of time series, and finally produce the classification error rate. In D-RPFB, the filter group will be firstly screened and then participate in the classification of time series. In this case, if the classification error rate of the D-RPFB is the same with that of the RPFB, it can verify that D-RPFB can reduce the number of redundancy and even potentially mislead filters, thus obtaining a better performance.

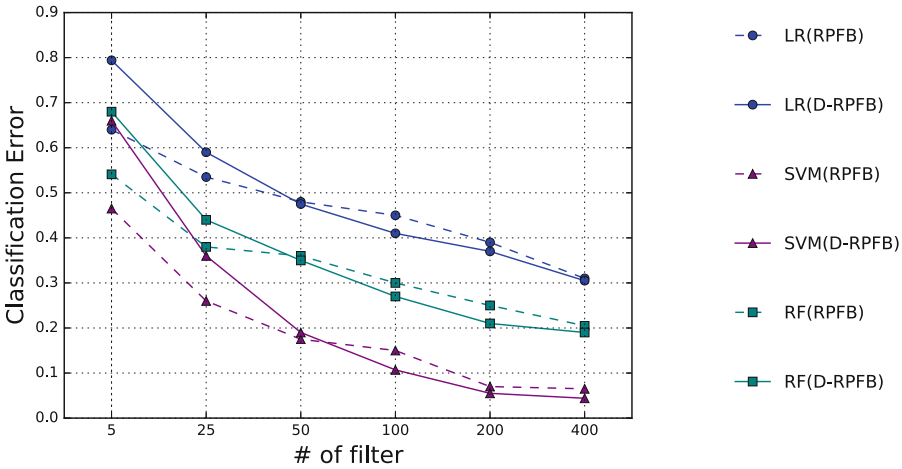


Fig. 3. The performance of classification comparison between the RPFB and D-RPFB with different classifiers on data set detection of bearing defects [5].

In our experiment, we empirically retain 75% filters (i.e., reduced number of filters in RPFB by 25%) in D-RPFB. As shown in Fig. 3, both of D-RPFB and RPFB are decreasing with the increasing of the number of filters. On this data set, the SVM can provide a lower error rate than the LR or RF. This conclusion is consistent in both the D-RPFB and RPFB. On the one hand, the error rate of the RPFB and D-RPFB is relatively high when the number of filters is relatively small. Besides, D-RPFB is worse than RPFB. This implies that the RPFB has a limited ability to summarize the time series when there are only a few filters. Meanwhile, D-RPFB further reduces the number of filters with relatively poor quality by distillation mechanism results in fewer filters, which reduces the accuracy of the D-RPFB. On the other hand, with the increasing of

the number of filters, the error rate of the D-RPFB and RPFB has decreased, but the D-RPFB declines more. This is because the D-RPFB has gradually obtained the filter which can accurately summarize the time series through the screening mechanism and remove some filters that can produce a misleading effect. The RPFB, because there is no screening mechanism to distinguish the redundant and misleading filters, the effect of some inefficient filters hinders classifier from getting a better performance.

4.3 Heart Rate Classification

To show more that the D-RPFB can improve the performance of classification, we apply the heart rate data set [5] used in the RPFB. There are two time series with a length of 1800, which belong to category A and B respectively. We firstly divide the time series of category A into 30 short time series with 60 length, 15 of which are training data sets and 15 others are test data sets. Next, we conduct the same operations on the time series of category B. After dividing two long time series, we get 30 training time series (15 of them are category A and the remaining 15 are category B) and 30 test time series (also 15 of them are category A and the remaining 15 are category B). Then, we generate a set of filter banks, each of which will be used in the D-RPFB and RPFB respectively. Finally, again, RPFB uses all the generated filters for classifier. And D-RPFB uses the screened filters for classifier.

In this experiment, we empirically retain 75% filters (i.e., reduced number of filters in RPFB by 25%) in D-RPFB. As shown in Fig. 4, with the small amounts of filters, the performance of D-RPFB is inferior to RPFB again. This implies that there is no need for distillation when the number of filter is very small. However, with the increase of the filters, most of the points on the classification

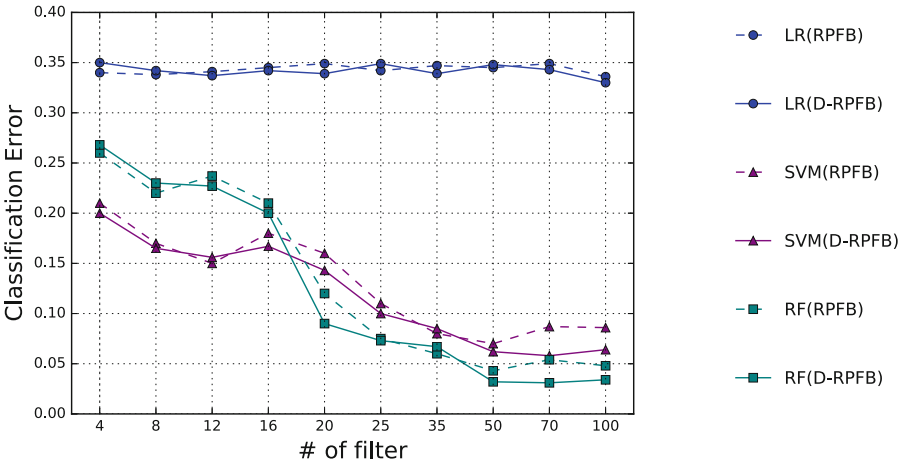


Fig. 4. The performance of classification comparison between the RPFB and D-RPFB with different classifiers on data set heart rate [5].

error curve using the features provided by D-RPFB are under the classification error curve of using the features provided by RPFB, even if some points are not under the classification error curve of RPFB, they are not much higher than in the original method. That is to say, such numerous filters randomly generated by RPFB are indeed redundant and have some misleading filters. D-RPFB distil the filters obtained by RPFB to reduce redundancy or some misleading filters to achieve the high quality of the filters and then input to the classifier, resulting a better performance.

4.4 Analyzing the Choosing of the Screening Percentage of the Filters on Hand Profile Data Set

How many filters can be kept to obtain a good summary of the input time series remains to be a question. The above reported result is under the 75% retainment (i.e., the corresponding percentage of screening is 25%) of the filters case. In this section, we analyze the choosing of the screening percentage of the filters on Hand profile data set [1]. We first generate 200 filters and then adjust the remaining filter ratio by selecting the high ranking filters, obtaining the corresponding results.

As shown in Fig. 5, if the number of filters retained is too small, the features extracted by these filters may not provide a good summary of the input time series, thus resulting a worse performance. With the percentage of retainment is increasing, the performance is better. Combined with the conclusions of experiments 4.2 and 4.3, there is no redundant or misleading information which could

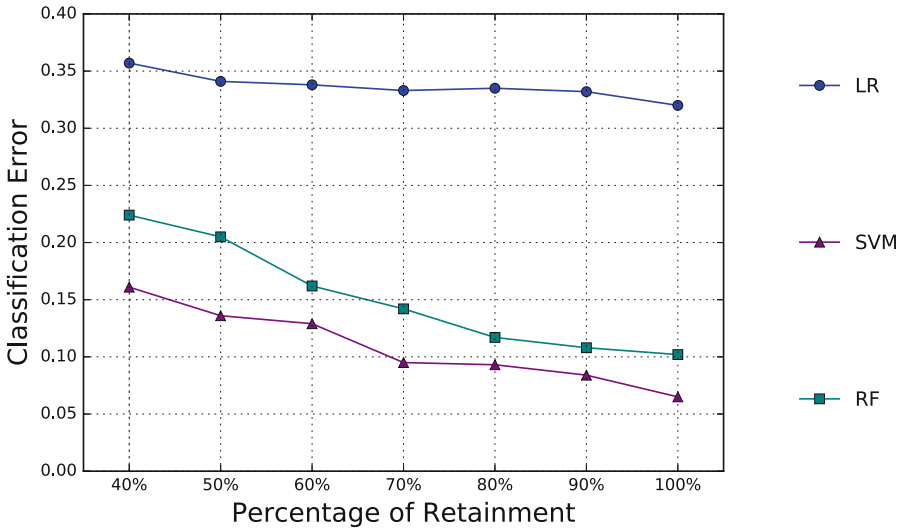


Fig. 5. The performance of classification obtained by using different classifiers under different percentages of retainment on D-RPFB.

harm the performance among such 200 filters. We can see that the original time series has been well summarized at the 80% of retainment (i.e., the corresponding percentage of screening is 20%), because the benefits from retaining more are already very small. Besides, more filters retained mean more running-time consuming when combined with specific classifier. So, in our experiment, we retain the number of filters at the original 80% while making further adjustments and finally retain 75% (i.e., the corresponding percentage of screening is 25%) to get a better performance.

5 Conclusion

In this paper, we proposed the distillation of random projection filter bank (D-PRFB) for time series classification, which is an improvement method of the random projection filter bank (PRFB). Before directly applying the features generated by the randomly generated numerous autoregressive filters that are convolved with the input time series, we add filter screening in the original method for screening the filters that are most capable of guiding the classifier to get better performance. We evaluated the D-PRFB in three different areas of time series data with three traditional classifiers. Extensive experimental results demonstrate that D-PRFB can reduce redundancy and even potentially misleading filters, thus improving the quality of the features provided to the classifier which directly improves the learning ability of the classifier to obtain a better performance.

Acknowledgment. The work described in this paper was partially funded by the National Natural Science Foundation of China (Grant No. 61502174, 61872148), the Natural Science Foundation of Guangdong Province (Grant No. 2017A030313355, 2017A030313358), the Guangzhou Science and Technology Planning Project (Grant No. 201704030051), the Opening Project of Guangdong Province Key Laboratory of Big Data Analysis and Processing (Grant No. 2017014) and the Guangdong University of Finance & Economics Big Data and Educational Statistics Application Laboratory (Grant No. 2017WSYS001).

References

1. Chen, Y., et al.: The UCR time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/
2. Cucchiaro, A.: Applied logistic regression. *Technometrics* **44**(1), 81–82 (1989)
3. Elmoaqt, H., Tilbury, D.M., Ramachandran, S.K.: Multi-step ahead predictions for critical levels in physiological time series. *IEEE Trans. Cybern.* **46**(7), 1704–1714 (2016)
4. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in time-series databases, vol. 23 (1994)
5. Farahmand, A.M., Pourazarm, S., Nikovski, D.: Random projection filter bank for time series data. In: *Advances in Neural Information Processing Systems*, pp. 6565–6575 (2017)

6. Jaynes, E.T.: Information theory and statistical mechanics. *Phys. Rev.* **106**(4), 620 (1957)
7. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. *Knowl. Inform. Syst.* **3**(3), 263–286 (2001)
8. Liaw, A., Wiener, M.: Classification and regression by randomForest. *R news* **2**(3), 18–22 (2002)
9. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 2–11 (2003)
10. Ma, Q., Shen, L., Chen, E., Tian, S., Wang, J., Cottrell, G.W.: Walking walking walking: action recognition from action echoes. In: *International Joint Conference on Artificial Intelligence*, pp. 2457–2463 (2017)
11. Oppenheim, A.V., Schaffer, R.W.: *Discrete-time signal processing* **23**(2), 157 (1989)
12. Quinlan, J.R.: Induction of decision trees. *Mach. Learn.* **1**(1), 81–106 (1986)
13. Susto, G.A., Schirru, A., Pampuri, S., Mcloone, S.: Supervised aggregative feature extraction for big data time series regression. *IEEE Trans. Ind. Inform.* **12**(3), 1243–1252 (2016)
14. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. *Neural Process. Lett.* **9**(3), 293–300 (1999)
15. Xu, Z., Kersting, K., von Ritter, L.: Stochastic online anomaly analysis for streaming time series. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3189–3195 (2017)
16. Zhan, H., Gomes, G., Li, X.S., Madduri, K., Sim, A., Wu, K.: Consensus ensemble system for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 1–12 (2018)



Jointly Sparse Reconstructed Regression Learning

Dongmei Mo¹, Zhihui Lai¹(✉), and Heng Kong²

¹ College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

lai_zhi_hui@163.com

² School of Medicine, Shenzhen University, Shenzhen 518060, China

Abstract. Least squares regression and ridge regression are simple and effective methods for feature selection and classification and many methods based on them are proposed. However, most of these methods have small-class problem, which means that the number of the projection learned by these methods is limited by the number of class. In this paper, we propose a jointly sparse reconstructed regression (JSRR) to solve this problem. Moreover, JSRR uses $L_{2,1}$ -norm as the basic measurement so that it can enhance robustness to outliers and guarantee joint sparsity for discriminant feature selection. In addition, by integrating the property of robust feature selection (RFS) and principle component analysis (PCA), JSRR is able to obtain the projections that have minimum reconstructed error and strong discriminability for recognition task. We also propose an iterative algorithm to solve the optimization problem. A series of experiments are conducted to evaluate the performance of JSRR. Experimental results indicate that JSRR outperforms the classical RR and some state-of-the-art regression methods.

Keywords: Regression · Feature selection · Joint sparsity · Classification Robustness

1 Introduction

During the last decades, many methods are proposed for feature selection. Taking the label information into consideration or not, the feature selection methods can be divided into three categories: supervised algorithms, semi-supervised algorithms and unsupervised algorithms. For unsupervised learning, the classical method is principle component analysis (PCA) [1] which projects high dimensional data into a lower dimensional space via seeking the maximum the variance of the data [2]. For supervised learning, linear discriminant analysis (LDA) [3] is the representative method that utilizes label information to learn an optimal matrix that maximizes the between-class scatter and at the same time minimizes the within-class scatter in feature space [4]. Besides, least squares regression (LSR) and ridge regression (RR) are also the classical supervised learning methods.

Although PCA, LDA and LSR are simple and effective in dealing with problems in data analysis and machine learning, they still have a major disadvantage. That is, they

do not have sparsity property. Actually, the methods with sparsity are able to learn a series of sparse projections for feature presentation. To solve this problem, the sparse RR [5] as well as elastic net [6] was proposed. These methods are very classical and widely used in many cases. Inspired by them, many regression based methods are also developed to learn sparse approximation projections for feature selection [7–10]. However, these sparse learning methods based on L_1 -norm regularization have two drawbacks. First, the L_1 -norm based methods do not have joint sparsity. Second, since these methods use L_1 -norm regularization on the projections, they need to compute the projection vectors one by one during the procedure of feature selection, which leads to higher training time. Recently, $L_{2,1}$ -norm regularization has attracted great attention in the field of feature selection, with which we can obtain joint sparsity to improve the performance of feature selection and classification. Moreover, the $L_{2,1}$ -norm based methods are less time-consuming than the methods based on L_1 -norm regularization. Nie et al. proposed robust feature selection (RFS) [11] by using $L_{2,1}$ -norm on both of loss function and the regularization term. Yang et al. proposed unsupervised discriminative feature selection (UDFS) [12] to extend the $L_{2,1}$ -norm regularization to unsupervised learning. Xiang et al. proposed discriminative least squares regression (DLSR) [13] to enlarge the distance between different classes based on the framework of LSR. In addition, many $L_{2,1}$ -norm based methods are also proposed to deal with different classification tasks [14–19].

Even though the above $L_{2,1}$ -norm based methods are able to obtain jointly sparse projections for discriminative feature selection, they ignore the small-class problem. That is, the number of the learned projections is limited by the number of class. For example, suppose the number of the class is c , RR, RFS and even DSLR cannot obtain more than c projections for feature selection, which indicates that they cannot obtain enough projections if the number of the class is small. In addition, all of the existing methods do not consider the property of supervised and unsupervised learning in a unified regression form.

Based on this regard, in this paper we propose a reconstructed regression method for jointly sparse feature selection. The proposed method called Jointly Sparse Reconstructed Regression (JSRR) integrates the property of RFS and PCA in regression form, by which the joint sparsity is obtained and the small-class problem is solved. Moreover, compared with PCA, JSRR is able to embed the label information in the loss function so as to obtain discriminative projection for feature selection. In summary, the contributions of the proposed JSRR can be described as follows:

- (1) JSRR is able to enhance the robustness to outliers by using $L_{2,1}$ -norm instead of the L_2 -norm as the basic measurement on the loss function. Moreover, it can guarantee the joint sparsity for discriminative feature selection by imposing $L_{2,1}$ -norm penalty on the regularization term.
- (2) Compared with LSR, RR and their extensions, JSRR can solve the small-class problem, by which it can obtain more than c projections to improve the performance of feature selection and classification.
- (3) Compared with PCA, JSRR considers the label information on the loss function, so that it can obtain more discriminative information for effective feature selection.

2 The Proposed Method

In this section, we first present the notation of the variables in this paper and briefly review the classical ridge regression. Then we propose the jointly sparse reconstructed regression (JSRR) for feature selection and give the corresponding optimization procedure.

2.1 Notation

In this paper, we denote all the matrices as bold uppercase italic letters, i.e. \mathbf{X} , \mathbf{Y} , etc., while vectors are denoted as bold lowercase italic letters, i.e. \mathbf{x} , \mathbf{y} , etc. and scalars are presented as lowercase italic letters, i.e. i , j , c , n .

The sample matrix is denoted as $\mathbf{X} \in R^{d \times n}$, where d is the dimension of the data and n is the number of samples. The label matrix is presented as $\mathbf{Y} \in R^{n \times c}$ with $Y_{ij} = 1$ where \mathbf{x}_i belongs to j -th class, otherwise, $Y_{ij} = 0$.

2.2 Ridge Regression Revisit

Least squares regression is simple and effective technic for data analysis and classification. The optimization problem of LSR is as follows

$$\mathbf{W}^\# = \arg \min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}^T \mathbf{W}\|_F^2 \quad (1)$$

where $\mathbf{Y} \in R^{n \times c}$ is the label matrix and $\mathbf{X} \in R^{d \times n}$ is the sample matrix, $\mathbf{W} \in R^{d \times c}$ is the projection matrix used for feature selection.

LSR can obtain its optimal solution only when $\mathbf{X}\mathbf{X}^T$ is full-rank, that is, it exists the singular problem. To solve this problem, a L_2 -norm based penalty is added to the objective function and that comes to the optimization problem of the ridge regression.

$$\mathbf{W}^* = \arg \min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}^T \mathbf{W}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \quad (2)$$

where λ is the parameter to balance the two terms. The second term in (2) acts as bias term and it can avoid the singular problem in LSR.

2.3 Jointly Sparse Reconstructed Regression

Motivated by the previous work in RFS [11] that using $L_{2,1}$ -norm on both of the loss function and regularization term can not only enhance the robustness to outliers but also obtain jointly sparse projections for feature selection. In this paper, we integrate the property of RFS and PCA to design a more complete model to not only inherit the property of RFS and PCA, but also solve the small-class problem in RR or its extensions so as to obtain enough projections to improve the performance of feature selection and classification. The objective function of the proposed JSRR is presented as follows

$$\begin{aligned}
 (\mathbf{Q}^*, \mathbf{P}^*, \mathbf{A}^*) = \arg \min_{\mathbf{Q}, \mathbf{P}, \mathbf{A}} & \alpha \|\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A}\|_{2,1} + (1 - \alpha) \|\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X}\|_{2,1} + \beta \|\mathbf{Q}\|_{2,1} \\
 \text{s.t. } & \mathbf{A} \mathbf{A}^T = \mathbf{I}
 \end{aligned} \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{d \times n}$ is the sample matrix, $\mathbf{Y} \in \mathbb{R}^{n \times c}$ is label matrix, $\mathbf{A} \in \mathbb{R}^{k \times c}$ is the orthogonal matrix, $\mathbf{P} \in \mathbb{R}^{d \times k}$ is the auxiliary matrix and $\mathbf{Q} \in \mathbb{R}^{d \times k}$ is projection matrix. α and β are the parameters to balance the three terms. $\|\cdot\|_{2,1}$ is the $L_{2,1}$ -norm definition.

In (3), the first term is the loss function as in LSR where matrix $\mathbf{Q} \mathbf{A}$ with size $d \times c$ is similar to the matrix $\mathbf{W} \in \mathbb{R}^{d \times c}$. The difference between JSRR and LSR is that the size of the projection matrix \mathbf{Q} in JSRR is $d \times k$ while the projection matrix in LSR or RR is $d \times c$, which means the JSRR can obtain k projections for feature selection while LSR and RR can only obtain at most c projections (note that k is a variable and it can be set as any integer). If we set $k > c$, then JSRR can break through the limitation of the class number and thus solve the small-class problem. Another difference between JSRR and LSR or RR is that JSRR uses $L_{2,1}$ -norm as the basic measurement on the loss function, by which the model is more robust to outliers. Compared with RFS, JSRR does not have the small-class problem. Moreover, JSRR can degrade to conceptual framework of RFS when $\alpha = 1$. That is, the second term in (3) is released, JSRR is the $L_{2,1}$ -norm based RR that has similar property with RFS. When $\alpha = 0$, JSRR becomes a unsupervised learning method. In this case, JSRR is the joint sparse principle component analysis method as proposed in [19]. Otherwise, if $\alpha \neq 0$ and $\alpha \neq 1$, JSRR holds the property of RFS and PCA, which enhances the robustness to outliers and at the same time obtains sparse principle components (PCs) for feature selection.

In summary, from the objective function of JSRR, we can know that it uses $L_{2,1}$ -norm instead of L_2 -norm on all terms to enhance robustness to outlier and simultaneously guarantee joint sparsity for discriminative feature selection. Also, since the projection matrix is \mathbf{Q} with size $d \times k$, JSRR can obtain k projections instead of c projection to solve the small-class problem so as to obtain enough projections for feature selection and classification. In addition, JSRR enjoys the property of RFS and PCA when $\alpha \neq 1$ and $\alpha \neq 0$, or it can also be used for unsupervised learning when $\alpha = 0$.

2.4 The Optimal Solution

There are three variables in (3) and the optimization problem is not convex, which means that we cannot obtain the optimal solution directly. Therefore, we need to develop an iterative algorithm to solve the optimization problem.

First, from (3), we have

$$\begin{aligned}
 & \alpha \|\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A}\|_{2,1} + (1 - \alpha) \|\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X}\|_{2,1} + \beta \|\mathbf{Q}\|_{2,1} \\
 & = \alpha \text{tr}[(\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A})^T \mathbf{D} (\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A})] \\
 & + (1 - \alpha) \text{tr}[(\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X})^T \mathbf{D}_1 (\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X})] + \beta \text{tr}(\mathbf{Q}^T \mathbf{D}_2 \mathbf{Q})
 \end{aligned} \tag{4}$$

where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with the elements represented as

$$\mathbf{D}_{ii} = \frac{1}{2\|(\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A})^i\|_2} \quad (5)$$

where $(\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A})^i$ denotes the i -th row of the matrix $(\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A})$.

Similarly, the elements of diagonal matrix $\mathbf{D}_1 \in \mathbb{R}^{d \times d}$ and $\mathbf{D}_2 \in \mathbb{R}^{d \times d}$ are represented as

$$(\mathbf{D}_1)_{ii} = \frac{1}{2\|(\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X})^i\|_2} \quad (6)$$

$$(\mathbf{D}_2)_{ii} = \frac{1}{2\|\mathbf{Q}^i\|_2} \quad (7)$$

where $(\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X})^i$ and \mathbf{Q}^i denote the i -th row of the matrix $(\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X})$ and \mathbf{Q} , respectively.

From (4), we have

$$\begin{aligned} & \alpha \|\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A}\|_{2,1} + (1 - \alpha) \|\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X}\|_{2,1} + \beta \|\mathbf{Q}\|_{2,1} \\ & = \alpha \text{tr}[(\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A})^T \mathbf{D} (\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A})] \\ & \quad + (1 - \alpha) \text{tr}[(\sqrt{\mathbf{D}_1} (\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X}))^T (\sqrt{\mathbf{D}_1} (\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X}))] + \beta \text{tr}(\mathbf{Q}^T \mathbf{D}_2 \mathbf{Q}) \end{aligned} \quad (8)$$

From (8), we can know that the objective function in (3) can be rewritten as the following optimization problem

$$\begin{aligned} (\mathbf{Q}^*, \mathbf{P}^*, \mathbf{A}^*) = \arg \min_{\mathbf{Q}, \mathbf{P}, \mathbf{A}} & \alpha \|\sqrt{\mathbf{D}} (\mathbf{Y} - \mathbf{X}^T \mathbf{Q} \mathbf{A})\|_F^2 + (1 - \alpha) \|\sqrt{\mathbf{D}_1} (\mathbf{X} - \mathbf{P} \mathbf{Q}^T \mathbf{X})\|_F^2 + \beta \|\sqrt{\mathbf{D}_2} \mathbf{Q}\|_F^2 \\ & \text{s.t. } \mathbf{A} \mathbf{A}^T = \mathbf{I} \end{aligned} \quad (9)$$

Let $\bar{\mathbf{P}} = \sqrt{\mathbf{D}_1} \mathbf{P}$, $\bar{\mathbf{Q}} = \sqrt{\mathbf{D}_1}^{-1} \mathbf{Q}$, then (9) is equal to

$$\begin{aligned} (\bar{\mathbf{Q}}^*, \bar{\mathbf{P}}^*, \mathbf{A}^*) = \arg \min_{\bar{\mathbf{Q}}, \bar{\mathbf{P}}, \mathbf{A}} & \alpha \|\sqrt{\mathbf{D}} (\mathbf{Y} - \mathbf{X}^T \sqrt{\mathbf{D}_1} \bar{\mathbf{Q}} \mathbf{A})\|_F^2 \\ & + (1 - \alpha) \left\| \sqrt{\mathbf{D}_1} \mathbf{X} - \bar{\mathbf{P}} \bar{\mathbf{Q}}^T \sqrt{\mathbf{D}_1} \mathbf{X} \right\|_F^2 + \beta \|\sqrt{\mathbf{D}_2} \sqrt{\mathbf{D}_1} \bar{\mathbf{Q}}\|_F^2 \\ & \text{s.t. } \mathbf{A} \mathbf{A}^T = \mathbf{I} \end{aligned} \quad (10)$$

By imposing the orthogonal constraint $\bar{\mathbf{P}}^T \bar{\mathbf{P}} = \mathbf{I}$ to (10), we have

$$\begin{aligned} (\bar{\mathbf{Q}}^*, \bar{\mathbf{P}}^*, \mathbf{A}^*) = \arg \min_{\bar{\mathbf{Q}}, \bar{\mathbf{P}}, \mathbf{A}} & \alpha \|\sqrt{\mathbf{D}} (\mathbf{Y} - \mathbf{X}^T \sqrt{\mathbf{D}_1} \bar{\mathbf{Q}} \mathbf{A})\|_F^2 \\ & + (1 - \alpha) \|\sqrt{\mathbf{D}_1} \mathbf{X} - \bar{\mathbf{P}} \bar{\mathbf{Q}}^T \sqrt{\mathbf{D}_1} \mathbf{X}\|_F^2 + \beta \|\sqrt{\mathbf{D}_2} \sqrt{\mathbf{D}_1} \bar{\mathbf{Q}}\|_F^2 \\ & \text{s.t. } \mathbf{A} \mathbf{A}^T = \mathbf{I}, \bar{\mathbf{P}}^T \bar{\mathbf{P}} = \mathbf{I} \end{aligned} \quad (11)$$

Take the optimization problem in (11) as the objective optimization problem in this paper, we can first obtain the optimal solution of $\bar{\mathbf{P}}$ and $\bar{\mathbf{Q}}$, and then we obtain the optimal solution of \mathbf{P} and \mathbf{Q} as $\mathbf{P} = \sqrt{\mathbf{D}_1}^{-1}\bar{\mathbf{P}}$, $\mathbf{Q} = \sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}$.

$\bar{\mathbf{Q}}$ Step: Suppose $\bar{\mathbf{P}}$ and \mathbf{A} are fixed, there exists an optimal matrix $\bar{\mathbf{P}}_\perp$ that guarantees $[\bar{\mathbf{P}}, \bar{\mathbf{P}}_\perp]$ is a $d \times d$ column orthogonal matrix. From the optimization problem in (11), we have

$$\begin{aligned} & \|\sqrt{\mathbf{D}_1}\mathbf{X} - \bar{\mathbf{P}}\bar{\mathbf{Q}}^T\sqrt{\mathbf{D}_1}\mathbf{X}\|_F^2 = \|\mathbf{X}^T\sqrt{\mathbf{D}_1} - \mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}\bar{\mathbf{P}}^T\|_F^2 \\ & = \|\mathbf{X}^T\sqrt{\mathbf{D}_1}[\bar{\mathbf{P}}, \bar{\mathbf{P}}_\perp] - \mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}\bar{\mathbf{P}}^T[\bar{\mathbf{P}}, \bar{\mathbf{P}}_\perp]\|_F^2 \\ & = \|\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}} - \mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}\bar{\mathbf{P}}^T\bar{\mathbf{P}}\|_F^2 + \|\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}}_\perp - \mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}\bar{\mathbf{P}}^T\bar{\mathbf{P}}_\perp\|_F^2 \\ & = \|\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}} - \mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}\|_F^2 + \|\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}}_\perp\|_F^2 \end{aligned} \tag{12}$$

In (12), since $\bar{\mathbf{P}}$ is given and $\|\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}}_\perp\|_F^2$ is a constant, the optimization problem in (11) becomes

$$\begin{aligned} \bar{\mathbf{Q}}^* = \arg \min_{\bar{\mathbf{Q}}} \alpha & \|\sqrt{\mathbf{D}}(\mathbf{Y} - \mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}\mathbf{A})\|_F^2 \\ & + (1 - \alpha)\|\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}} - \mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}\|_F^2 + \beta\|\sqrt{\mathbf{D}_2}\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}\|_F^2 \\ & \text{s.t. } \mathbf{A}\mathbf{A}^T = \mathbf{I} \end{aligned} \tag{13}$$

From (13), we have

$$\begin{aligned} & \alpha \text{tr}(\mathbf{Y}^T\mathbf{D}\mathbf{Y} - 2\bar{\mathbf{Q}}^T\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{D}\mathbf{Y}\mathbf{A}^T + \bar{\mathbf{Q}}^T\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{D}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}) \\ & + (1 - \alpha)\text{tr}(\bar{\mathbf{P}}^T\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}} - 2\bar{\mathbf{Q}}^T\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}} + \bar{\mathbf{Q}}^T\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}) \\ & + \beta\text{tr}(\bar{\mathbf{Q}}^T\sqrt{\mathbf{D}_1}\mathbf{D}_2\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}) \\ & \text{s.t. } \mathbf{A}\mathbf{A}^T = \mathbf{I} \end{aligned} \tag{14}$$

By the derivative of (14) with respect to $\bar{\mathbf{Q}}$ to be 0, we have

$$\begin{aligned} & \alpha(\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{D}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}} - \sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{D}\mathbf{Y}\mathbf{A}^T) \\ & + (1 - \alpha)(\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}} - \sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}}) + \beta\text{tr}(\sqrt{\mathbf{D}_1}\mathbf{D}_2\sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}) = 0 \end{aligned} \tag{15}$$

Then, we have

$$\bar{\mathbf{Q}} = [\sqrt{\mathbf{D}_1}(\alpha\mathbf{X}\mathbf{D}\mathbf{X}^T + (1 - \alpha)\mathbf{X}\mathbf{X}^T + \beta\mathbf{D}_2)\sqrt{\mathbf{D}_1}]^{-1}(\alpha\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{D}\mathbf{Y}\mathbf{A}^T + (1 - \alpha)\sqrt{\mathbf{D}_1}\mathbf{X}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}}) \tag{16}$$

Since $\mathbf{Q} = \sqrt{\mathbf{D}_1}\bar{\mathbf{Q}}$, then

$$\mathbf{Q} = [\alpha\mathbf{X}\mathbf{D}\mathbf{X}^T + (1 - \alpha)\mathbf{X}\mathbf{X}^T + \beta\mathbf{D}_2]^{-1}(\alpha\mathbf{X}\mathbf{D}\mathbf{Y}\mathbf{A}^T + (1 - \alpha)\mathbf{X}\mathbf{X}^T\sqrt{\mathbf{D}_1}\bar{\mathbf{P}}) \tag{17}$$

\bar{P} Step: Suppose \bar{Q} and A are given, from (14), we have

$$\bar{P}^* = \arg \max_P \text{tr}(\bar{P}^T \sqrt{D_1} X X^T \sqrt{D_1} \bar{Q}) \quad (18)$$

Theorem 1. [20] Suppose $G \in R^{c \times k}$ is a matrix with rank of k and $Z \in R^{c \times k}$ is an orthogonal matrix. The optimization problem

$$Z = \arg \min \text{tr}(Z^T G) \quad s.t. \quad Z^T Z = I_k \quad (19)$$

Algorithm 1 The algorithm of JSRR

Input: The data matrix $X \in R^{d \times n}$, the label matrix $Y \in R^{n \times c}$, the projection number k , the parameter α and β .

Initialize $A \in R^{k \times c}$, $P \in R^{d \times k}$, $Q \in R^{d \times k}$, $D \in R^{n \times n}$, $D_1 \in R^{d \times d}$ and $D_2 \in R^{d \times d}$.

repeat

 compute \bar{Q} using (16); update Q using (17); compute \bar{P} using (21); update P using (22);
 update A using (25); update D using (5); update D_1 using (6); update D_2 using (7);

until converge

Output: $Q \in R^{d \times k}$ with k sparse projections.

can be solved by singular value decomposition (SVD) of G , i.e. $G = \check{U} \check{D} \check{V}^T$, then $Z = \check{U} \check{V}^T$.

From Theorem 1, we can know that the optimal solution of \bar{P} can be obtain by SVD of $\sqrt{D_1} X X^T \sqrt{D_1} \bar{Q}$, that is,

$$\sqrt{D_1} X X^T \sqrt{D_1} \bar{Q} = U D V^T \quad (20)$$

then

$$\bar{P} = U V^T \quad (21)$$

Since $P = \sqrt{D_1}^{-1} \bar{P}$, we have

$$P = \sqrt{D_1}^{-1} U V^T \quad (22)$$

A Step: Suppose \bar{Q} and \bar{P} are given, from (14), we have

$$\begin{aligned} A^* = \arg \max_A \alpha \text{tr}(A Y^T D X^T \sqrt{D_1} \bar{Q}) \\ s.t. \quad A A^T = I \end{aligned} \quad (23)$$

Similarly, according to Theorem 1, we know that the optimal solution of \mathbf{A} can be obtain by SVD of $\mathbf{Y}^T \mathbf{D} \mathbf{X}^T \sqrt{\mathbf{D}_1} \bar{\mathbf{Q}}$, that is,

$$\mathbf{Y}^T \mathbf{D} \mathbf{X}^T \sqrt{\mathbf{D}_1} \bar{\mathbf{Q}} = \tilde{\mathbf{U}} \tilde{\mathbf{D}} \tilde{\mathbf{V}}^T \quad (24)$$

Then, we have

$$\mathbf{A} = \tilde{\mathbf{U}} \tilde{\mathbf{V}}^T \quad (25)$$

The details of the iterative algorithm that solves optimization problem in (3) are shown in Algorithm 1.

3 Experiments

In this section, several benchmark datasets with varying image types are used to evaluate the performance of the proposed JSRR on feature selection and recognition. These datasets include AR dataset, CMU PIE dataset and LFW dataset. In addition, in all experiments, we compare the proposed JSRR with some classical methods including RR, and some state-of-the-art methods including RFS [11], UDFS [12], DLSR [13] and RIPCA [21].

3.1 Datasets Description

The AR database [22] is consist of over 4000 images from 126 individuals. In our experiment, we use the subsection that contains 2,400 images from 120 individuals. All of these images are normalized to 50×40 pixels. This dataset is used to evaluate the performance of JSRR with varying facial expressions, lighting conditions and occlusions.

The CMU PIE dataset [23] contains 41,368 face images from 68 individuals. We use a subset (C29) which has 1632 images from 68 individuals in our experiment. All of these images are cropped to 32×32 pixels.

Labeled Faces in the Wild (LFW) databases [24] is consist of images from 5,749 subjects in uncontrolled environment. In our experiment, we select 4,324 images from 158 subjects in LFW-a dataset.

Recently, deep learning technique is very famous in the field of machine learning and computer vision. Therefore, to explore the performance of JSRR based on the background of deep learning, we use deep features on LFW dataset instead of original image features as input. Similar to [25], we use deep convolutional neural network (CNN) as feature extractor to obtain deep features. After that, JSRR, RR, RFS, UDFS, DLSR and RIPCA are used to perform further feature selection and extraction. The property of the datasets is summarized in Table 1.

Table 1. Description of datasets.

Datasets	# of Samples	Features	classes
AR	2,400	2,000	120
CMU PIE	1,632	1,024	68
LFW	4,324	1,024	158

3.2 Experimental Setting

In our experiments, PCA is used as pre-processing to perform dimensionality reduction. After that, all methods including the comparative methods and the proposed JSRR are used to perform feature selection and extraction and nearest neighbor (NN) classifier is used for classification. The recognition rate is used as the criteria to evaluate the performance of all methods. Each method independently runs 10 times to conduct feature selection and the mean recognition rate is computed.

For JSRR, since there are two parameters, i.e. α and β , need to optimize, we analyze their values in the area of $[-3, \dots, 3]$ and $[10^{-3}, \dots, 10^3]$, respectively. For the comparative methods, their parameters are set as the value that introduced in the original paper. For example, the value of the parameter in RFS, UDFS and DLSR is set as $[10^{-3}, \dots, 10^3]$, $[10^{-3}, \dots, 10^3]$ and $[10^{-4}, \dots, 10^1]$, respectively.

3.3 Experimental Results and Comparison

On AR, CMU PIE and LFW dataset, l_1 ($l_1 = 2, 3$), l_2 ($l_2 = 4, 5$) l_3 ($l_3 = 4, 5$) images of each individual are selected for training while the rest of the images are used for testing.

To explore the optimal values of α and β , we report in Fig. 1(a) the recognition rates with varying values of $\alpha \in [-3, -2, \dots, 3]$ and $\beta \in [10^{-3}, 10^{-2}, \dots, 10^3]$ on AR database. From Fig. 1(a), we can know that JSRR obtain the best performance while α lies in the area of $[-3, -2, -1]$ and β lies in the area of $[10^{-3}, 10^{-2}, \dots, 10^3]$. Therefore, we use these values for α and β on all experiments for simplicity.

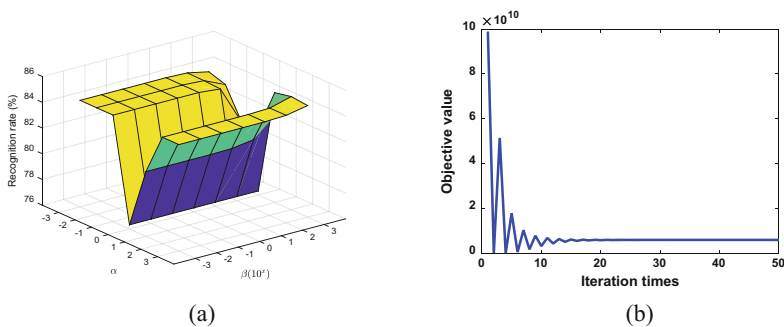


Fig. 1. (a) Sensitivity analysis of parameters, (b) convergence curve on AR database.

The average recognition rates with varying dimension on AR and CMU PIE dataset are shown in Fig. 2 while the maximum average recognition versus the dimension and the standard deviation on AR, CUM PIE and LFW dataset are listed in Tables 2, 3 and 4, respectively. The convergence curve of the proposed JSRR on AR dataset is presented in Fig. 1(b).

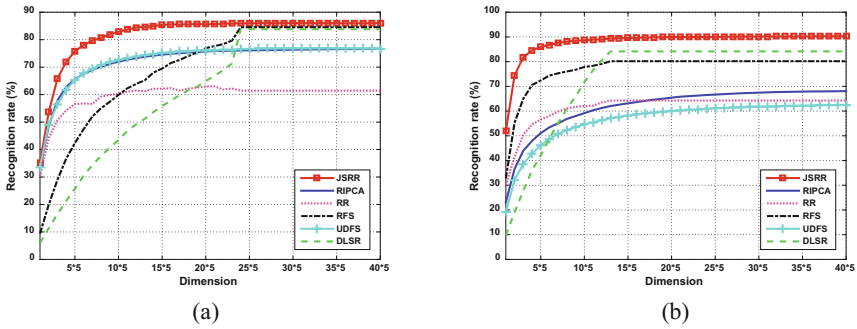


Fig. 2. The recognition rate versus the dimension on (a) AR, (b) CUM PIE dataset.

Table 2. On AR dataset, the performance (recognition rate, standard deviation (%), dimension) of all methods.

l_1	RR	RFS	UDFS	DLSR	RIPCA	JSRR
2	49.62 ± 6.69 105	80.74 ± 12.00 120	71.20 ± 5.01 180	81.74 ± 9.03 120	71.03 ± 5.02 195	83.93 ± 8.25 195
3	62.97 ± 5.21 105	84.54 ± 12.37 120	76.92 ± 5.12 180	83.81 ± 10.08 120	76.70 ± 5.26 200	86.08 ± 9.99 135

Table 3. On CMU PIE dataset, the performance (recognition rate, standard deviation (%), dimension) of all methods.

l_2	RR	RFS	UDFS	DLSR	RIPCA	JSRR
4	64.28 ± 9.71 65	80.18 ± 8.81 65	62.54 ± 10.15 200	84.15 ± 8.92 65	68.09 ± 9.33 200	90.29 ± 3.58 185
5	72.56 ± 8.98 65	87.40 ± 7.42 65	72.46 ± 13.57 200	86.04 ± 6.57 65	71.82 ± 7.34 200	90.75 ± 3.59 200

Table 4. On LFW dataset, the performance (recognition rate, standard deviation (%), dimension) of all methods.

l_3	RR	RFS	UDFS	DLSR	RIPCA	JSRR
4	94.21 ± 1.36 155	98.24 ± 0.00 155	97.86 ± 0.00 70	96.34 ± 0.00 155	98.35 ± 0.00 130	98.44 ± 0.04 170
5	92.64 ± 3.63 155	98.53 ± 0.00 155	98.02 ± 0.00 90	97.62 ± 0.00 155	98.64 ± 0.00 90	98.70 ± 0.05 135

According to the experimental results, we have the following interesting observations:

- (1) In all experiments, JSRR obtains the best performance. The potential reason for this phenomenon is that JSRR integrates the property of RFS and PCA, with which it can obtain joint sparse projections for discriminative feature selection and extraction. Furthermore, by considering the label information on the loss function, JSRR is able to enhance the discriminability of the sparse PCs.
- (2) On AR and CMU PIE dataset, JSRR, RFS and DLSR obtain better performance than other methods, which indicates that utilizing $L_{2,1}$ -norm on loss function and regularization term is able to enhance the robustness and guarantee joint sparsity, such that they are superior for feature selection and classification.
- (3) The experimental results demonstrate that JSRR can solve the small-class problem. For example, Fig. 2(b) and Table 3 show that RR, RFS and DLSR obtain the best recognition rate when dimension is 65 (the class number is 68). However, JSRR obtains its highest recognition rate when dimension is 185. It indicates that the number of the projection learned by JSRR is not limited by the number of class, i.e. JSRR can solve the small-class problem.

4 Conclusion

In this paper, we propose a method called jointly sparse reconstructed regression (JSRR) which uses joint $L_{2,1}$ -norm as the basic measurement on the objective function. By doing so, JSRR is more robust to outliers than the L_2 -norm based methods. Also, it can obtain jointly sparse projection for discriminative feature selection. Different from LSR, RR and their extensions, JSRR is able to solve the small-class problem, which enables it to obtain enough projections to perform feature selection and extraction even though the class number is small. Under some certain conditions, JSRR can degrade to RFS or sparse version of PCA, which indicates that JSRR at least guarantees the effectiveness of RFS and PCA. To solve the optimization problem of JSRR, an iterative algorithm is proposed. Experimental results on three well-known facial datasets demonstrate that JSRR is superior to the classical RR and some state-of-the-art feature selection methods.

Acknowledgment. This work was supported in part by the Natural Science Foundation of China (Grant 61573248, Grant 61773328, Grant 61773328 and Grant 61703283), Research Grant of The Hong Kong Polytechnic University (Project Code:G-UA2B), China Postdoctoral Science Foundation (Project 2016M590812 and Project 2017T100645), the Guangdong Natural Science Foundation (Project 2017A030313367 and Project 2017A030310067), and Shenzhen Municipal Science and Technology Innovation Council (No. JCYJ20170302153434048 and No. JCYJ20160429182058044).

References

1. Yang, J., Zhang, D., Frangi, A.F., Yang, J.: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 131–137 (2004)
2. Fan, Z., et al.: Modified principal component analysis: an integration of multiple similarity subspace models. *IEEE Trans. Neural Netw. Learn. Syst.* **25**, 1538–1552 (2017)
3. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 711–720 (1997)
4. Zhong, F., Zhang, J., Li, D.: Discriminant locality preserving projections based on L1-norm maximization. *IEEE Trans. Neural Netw. Learn. Syst.* **25**, 2065–2074 (2014)
5. Tibshirani, R.: Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288 (1996)
6. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. Ser. B.* **67**, 301–320 (2005)
7. Majumdar, A., Ward, R.K.: Classification via group sparsity promoting regularization. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 861–864 (2009)
8. Li, C., Shao, Y., Deng, N.: Robust L1-norm two-dimensional linear discriminant analysis. *Neural Netw.* **65**, 92–104 (2015)
9. Gong, P., Zhang, C., Lu, Z., Huang, J.Z., Ye, J.: A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In: *International Conference on Machine Learning*, vol. 28, pp. 37–45 (2013)
10. Xu, Y., Zhang, B., Zhong, Z.: Multiple representations and sparse representation for image classification. *Pattern Recognit. Lett.* **68**, 9–14 (2015)
11. Nie, F., Huang, H., Cai, X., Ding, C.: Efficient and robust feature selection via joint L2,1-norms minimization. *Adv. Neural Inf. Process. Syst.* **23**, 1813–1821 (2010)
12. Yang, Y., Shen, H.T., Ma, Z., Huang, Z., Zhou, X.: L2,1-norm regularized discriminative feature selection for unsupervised learning. In: *International Joint Conference on Artificial Intelligence*, pp. 1589–1594 (2011)
13. Xiang, S., Nie, F., Meng, G., Pan, C., Zhang, C.: Discriminative least squares regression for multiclass classification and feature selection. *IEEE Trans. Neural Netw. Learn. Syst.* **23**, 1738–1754 (2012)
14. He, R., Tan, T., Wang, L., Zheng, W.: L2,1 regularized coreentropy for robust feature selection. In: *Computer Vision and Pattern Recognition*, pp. 2504–2511 (2012)
15. Shi, X., Yang, Y., Guo, Z., Lai, Z.: Face recognition by sparse discriminant analysis via joint L2,1-norm minimization. *Pattern Recognit.* **47**, 2447–2453 (2014)
16. Gu, Q., Li, Z., Han, J.: Joint feature selection and subspace learning. In: *International Joint Conference on Artificial Intelligence*, vol. 55, pp. 1294–1299 (2011)
17. Huang, J., Li, G., Huang, Q., Member, S., Wu, X.: Joint feature selection and classification for multilabel learning. *IEEE Trans. Cybern.* 1–14 (2017)
18. Yang, J., Chu, D., Zhang, L., Xu, Y., Yang, J.: Sparse representation classifier steered discriminative projection with applications to face recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **24**, 1023–1035 (2013)
19. Yi, S., Lai, Z., He, Z., Cheung, Y.M., Liu, Y.: Joint sparse principal component analysis. *Pattern Recognit.* **61**, 524–536 (2017)
20. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. *J. Comput. Graph. Stat.* **15**, 1–30 (2004)

21. Lai, Z., Xu, Y., Yang, J., Shen, L., Zhang, D.: Rotational invariant dimensionality reduction algorithms. *IEEE Trans. Cybern.* **47**, 3733–3746 (2017)
22. Martinez, A.A., Benavente, R.: The AR face database. CVC Technical report #24 (1998)
23. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 1615–1618 (2003)
24. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07-49. University Massachusetts, Amherst (2007)
25. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31



Multi-scale Attributed Graph Kernel for Image Categorization

Duo Hu, Qin Xu^(✉), Jin Tang, and Bin Luo

School of Computer Science and Technology, Anhui University, Hefei, China
michaelhd524@163.com, {xuqin,tj,binluo}@ahu.edu.cn

Abstract. The spatial pyramid matching has been widely adopted for scene recognition and image retrieval. It splits the image into sub-regions and counts the local features within the sub-region. However, it has not captured the spatial relationship between the local features located in the sub-region. This paper proposes to construct the multi-scale attributed graphs which involve the vocabulary label to characterize the spatial structure of the local features at different scales. We compute the distances of any two attributed graph corresponding to the image grids and find the optimal matching to aggregate. Then we poll the distances of graphs at different scales to build the kernel for image classification. We conduct our method on the Caltech 101, Caltech 256, Scene Categories, and Six Actions datasets and compare with five methods. The experiment results demonstrate that our method can provide a good accuracy for image categorization.

Keywords: Image classification · Multi-scale attributed graph
Graph distance

1 Introduction

Image Categorization, which has a quite wide range of applications, such as face recognition, scene classification and pedestrian tracking, is a challenging task in computer vision. It is undoubtedly of great theoretical and practical significance to study the robust and accurate image classification algorithm. How to find the correct classification of an unlabel image from a large scale image database has been a research spot for several decades and numerous methods have been developed.

The approach bag of words (BoW) has been widely used in image classification [1–3]. BoW based methods use image visual features (e.g. SIFT [4]) to build a dictionary of visual words and computing a histogram for each image for recognition. However, the BoW method does not contain spatial and structural information of the image. In this respect, one limitation of the BoW approach is that it can not encode the spatial distribution of visual words within an image.

To characterize the spatial layout of the local features, the spatial pyramid [5] divides the image into different regions at different levels and computes a

BoW for each region, and the final image descriptor as the concatenation of the histograms from all regions. For the same reason, latent pyramidal regions (LPR) [6] are trained by combining the benefits of spatial pyramid representation using nonlinear feature coding and latent SVM. Yang et al. [7] proposed the linear spatial pyramid matching using sparse coding (ScSPM) and Wang et al. [8] proposed the locality-constrained linear coding method to improve the ScSPM method by adding the local constraints. In order to obtain the vector based on BoW with certain invariance, Cao et al. presented two methods of linear BoW and annular BoW to improve the robustness to some degree [9].

In recent years, graph matching algorithms have been applied to solve image classification [10, 11]. One of the most popular methods to perform graph matching is the graph edit distance [12–15]. Jouili et al. [12] used Hungarian method with a vector which encodes vertices and edges of the same representation to compute a suboptimal cost of edit distance. Zhou et al. [16] proposed a deformable graph matching method to match graphics that are subject to global rigid and non rigid geometric constraints. The bag of graph [13] and bag of visual graphs [14] combines the spatial locations of interest points and their labels defined in terms of the traditional visual-word codebook to define a set of connected graphs, then defines descriptors for image classification based on graph local structures. Lee et al. [17] generalizes the formula of hyper-graph matching to cover arbitrary sequence of feature relations and obtained a new graph matching algorithm by reinterpreting the concept of random walk on hyper-graph. Zhang et al. [18] proposed a saliency-guided graphlet selection algorithm for image categorization. In the multi-graph-view respects, Wu et al. [19] proposes a multi-graph-view model to represent and classify complex targets. Mousavi et al. [20] generated a graph pyramid based on the selected graph summarization algorithm to provide the required information for classification.

The matching node embeddings [21] is presented as the graph kernel based on the pyramid match kernel. It restricts the matchings only between vertices that share same labels. However, the interest points have not assigned labels. Thus this method is not competent for the graph based on interest points without tags. Our approach takes this into account that applying the weighted Hungarian method to find the most similar graph, that can be a good way to overcome this problem.

In this paper, we propose to construct a multi-scale attributed graph model for image classification, where the spatial structure relation between the interest points of the image at different scales are captured. The graphs are pruned to give more efficient structure information for categorization. At each scale, the distance of the attributed graphs are calculated to find the optimal matched graphs. Final the distances are accumulated with weight to built the kernel for SVM.

The rest of the paper is organized as follows. We first present the proposed multi-scale attributed graph for image representation in Sect. 2, and then compute the distance between the attributed graphs corresponding to image grids in Sect. 3. In Sect. 4, the kernel for classification is built by accumulating the

distances between the matched graphs. The experimental results on four public datasets are presented and discussed in Sect. 5. Finally, conclusions are drawn in Sect. 6.

2 The Multi-scale Attributed Graph Model for Image Representation

To describe the structure and spatial features of the images at different scales, we define multi-scale attributed graphs $G^l = (V^l, E^l, A^l)$, where l denotes the scale or level factor, the nodes set $V^l = \{v_1, v_2, \dots, v_n\}$ corresponds to the image feature points $F = \{f_1, f_2, \dots, f_n\}$, which obtained by extracting the SIFT features of the images in our experiment, the edges set $E^l = \{e_{ij}\}$ are constructed by delaunay triangulation, and A^l denotes the attribute of the node set V^l , for a node v_i , its attribute is defined as,

$$A^l_{v_i} = \{av_i, degree(v_i), \{ae_i\}\} \tag{1}$$

where av_i is the label of node v_i which corresponds to the feature point f_i . In terms of the widely used bag of words, we assign a vocabulary label to each node, $degree(v_i)$ is the degree of the node v_i , $\{ae_i\}$ is the attribute set of all the edges which are adjacent to the node v_i . There exist many methods for constructing graph based on images, such as k-nearest neighbor graph [22] and deep learning hash [23]. We use the delaunay triangulation method here for its stability and efficiency. To characterize the image structure at different scale, we split an image into a sequence of grids at each scale $l \in (0, \dots, L)$, such that a total of $S = 2^{sl}$ image grids are obtained, where s is the dimension of the images. For each grid, we construct an attributed graph on the feature points, as shown in Fig. 1. These graphs form the multi-scale structure representation of an image.

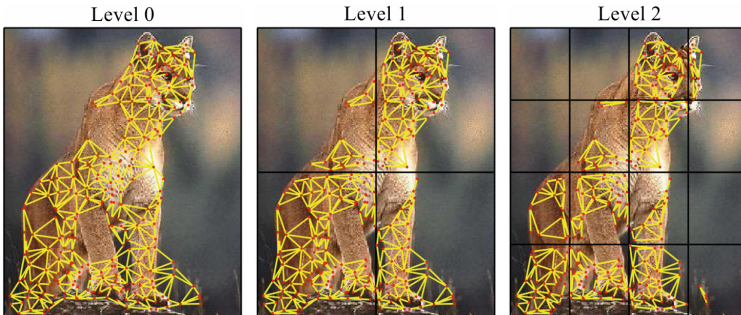


Fig. 1. The multi-scale attributed graph extraction from an cougar body.

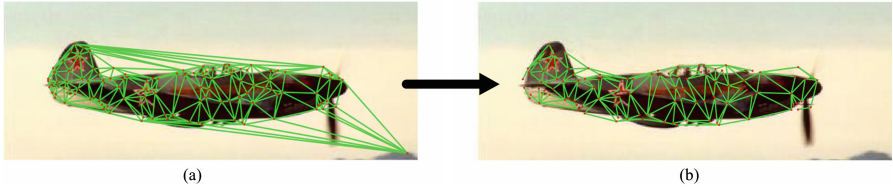


Fig. 2. The graph model for an image in Caltech 101. (a) The delaunay triangulation graph on the feature points; (b) Our graph constructed after pruning.

Since the images have not been preprocessed as segmentation or salient analysis, the images usually have the objective, background and noise. The multi-scale attributed graphs built from the original image will contains the additional structure information which is not related to the objective. For example, Fig. 2(a) shows the attributed graph constructed from an image in Caltech 101 at level 0. We can see that one point in the background in the lower right corner of the image is connected to feature of the aircraft. Furthermore, the feature points of the tail and the head of the aircraft are also connected. However, these edges are useless to reflect the structure of the aircraft and not helpful for image classification. Therefore, we consider to prune the graph, specifically, remove the edges which connect the points with long distance and short distance, as shown in Fig. 2(b). Let m be the value of the longest edge of the constructed graphs for one image, we delete the edges longer than βm and shorter than αm , where $0 < \alpha < \beta < 1$. In the experiments, we choose $\alpha = 0.1$ and $\beta = 0.6$. Because experiments show that the short edges can not improve the classification but increase the computational complexity. We can effectively avoid the error structure with the complicated background, and focus on the local structure of the image by pruning edges and constructing the multi-scale attributed graphs. Moreover, the graph after pruning becomes sparse and computational efficient.

3 Graph Distance Based on Node Attributes

To match the multi-scale structure between two images, we compute the distance between the multi-scale graphs constructed from two images. The graph distance is obtained based on the node attributes using the heterogeneous euclidean overlap metric (HEOM) [12], which can handle the numeric and symbolic attributes of nodes. The distance of two nodes v_i and v_j is defined as their distance between the node attribute \mathbf{A}_i and \mathbf{A}_j ,

$$d(\mathbf{A}_i, \mathbf{A}_j) = \sqrt{\sum_{k=0}^N q(\mathbf{A}_i(k), \mathbf{A}_j(k))^2} \quad (2)$$

where N refers the length of the longest node signature of v_i and v_j , and

$$q(\mathbf{A}_i(k), \mathbf{A}_j(k)) = \begin{cases} \frac{|\mathbf{A}_i(k) - \mathbf{A}_j(k)|}{\text{range}} & \text{if } \mathbf{A}_i(k) \text{ and } \mathbf{A}_j(k) \text{ are both numeric} \\ R(\mathbf{A}_i(k), \mathbf{A}_j(k)) & \text{if } \mathbf{A}_i(k) \text{ and } \mathbf{A}_j(k) \text{ are both symbolic} \\ 1 & \text{if } \mathbf{A}_i(k) \text{ or } \mathbf{A}_j(k) \text{ is missing} \end{cases} \quad (3)$$

where

$$R(\mathbf{A}_i(k), \mathbf{A}_j(k)) = \begin{cases} 0 & \text{if } \mathbf{A}_i(k) = \mathbf{A}_j(k) \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

and *range* is used to normalize the distance of the numeric attribute.

The distance between an attributed graph corresponding to the grid i in image I_1 and an attributed graph corresponding to the grid j in image I_2 at the same scale is computed as [12]:

$$D(G_1(i), G_2(j)) = \frac{\bar{M}}{|\mathbf{M}|} + ||G_1(i)| - |G_2(j)|| \quad (5)$$

where \bar{M} is the optimum graph matching cost of two attributed graphs $G_1(i)$ and $G_2(j)$, the \mathbf{M} is the distance matrix of two attributed graphs that each element of matrix corresponds to the distance between a vertex of graph $G_1(i)$ and a vertex of graph $G_2(j)$. The node matching between two attributed graphs $G_1(i)$ and $G_2(i)$ is carry out by the hungarian method. Then the optimum graph matching cost \bar{M} is computed by calculating the sum of the distance between two correspondence points. $|\mathbf{M}|$ is a normalization constant that refers to the number of matched vertices. $|G_1(i)|$ is the number of vertices in graph $G_1(i)$. The Eq. (5) represents the matching cost normalized by the matching size, and is effected by the sizes of the two graphs.

4 Multi-scale Attributed Graph Kernel Computation

When the distances between any two attributed graphs corresponding to two grids in the different images are computed at a scale, for convenience, we use the efficient hungarian method to find the optimal matched graphs correspond to two different images and get c distances $\{D_1, D_2, \dots, D_c\}$ between the matched graphs. Inspired by the concept of graph kernel [24], which compares and counts the common subgraphs between two graphs. We compute a kernel by accumulating the distances between the matched graphs from two images, i.e.

$$\kappa(G_1^l, G_2^l) \propto \exp\left(-\frac{\sum_{i=1}^c w_i \cdot D_i}{c}\right) \quad (6)$$

where w_i is the reciprocal of the total number of vertices of the matched graphs corresponding to two grids.

The final kernel is then the sum of all the level kernels,

$$K(G_1, G_2) = \sum_{l=0}^L \frac{1}{2^{L-l}} \kappa(G_1^l, G_2^l) \quad (7)$$

where the weight associated with level l is set to $\frac{1}{2^{L-l}}$, which are inversely proportional to the number of the grids which increases as the level increases. The multi-scale attributed graph match kernel we built is a positive semidefinite kernel matrix which can be used by SVM for classification. We summarize the proposed image categorization model in Algorithm 1.

Algorithm 1. The Multi-scale Attributed Graph Kernel algorithm.

Input: H category-labeled training images $\{I^1, I^2, \dots, I^H\}$;

Output: The multi-scale attributed graph match kernel;

(1) Split each image into a sequence of grids at each scale $l \in (0, \dots, L)$;

(2) Use the delaunay triangulation method to characterize the image structure at different scale, built the multi-scale attributed graphs;

(3) Remove the long edges and short edges of the constructed graphs;

(4) Compute the distance between the multi-scale graphs based on the node attributes, use the hungarian method to find the optimal matched graph;

(5) Build the multi-scale attributed graph match kernel by accumulating the distances between the multi-scale matched graphs from two images.

5 Experiments

In this section, we conduct comparative experiments on four benchmark datasets: Caltech 101 [27], Caltech 256 [28], Scene Categories [29], and Six Actions [30]. The performance of the proposed multi-scale attributed graph match kernel is evaluated and compared with traditional bag of words (BoW) [3], the spatial pyramids (SP) [5], BoVG-SP [14], fine-grained dictionary learning (FDL) [25] and word spatial arrangement (WSA) [32] respectively. The experimental results are summarized and analyzed. All experiments are implemented in Matlab 8.6 and executed on a Intel Core i7-6700 3.4 GHz CPU with 16 GB of memory and no effort made to optimize algorithm speed.

5.1 Dataset

The Scene Categories dataset is composed of fifteen scene categories. Each category has 200 to 400 images, and average image size is 300×250 pixels. In experiments, we randomly select 40 images of each class for training and 20 images per class for testing to evaluate the impact of different approaches in image categories.

The Caltech 101 dataset consists of a total of 9146 images, split between 101 different object categories. Each object category contains between 40 and 800 images on average. Each image is about 300×200 pixels in dimension. We use SIFT detector, a codebook of size 300 and 30 images per class for training and the rest for testing.

The Caltech 256 dataset is collected in a similar manner of Caltech 101 which split into a set of 256 object categories containing a total of 30607 images.

The Six Actions dataset collect about 2400 images in total for six action queries, each action class contains about 400 images and the size of each class are 200×200 pixels.

5.2 Baseline

This paper adopts the method in [3] as the baseline approach. The 128-D SIFT descriptors are used for feature extraction and the experiment uses K-means method to get the codebook of size 300. With the increase of scale l , the effect of characterizing image structure is better, but when the scale is larger than 3, the number of grids is too large, the complexity of the algorithm is greatly increased but the improvement of accuracy is limited. Thus the scale level of the multi-scale attributed graph is set to $L = 3$. The LIB-SVM [31] is employed for classification training.

5.3 Results

Table 1 shows the classification results on four datasets. As we can see our method and FDL produce the higher classification accuracy than other methods. Our method achieves highest recognition rates on Scene Categories, Caltech 101 and Six Actions dataset. Taking Scene Categories for example, it is clear that the classification accuracy of MsAG is 79.67%, which is higher than others.

Table 1. Categorization accuracies on four datasets

Dataset	BoW	SP	BOVG-SP	FDL	WSA	MsAG
Scene Categories	69.00%	73.33%	78.00%	82.96%	78.43%	79.67%
Caltech 101	26.43%	34.75%	38.75%	43.73%	39.74%	44.56%
Caltech 256	9.92%	15.70%	16.19%	19.93%	16.08%	18.52%
Six Actions	77.50%	82.50%	83.33%	85.67%	85.16%	86.02%

Figure 3 shows a confusion matrix between the fifteen scene categories, confusion occurs between the classes like kitchen, bedroom, living room, and also between some natural classes, such as coast and open country. The curves in Fig. 4 shows the classification accuracy for different training set sizes on Caltech 101. We partition the dataset into train images (5, 10, 15, 20, 25 and 30 images per class) and test images (limit the number of test images to 30 per class). The figure shows that the accuracy increases with the training size. Our approach has always been better than the other methods when the number increases from 5 to 30. In Fig. 5, the experimental results on Six Actions show that the results of our method is consistent with that on Caltech 101.

Then we compare the classification accuracies of each method for different codebook sizes, the FDL and WSA methods do not involve codebook, so we do

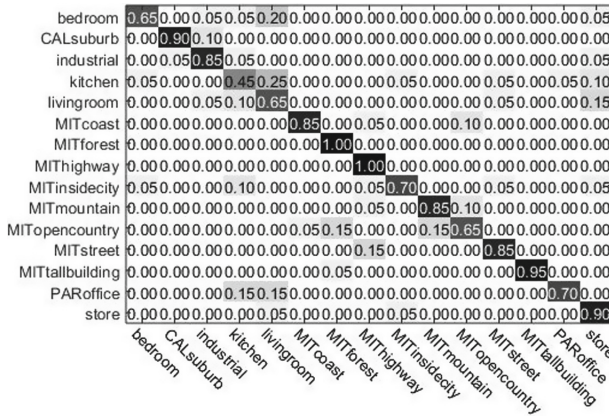


Fig. 3. Confusion matrix for the Scene Category dataset. Average classification rates for individual classes are listed along the diagonal. The entry in the i^{th} row and j^{th} column is the percentage of images from class i that were misidentified as class j .

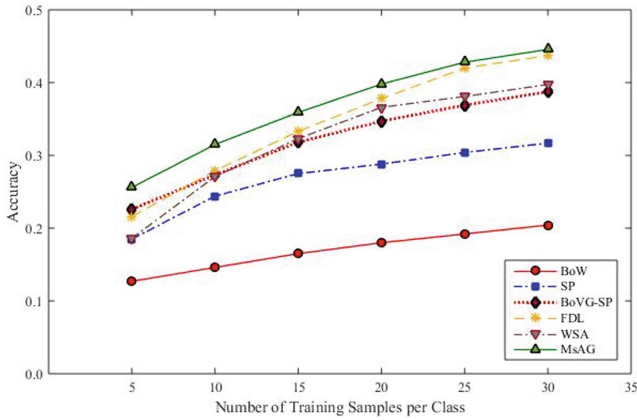


Fig. 4. Classification accuracy for different training set sizes on Caltech 101.

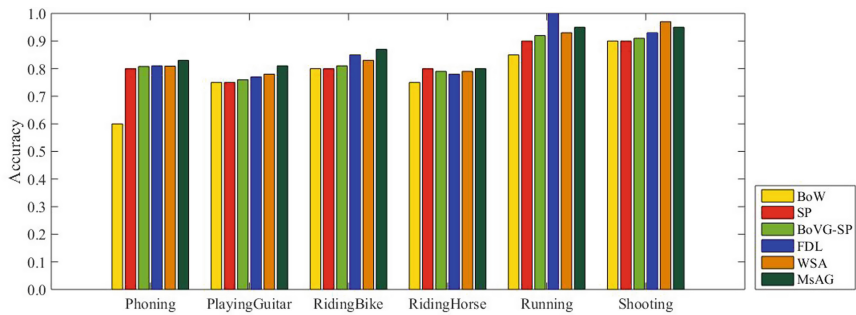


Fig. 5. Performance of BoW, SP, BoVG-SP, FDL, WSA and MsAG on Six Actions.

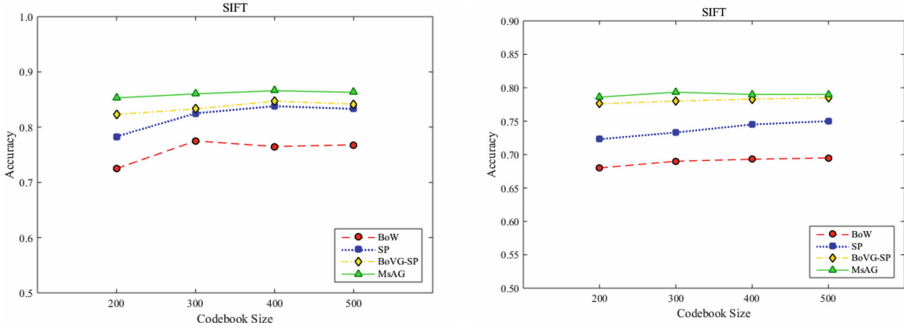


Fig. 6. Different codebook size on the performance of BoW, SP, BoVG-SP and MsAG on (a) Six Actions, (b) Scene Category.

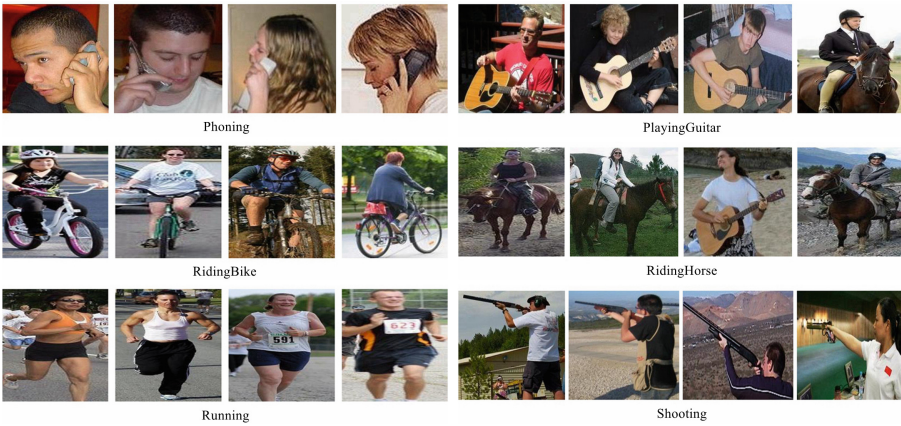


Fig. 7. Partial results of image classification on the Six Actions database.

not compare them in this experiment. As shown on Fig. 6, classification accuracy increases when the codebook size increases from 200 to 500 and remains obtain similar results on both datasets when the size is larger than 300. Comparatively, considering the time consumption of the algorithm, we set the size of the codebook to 300.

Figure 7 shows partial results of image classification on Six Actions database using the MsAG, which show that our method had better recognition accuracy on each label category. Meanwhile, the performance of our method is stable in similar categories problem. We believe that our approach is still very competitive in other conditions.

6 Conclusion

In this paper, we explore the multi-scale attributed graph construction and matching kernel for image classification. This may provide a further step to

utilize the structure information for image recognition. The comparisons on four standard datasets with five approaches, which are BoW, SP, FDL, WSA and BoVG-SP, show the efficiency of our approach.

Our work has been limited the simple edge construction using delaunay triangulation, there are several nature extension that can be taken advantage of. First, we can build different edge sets to form the local structure for image. Second, one can use various graph distance computation for more accurate graph matching.

Acknowledgment. The authors would like to thank the anonymous referees for their constructive comments which have helped improve the paper. The research is supported by the National Natural Science Foundation of China (Nos. 61502003, 71501002, 61472002 and 61671018), Natural Science Foundation of Anhui Province (No. 1608085QF133).

References

1. Penatti, O.A.B., Valle, E., da S. Torres, R.: Encoding spatial arrangement of visual words. In: San Martin, C., Kim, S.-W. (eds.) CIARP 2011. LNCS, vol. 7042, pp. 240–247. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25085-9_28
2. Boureau, Y.L., Bach, F., Lecun, Y., Ponce, J.: Learning mid-level features for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 26, pp. 2559–2566 (2010)
3. Sivic, J., Russell, B.C., Efros, A.A., et al.: Discovering objects and their location in images. In: Tenth IEEE International Conference on Computer Vision, vol. 1, pp. 370–377 (2005)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, no. (1/2), pp. 2169–2178 (2006)
6. Sadeghi, F., Tappen, M.F.: Latent pyramidal regions for recognizing scenes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 228–241. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_17
7. Yang, J., Yu, K., Gong, Y., et al.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1794–1801 (2009)
8. Wang, J., Yang, J., Yu, K., et al.: Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 119, pp. 3360–3367 (2010)
9. Cao, Y., Wang, C., Li, Z., et al.: Spatial-bag-of-features. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 238, pp. 3352–3359 (2010)
10. Silva, F.B., Werneck, R.D.O., Goldenstein, S., et al.: Graph-based bag-of-words for classification. In: International Conference on Pattern Recognition, vol. 74, pp. 266–285 (2018)

11. Bunke, H., Allermann, G.: Inexact graph matching for structural pattern recognition. In: International Conference on Pattern Recognition Letters, vol. 1, no. 4, pp. 245–253 (1983)
12. Jouli, S., Mili, I., Tabbone, S.: Attributed graph matching using local descriptions. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2009. LNCS, vol. 5807, pp. 89–99. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04697-1_9
13. Silva, F.B., Tabbone, S., Torres, R.D.S.: Bog: a new approach for graph matching. In: International Conference on Pattern Recognition, pp. 82–87 (2014)
14. Silva, F.B., Goldenstein, S., Tabbone, S., et al.: Image classification based on bag of visual graphs. In: IEEE International Conference on Image Processing, vol. 2010, pp. 4312–4316 (2014)
15. Hashimoto, M., Cesar, R.M.: Object detection by keygraph classification. In: Torsello, A., Escolano, F., Brun, L. (eds.) GbRPR 2009. LNCS, vol. 5534, pp. 223–232. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02124-4_23
16. Zhou, F., Torre, F.D.L.: Deformable graph matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 9, pp. 2922–2929 (2013)
17. Lee, J., Cho, M., Lee, K.M.: Hyper-graph matching via reweighted random walks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 42, pp. 1633–1640 (2011)
18. Zhang, L., Hong, R., Gao, Y.: Image categorization by learning a propagated graphlet path. *IEEE Trans. Neural Netw. Learn. Syst.* **27**(3), 674–685 (2016)
19. Wu, J., Pan, S., Zhu, X., et al.: Multi-graph-view learning for complicated object classification. In: International Conference on Artificial Intelligence, pp. 3953–3959. AAAI Press (2015)
20. Mousavi, S.F., Safayani, M., Mirzaei, A., et al.: Hierarchical graph embedding in vector space by graph pyramid. In: International Conference on Pattern Recognition, vol. 61, pp. 245–254 (2017)
21. Nikolentzos, G., Meladianos, P., Vazirgiannis, M.: Matching node embeddings for graph similarity. In: Proceedings of the 31st Conference on Artificial Intelligence, AAAI, pp. 2429–2435 (2017)
22. Dong, W., Moses, C., Li, K.: Efficient k-nearest neighbor graph construction for generic similarity measures. In: International Conference on World Wide Web, pp. 577–586. ACM (2011)
23. Song, J., Gao, L., Zou, F.: Deep and fast: deep learning hashing with semi-supervised graph construction. *Image Vis. Comput.* **55**, 101–108 (2016)
24. Harchaoui, Z., Bach, F.: Image classification with segmentation graph kernels. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 76, pp. 1–8 (2007)
25. Shu, X., Tang, J., Qi, G.J.: Image classification with tailored fine-grained dictionaries. *IEEE Trans. Circuits Syst. Video Technol.* **28**(2), 454–467 (2018)
26. Grauman, K., Darrell, T.: The pyramid match kernels: discriminative classification with sets of image features. In: Proceedings of the Tenth IEEE International Conference on Computer Vision, vol. 2, pp. 1458–1465 (2005)
27. Li, F.F., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: IEEE CVPR Workshop on Generative-Model Based Vision, vol. 106, no. 1, pp. 59–70 (2007)

28. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. In: California Institute of Technology (2007)
29. Li, F.F., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 524–531 (2005)
30. Li, P., Ma, J.: What is happening in a still picture? In: International Conference on Pattern Recognition, pp. 32–36 (2011)
31. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**(3), 1–27 (2011)
32. Penatti, O.A.B., Silva, F.B., Valle, E., et al.: Visual word spatial arrangement for image retrieval and classification. In: International Conference on Pattern Recognition, vol. 47, no. 2, pp. 705–720 (2014)

Author Index

- Batur, Aliya 99
Bi, Ning 351
- Cai, Zhengqi 88
Cao, Zhiguo 15
Chen, Jingying 245
Chen, Rui 257
Chen, Weifu 491
Chen, Wen-Sheng 292
Chen, Yajun 465
Cheng, Danni 27
- Deng, Huiqi 491
Deng, Liping 292
Du, Heran 183
Du, Jiang 268
- Fan, Dongjie 351
Fan, Jianping 478
Feng, Guocan 491
Feng, Zhanxiang 137
Feng, Zhen-Hua 388
- Gao, Xinbo 478
Gao, Zhi 452
Guo, Zhenhua 207
- Han, Yuehui 52, 63, 74
Hao, You 280
Hao, Zhanjun 52
He, Guoliang 503
He, Jinrong 503
He, JunHua 375
He, Zhihai 403
Hu, Duo 610
Hu, Haifeng 219
Hu, Wenhui 550
Hua, Wen-Wen 233
Huang, Rong 403
Huang, Rui 137, 539
- Huang, Yaoxiong 41, 112
Huang, Yibin 440
Huang, Zengxi 388
Huangfu, Zhenzhen 539
- Ji, Jian 514
Jia, Yunde 452
Jian, Meng 363
Jiang, Hui 527
Jiang, Mingyan 127
Jin, Lianwen 41, 112
Jing, Liping 328
- Ke, Kangyin 573
Kittler, Josef 388
Kong, Heng 597
Kong, Lizhi 162
- Lai, Jianhuang 137, 150
Lai, Zhihui 597
Law, Jarvan 375
Li, Fake 27
Li, Hua 280
Li, Mei 550
Li, Qi 280
Li, Sen 586
Li, Sengping 3
Li, Wei-Hong 27
Li, Xiang 27
Li, Xingxing 328
Li, Xutao 3
Li, Zhenjiang 52, 74
Liang, Lingyu 41, 173
Lin, Guangfeng 465
Lin, Qingxiang 41
Lin, Yufei 586
Ling, Yu 503
Liu, Haiying 363
Liu, Hao 403
Liu, Huafeng 328

- Liu, Leyuan 245
 Liu, Manfei 112
 Liu, Suolan 162
 Liu, Yiguang 388
 Lu, Bibo 539
 Lu, Chan 27
 Luo, Bin 610
- Ma, Andy J. 491
 Ma, Qianli 586
 Ma, Wei 150
 Mamat, Patigul 99
 Mei, Chengjiu 527
 Mi, Zhenxing 415
 Miao, Qiguang 514
 Mo, Dongmei 597
 Mo, Hanlin 280
- Pan, Binbin 292
 Pan, Heng 503
- Qi, Xinyuan 15
- Ren, Chuan-Xian 233
 Ren, Silin 503
 Rong, Erhu 375
- Shen, Chunxu 440
 Shen, Qi 491
 Sheng, Xiaoliang 304
 Shi, Guangming 268
 Shi, Huabei 440
 Shi, Rui 514
 Shi, Wenhui 127
 Shi, Zhongchao 550
 Song, Jinjie 195
 Su, Yingcheng 207
 Sun, Tao 440
- Tan, Jun 351
 Tang, Jin 610
 Tao, Liang 195
 Tao, Wenbing 415
- Ubul, Kurban 99
- Wan, Xiaopei 207
 Wang, Chenye 268
 Wang, Fangzhao 427
 Wang, Hongyuan 162
 Wang, Huabin 195
 Wang, Jian 15
 Wang, Peng 550
 Wang, Qi 195
 Wang, Qigang 550
 Wang, Weilan 52, 63, 74, 88
 Wang, Xiaojuan 52, 63, 74
 Wang, Yiqun 52, 63, 74
 Wu, Jian 440
 Wu, Lifang 363
 Wu, Yongbo 219
 Wu, Yuwei 452
- Xiang, Hui 550
 Xiao, Yang 15
 Xie, Xiaohua 137, 150
 Xie, Xuemei 268
 Xie, Zecheng 112
 Xu, Can 245
 Xu, Jiamiao 427
 Xu, Lijun 340
 Xu, Qin 610
 Xu, Ruyi 245
 Xu, Tonglin 304
- Yan, Xiang 195
 Yang, Bowen 363
 Yang, Changshui 257
 Yang, Meng 560, 573
 You, Xinge 427
 Yousefnezhad, Muhammad 304
 Yu, Mingjing 183
 Yuan, Ning 304
 Yuen, Pong C. 491
- Zhang, Chao 15
 Zhang, Chunhui 514
 Zhang, Dai 363
 Zhang, Daoqiang 304
 Zhang, He 280
 Zhang, Ji 478
 Zhang, Peng 427
 Zhang, Shuaitao 112
 Zhang, Tiancheng 560
 Zhang, Xinglin 173
 Zhao, Fan 465
 Zhao, Hua 27
 Zheng, Huicheng 183
 Zheng, Qi 427

Zheng, Wei-Shi 27
Zheng, Yu 478
Zhong, Qin 560
Zhou, Rigui 527
Zhou, Wenjie 99
Zhou, Yijia 340
Zhou, Zhengdong 550

Zhu, Bing 316
Zhu, Changming 527
Zhu, Haiqing 351
Zhu, Junyong 150
Zhu, Yali 99
Zhuo, Hankz Hankui 375
Zou, Wanxin 3