Hongbin Zha · Xilin Chen
Liang Wang · Qiguang Miao (Eds.)

Communications in Computer and Information Science 547

Computer Vision

CCF Chinese Conference, CCCV 2015
Xi'an, China, September 18–20, 2015
Proceedings, Part II

Part 2

中国计算机学会
CCF

Springer

Communications in Computer and Information Science 547

Commenced Publication in 2007
Founding and Former Series Editors:
Alfredo Cuzzocrea, Dominik Ślęzak, and Xiaokang Yang

More information about this series at http://www.springer.com/series/7899

Hongbin Zha · Xilin Chen
Liang Wang · Qiguang Miao (Eds.)

Computer Vision

CCF Chinese Conference, CCCV 2015
Xi'an, China, September 18–20, 2015
Proceedings, Part II

Springer

Editors

Hongbin Zha
Peking University
Beijing
China

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

Liang Wang
Chinese Academy of Sciences
Beijing
China

Qiguang Miao
Xidian University
Shaanxi
China

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-662-48569-9 ISBN 978-3-662-48570-5 (eBook)
DOI 10.1007/978-3-662-48570-5

Library of Congress Control Number: 2015951339

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media
(www.springer.com)

Preface

Welcome to the proceedings of the First Chinese Conference on Computer Vision (CCCV 2015) held in Xi'an!

The CCCV conference is organized by the Computer Vision Task Forces of China Computer Federation (CCF), and is an important event for the computer vision community in China. In recent years, computer vision has increasingly become an enabling technology for many important and often mission-critical applications such as video surveillance and human–machine interface. The aim of CCCV is to provide a forum for scientific exchanges between computer vision researchers in China, and will be held every two years from 2015.

CCCV 2015 invited regular submissions for presentations and as well internationally renowned researchers to give keynote speeches. We received 176 full submissions, each of which was reviewed by at least two reviewers selected from the Program Committee and by other qualified researchers. Based on the reviewers' reports, 89 papers were finally accepted for presentation at the conference, yielding an acceptance rate of 51 %.

We are grateful to the keynote speakers, Prof. Fatih Porikli from the Australian National University, Prof. Maja Pantic from Imperial College London, Prof. Xiaoou Tang from the Chinese University of Hong Kong, and Prof. Demetri Terzopoulos from UCLA, USA.

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

We hope you find the proceedings of CCCV 2015 enjoyable.

September 2015

Tieniu Tan
Xinbo Gao
Hongbin Zha
Xilin Chen
Liang Wang
Qiguang Miao

Organization

CCCV 2015 (CCF Chinese Conference on Computer Vision 2015) was sponsored by CCF, co-sponsored by the CCF Task Force on Computer Vision, and hosted by the School of Computer Science and Technology, the School of Electronic Engineering and the State Key Laboratory of Integrated Service Networks of Xidian University.

General Chairs

Tieniu Tan	Institute of Automation, Chinese Academy of Sciences
Xinbo Gao	Xidian University

Program Chairs

Hongbin Zha	Peking University
Xilin Chen	Institute of Computing Technology, Chinese Academy of Science
Liang Wang	Institute of Automation, Chinese Academy of Sciences
Qiguang Miao	Xidian University

Organizing Chairs

Quan Wang	Xidian University
Jianhuang Lai	Sun Yat-sen University
Tao Wang	IQIYI Inc.
Wen Lu	Xidian University

Finance Chair

Yining Quan	Xidian University

Publicity Chairs

Deyu Meng	Xi’an Jiaotong University
Peiyi Shen	Xidian University

Publications Chair

Cheng Deng	Xidian University

Program Committee

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

Xiaochun Cao	Chinese Academy of Sciences
Shengyong Chen	Zhejiang University of Technology
Songcan Chen	Nanjing University of Aeronautics and Astronautics
Xilin Chen	Chinese Academy of Science
Xiaowu Chen	Beihang University
Xiaoming Deng	Chinese Academy of Science
Jing Dong	Chinese Academy of Science
Fuqing Duan	Beijing Normal University
Lichun Fang	Shanghai University
Jufu Feng	Peking University
Xin Geng	Southeast University
Ping Guo	Beijing Normal University
Huiguang He	Chinese Academy of Science
Ran He	Chinese Academy of Science
Zhiqiang Hou	Air Force Engineering University
Baiming Huang	University of Wisconsin-Madison
Hua Huang	Beijing Institute of Technology
Kaiqi Huang	Chinese Academy of Science
Qingming Huang	University of Chinese Academy of Sciences
Yongzhen Huang	Chinese Academy of Science
Rongrong Ji	Xiamen University
Yunde Jia	Beijing Institute of Technology
Xiaoyuan Jing	Wuhan University
Xiangwei Kong	Dalian University of Technology
Jianhuang Lai	Sun Yat-sen University
Chunming Li	University of Electronic Science and Technology of China
Hua Li	Chinese Academy of Science
Jian Li	National University of Defense Technology
Lingling Li	Zhengzhou Institute of Aeronautical Industry Management
Peihua Li	Dalian University of Technology
Shiying Li	Hunan University
Yongjie Li	University of Electronic Science and Technology of China
Zongmin Li	China University of Petroleum
Liang Lin	Sun Yat-sen University
Zhouchen Lin	Peking University
Hong Liu	Peking University
Huafeng Liu	Zhejiang University
Huaping Liu	Tsinghua University
Li Liu	National University of Defense Technology
Qingshan Liu	Nanjing University of Information Science and Technology
Yiguang Liu	Sichuan University
Guoliang Lu	Shandong University
Huchuan Lu	Dalian University of Technology
Bin Luo	Anhui University
Ke Lv	University of Chinese Academy of Sciences
Qiguang Miao	Xidian University

Pinle Qin	North University of China
Qiuqi Ruan	Beijing Jiaotong University
Nong Sang	Huazhong University of Science and Technology
Shiguang Shan	Chinese Academy of Science
Linlin Shen	Shenzhen University
Peiyi Shen	Xidian University
Wei Shen	Shanghai University
Fei Su	Beijing University of Posts and Telecommunications
Dongmei Sun	Beijing Jiaotong University
Zhengxing Sun	Nanjing University
Taizhe Tan	Guangdong University of Technology
Tieniu Tan	Chinese Academy of Science
Xiaoyang Tan	Nanjing University of Aeronautics and Astronautics
Jianliang Tang	Shenzhen University
Jinhui Tang	Nanjing University of Science and Technology
Yandong Tang	Chinese Academy of Science
Zengfu Wang	Chinese Academy of Science
Hanzi Wang	Xiamen University
Hao Wang	Philips Research Institute of China
Liang Wang	Chinese Academy of Science
Qing Wang	Northwestern Polytechnical University
Ruiping Wang	Chinese Academy of Science
Shiquan Wang	Philips Research Institute of China
Tao Wang	IQIYI Inc.
Qian Wang	Shanghai Jiaotong University
Wei Wang	Chinese Academy of Science
Yuanquan Wang	Tianjin University of Technology
Yunhong Wang	Beijing University of Aeronautics and Astronautics
Shikui Wei	Beijing Jiaotong University
Wei Wei	Xi’an University of Technology
Gongjian Wen	National University of Defense Technology
Jianxin Wu	Nanjing University
Guisong Xia	Wuhan University
Shiming Xiang	Chinese Academy of Science
Jinbiao Xu	Agilent Technologies(America)
Zenglin Xu	University of Electronic Science and Technology of China
Chenhui Yang	Xiamen University
Guosheng Yang	Minzu University of China
Jie Yang	Shanghai Jiaotong University
Jinfeng Yang	Civil Aviation University of China
Jingyu Yang	Nanjing University of Science and Technology
Jufeng Yang	Nankai University
Xianghua Ying	Peking University
Xingang You	Beijing Institute of Electronics Technology and Application
Xinge You	Huazhong University of Science and Technology
Jian Yu	Beijing Jiaotong University

Jun Yu	Hanzhou Electronic Science and Technology University
Shiqi Yu	Shenzhen University
Xiaoyi Yu	Peking University
Kai Yu	Baidu Inc.
Zhiwen Yu	South China University of Technology
Hongshan Zha	Peking University
Guofeng Zhang	Zhejiang University
Honggang Zhang	Beijing University of Posts and Telecommunications
Junping Zhang	Fudan University
Yan Zhang	Nanjing University
Yanning Zhang	Northwestern Polytechnical University
Yimin Zhang	Intel China Research Center
Yongdong Zhang	Chinese Academy of Science
Zhang Zhang	Chinese Academy of Science
Zhaoxiang Zhang	Beijing University of Aeronautics and Astronautics
Yao Zhao	Beijing Jiaotong University
Yuqian Zhao	Central South University
Weishi Zheng	Sun Yat-sen University
Hanning Zhou	Zhigu Inc.
Zongtan Zhou	National University of Defense Technology
Zhenfeng Zhu	Beijing Jiaotong University
Hui Zeng	University of Science and Technology Beijing
Wangmeng Zuo	Harbin Institute of Technology

Organizers

Organized by

中国计算机学会

CCF

China Computer Federation, China

中国计算机学会 · 计算机视觉专业组

Computer Vision Task Forces, China Computer Federation

CCF Task Force on Computer Vision

Hosted by

西安電子科技大學

XIDIAN UNIVERSITY

Xidian University

Sponsoring Institutions

NVIDIA.

NVIDIA Corporation.

iQIYI 爱奇艺
悦 享 品 质

IQIYI Inc.

公安部第三研究所
The Third Research Institute of Ministry of Public Security

The Third Research Institute of Ministry of Public Security

海康威视®
HIKVISION

Hangzhou Hikvision Digital Technology Co., Ltd.

北京文安
VIONVISION

Vion Technology Inc.

HUAWEI

Huawei Technologies Co., Ltd.

The Right Answer in Data Center
PARATERA 并行®

Contents – Part II

Contents – Part I

Directional Segmentation Based on Shear Transform and Shape Features for Road Centerlines Extraction from High Resolution Images

Ruyi Liu, Qiguang Miao(✉), Jianfeng Song, and Qing Xue

School of Computer and Technology, Xidian University, Xi'an 710071, China
{ruyi198901210121,qgmiao}@126.com, {33760709,309094071}@qq.com

Abstract. Accurate extraction of road networks from high resolution remote sensing images is a problem not satisfactorily solved by existing approaches, especially when the color of road is close to that of background. This paper studies a new road networks extraction from remote sensing images based on the shear transform, the directional segmentation, shape features and a skeletonization algorithm. The proposed method includes the following steps. Firstly, we combine shear transform with directional segmentation to get road regions. Secondly, road shape features filtering are used to extract reliable road segments. Finally, the road centerlines are extracted by a skeletonization algorithm. Road networks are then generated by post-processing. Experimental results show that this method is efficient in road centerlines extraction from remote sensing images.

Keywords: Road centerlines extraction · Shear transform · Directional segmentation · Shape features

1 Introduction

Roads are the backbone and essential modes of transportation, providing many different supports for human civilization. Road extraction plays a very important role in vehicle navigation system, urban planning, disaster management system and traffic management system. Due to the improvement of image resolution, the image has all sorts of detailed information to obtain very good reflection, but these details characteristics are interference for the extraction of road. Also, high-resolution satellite images have serious shadows, particularly in urban areas, which have an impact on road extraction, so the road extraction from high-resolution satellite images has a great scientific significance. In recent years, various road extraction algorithms have been proposed. A variety of road detection techniques[1]include knowledge based methods[2], mathematical morphology[3],[4], snakes[5]–[6], classification[7]–[10], differential geometry[11], region competition [12], active testing [13], perceptual grouping [14], and dynamic programming [15]. Mena [16] and Fortier et al. [17] provide extensive surveys of the literature on road extraction technique.

Although the above methods show a good performance in road extraction, it is difficult to get a satisfactory result [18], and we need do some further research.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 1–9, 2015.
DOI: 10.1007/978-3-662-48570-5_1

Chaudhuri et al. [19] proposed a semi-automatic road detection method. In this method there were only a small set of directions to be used to detect the road segment. Thus some road segments are not detected. In order to solve the above problem, we would like to develop an efficient algorithm. This paper proposes a method based on shear transform, the directional segmentation, shape features and a skeletonization algorithm.

The organization of this paper is as follows: In Section 2, the new method is described. In Section 3, we compare the experimental result with a semi-automatic road detection method, and show that the method we proposed is efficient. Finally, the concluding remarks are given in Section 4.

2 Methodology

In this section, we first combine shear transform with directional segmentation to get road regions. Then, the road regions are refined by using shape features. Finally, we extract the road centerlines and make post-processing to get complete networks. Fig. 1 gives a summary of the proposed method.

2.1 Image Preprocessing and Directional Segmentation Based On Shear Transform

Image preprocessing consists of image enhancement and gray processing. In our implementation, Grey level transformation is used to adjust the contrast of details in an image. When we convert a color image to a gray image, some useful information will be lost. Therefore we apply grey level transformation on the color images.

Roads are mostly narrow and linear in the image. When we consider a small set of directions in the process of segmentation, some road segments are lost and we can't extract the whole complete road regions. However, looking for pixels in all directions can be computationally complex. Therefore the shear transform [20-25] is introduced here.

Let $W_{s,k}$ denote the multi-direction shear operation, where $s=0$ or 1, $k \in [-2^{(ndir)}, 2^{(ndir)}], k \in Z$, Z denotes the set of integers and $ndir$ is the direction parameter $(ndir \in N)$. $s=0$ means the shear transform is applied in the horizontal direction, and $s=1$ in the vertical direction. By taking the multi-direction shear transform on the image $f(x,y)$, the sheared images $f'_{s,k}(x,y)$ would be obtained, the number of which is $2\times(2^{(ndir+1)}+1)$, as shown in equation (1)

$$f'_{s,k}(x,y) = f(x,y) * W_{s,k} \tag{1}$$

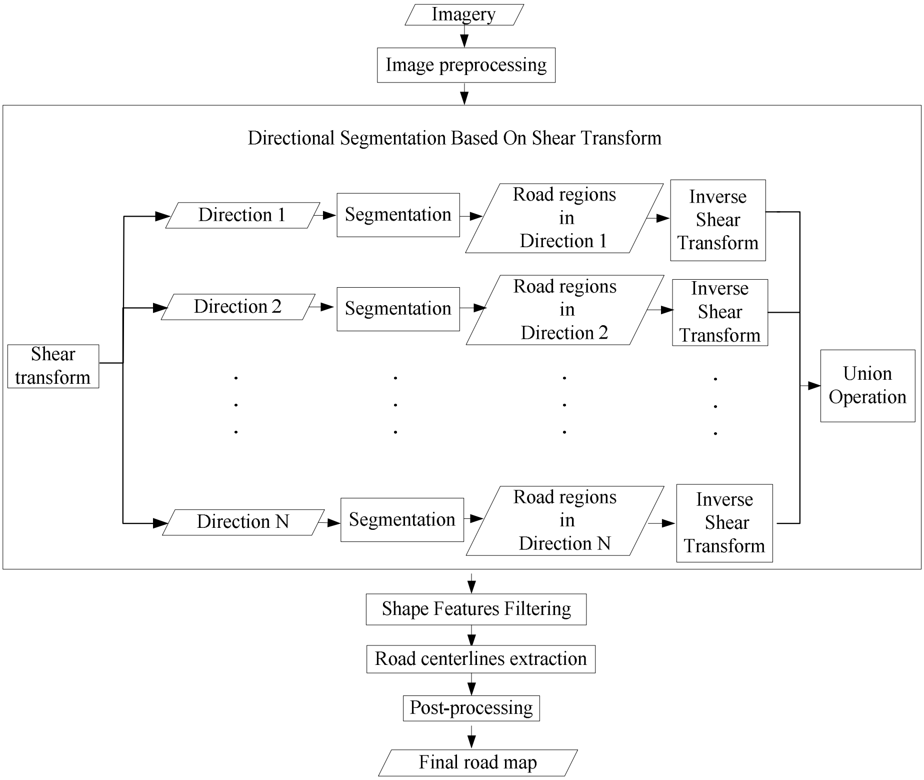


Fig. 1. Framework of the proposed method

In our implementation, shear matrix is $s_0 = \begin{bmatrix} 1 & 0 \\ \frac{\lfloor k \rfloor}{2^{ndir}} & 1 \end{bmatrix}$ or $s_1 = \begin{bmatrix} 1 & \frac{\lfloor k \rfloor}{2^{ndir}} \\ 0 & 1 \end{bmatrix}$.If $s = 0$ we take the shear transform according to the shear matrix s_0.

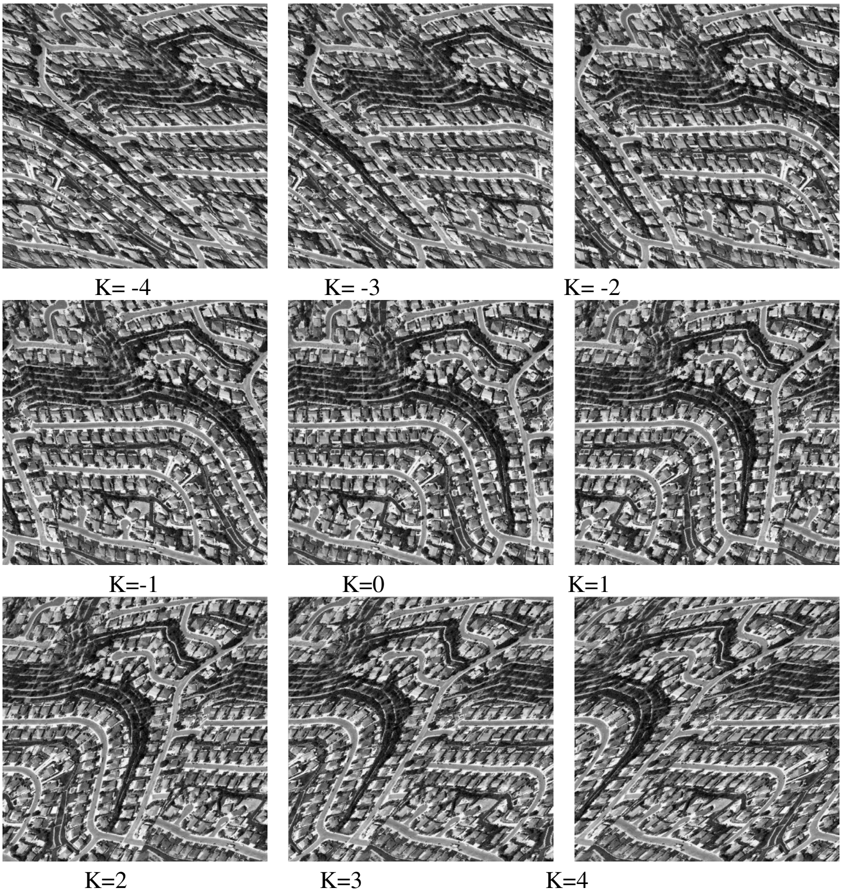
$$(x', y') = (x, y)\, s_0 = (x, y) \begin{bmatrix} 1 & 0 \\ \frac{\lfloor k \rfloor}{2^{ndir}} & 1 \end{bmatrix} = (x + y \times \frac{\lfloor k \rfloor}{2^{ndir}}, y) \tag{2}$$

$$f'_{0,k}(x', y') = f(x, y) \tag{3}$$

where (x', y') is the coordinate of a pixel in the sheared image and (x, y) is the coordinate of a pixel in the original image, but the values of all the pixels remain unchanged during this process. When $s = 1$, the shear transform is performed in the vertical direction according to s_1, and the procedure is similar.

A variety of sheared images which are obtained by the shear transform are shown in Fig. 2. The shear transform is applied to the image after preprocessing. Here we set $s = 0$ and $ndir = 2$, so the number of the sheared images is nine. It can be seen that the shear transform changes the entire neighborhood centered at a point and more directions for elongated regions will be estimated.

The algorithm to efficiently separate road segments in the sheared images is based on supervised directional homogeneity using a modified metric [19]. Two 5×5 road seed templates from the sheared images are chosen as the representatives. Considering only eight directions provides a good balance between computational efficiency and accurate extraction of road pixels, so we consider eight directions in every sheared image in this paper.


K= -4 K= -3 K= -2

K=-1 K=0 K=1

K=2 K=3 K=4

Fig. 2. Results of the shear transform

The inverse shear transform [20-25] is used for the images containing road segments in various directions, and then all of the images are fused into the final image by the union operation.

2.2 Shape Features Filtering

Some road regions identified after segmentation have some misclassified roads. To further discard misclassified roads, road shape features [26] are explored. These features are measured by the following: 1) area; 2) linear feature index (LFI).

1. Area: A road is commonly a continuous feature with a relatively larger area than other man-made features. Hence, segments with small area values can be regarded as noisy and should be removed.
2. LFI : Roads are generally narrower and longer than other artificial objects. This characteristic is described by LFI which is computed as follows.

- Using connected component analysis to divide pixels into connected components. The component is then converted to a rectangle which satisfies

$$LW = n_p \tag{4}$$

where L is the length of the new rectangle, W is the width of the new rectangle, n_p is the area of the road segment(also known as pixel number).

- LFI can be calculated by

$$LFI = \frac{L}{W} = \frac{L}{n_p / L} = \frac{L^2}{n_p} \tag{5}$$

In terms of roads' characteristics, they should have large values of LFI, so regions with small values of LFI can be regarded as nonlinear features and will be removed.

2.3 Road Centerlines Extraction and Post-processing

In our implementation, we have used a well-known skeletonization algorithm proposed by Uitert and Bitter [27]. We then remove two kinds of unwanted linear segments to improve the linear representation of the roads in the image. The branches which are connected to the main road skeleton at one end and are not connected at the other end in the thinned image need to be eliminated, so we remove the branches whose length is less than the minimum threshold. The long linear structures that are isolated need also to be eliminated. We make these post-processing based on the filtering method proposed in [19].

3 Experiments and Discussions

In this section, to test the performance of the proposed method we experiment with high-resolution aerial and satellite images and achieve satisfactory results. The proposed method is also compared with other's method to show the advantages and disadvantages of the proposed method. Fig. 3(a) shows a high resolution aerial image with spatial dimension of 300×300 pixels. The gray image is shown in Fig. 3(b). Fig. 3(c) shows the image after segmentation. As can be seen that road segments are mostly extracted. However, there are some small areas and noise in the segmented regions, which is convenient to remove in the following steps. Fig. 3(d) shows the image after shape features filtering. The final road networks result is shown in Fig. 3(e). In this experiment, all of the parameters and thresholds are set by the trial and error method in order to get a better result.

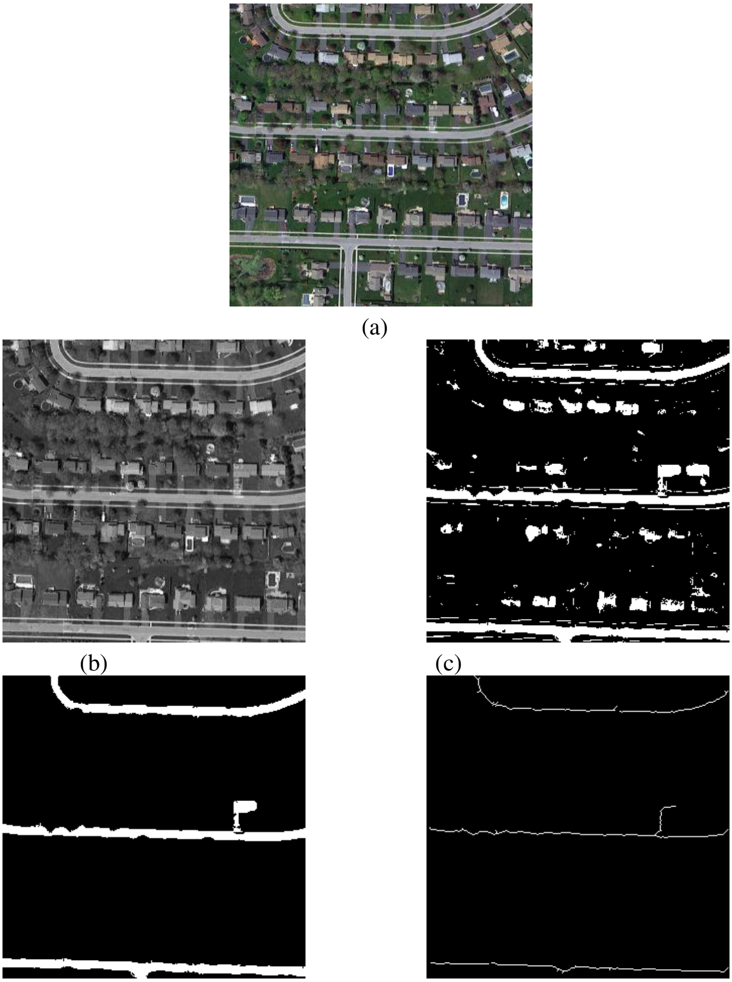
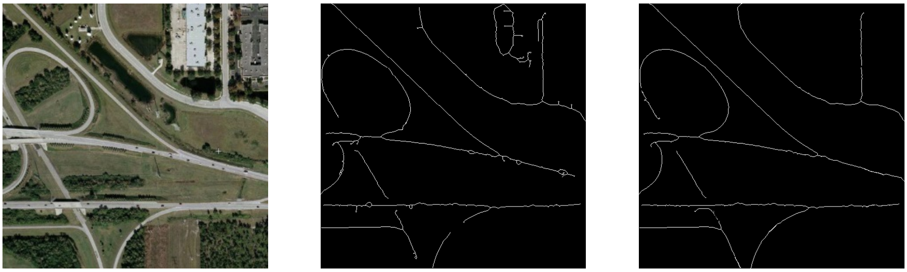


Fig. 3. Result with aerial image. (a) Original image, (b) Gray image, (c) Segmented image, (d) Image after shape features filtering, (e) Road networks.



(a) (b) (c)

Fig. 4. Comparison with Chaudhuri's method [13] on QuickBird image. (a)Input image, (b) Result of the Chaudhuri's method [13], (c) Result of the proposed method.

In order to prove the validity of the proposed algorithm, we compare our result with the method proposed by Chaudhuri et al. [19].As can be seen from the images shown in Fig. 4, the roads extracted by our algorithm are more complete and smooth than that generated by Chaudhuri's method. Clearly, most of the main road centerlines are extracted by the proposed algorithm.

We can conclude that the proposed method can perform more effectively than the other method. The computational complexity of the proposed method is higher than that of Chaudhuri's method, and it varies directly with the number of directions of the sheared images.

4 Conclusion

In this paper, we have presented a method to extract road centerlines from high-resolution images accurately and smoothly. The main steps in our algorithm are: image preprocessing, shear transform, road segmentation, shape features filling, road centerlines extraction and networks generation. In this method, the shear transform is introduced to overcome the limitation of directions for road segments. The experimental results have been evaluated to demonstrate the high accuracy of the proposed method. The extracted road centerlines retain smooth. However, the proposed method still has several flaws which we need to do some more research later. The main limitation of the proposed method is that all of the parameters and thresholds are set by the trial and error method, namely, be determined manually.

Acknowledgment. The work was jointly supported by the National Natural Science Foundations of China under grant No. 61472302, 61272280, 41271447, and 61272195; The Program for New Century Excellent Talents in University under grant No. NCET-12-0919; The Fundamental Research Funds for the Central Universities under grant No. K5051203020, K5051303016, K5051303018, BDY081422, and K50513100006; Natural Science Foundation of Shaanxi Province, under grant No.2014JM8310 ; The Creative Project of the Science and Technology State of Xi'an under grant No.CXY1341 (6) and CXY1440(1).

References

1. Das, S., Mirnalinee, T.T., Varghese, K.: Use of salient features for the design of a multistage framework to extract roads from high resolution multispectral satellite images. IEEE Trans. Geosci. Remote Sens. **49**(10), 3906–3931 (2011)
2. Trinder, J., Wang, Y.: Knowledge-based road interpretation in aerial images. Int. Arch. Photogramm. Remote Sens. **32**(4), 635–640 (1998)
3. Zhu, C., Shi, W., Pesaresi, M., Liu, L., Chen, X., King, B.: The recognition of road network from high-resolution satellite remotely sensed data using image morphological characteristics. Int. J. Remote Sens. **26**(24), 5493–5508 (2005)
4. Katartzis, A., Sahli, H., Pizurica, V., Cornelis, J.: A model-based approach to the automatic extraction of linear features from airborne images. IEEE Trans. Geosci. Remote Sens. **39**(9), 2073–2079 (2001)
5. Laptev, I., Mayer, H., Lindeberg, T., Eckstein, W., Steger, C., Baumgartner, A.: Automatic extraction of roads from aerial images based on scale space and snakes. Mach. Vis. Appl. **12**(1), 23–31 (2000)
6. Baumgartner, A., Hinz, S., Wiedemann, C.: Efficient methods and interfaces for road tracking. Proc. Int. Soc. Photogramm. Remote Sens., 28–31 (2002)
7. Yager, N., Sowmya, A.: Support vector machines for road extraction from remotely sensed images. In: Petkov, N., Westenberg, M.A. (eds.) CAIP 2003. LNCS, vol. 2756, pp. 285–292. Springer, Heidelberg (2003)
8. Song, M., Civco, D.: Road extraction using SVM and image segmentation. Photogramm. Eng. Remote Sens. **70**(12), 1365–1371 (2004)
9. Zhang, Q., Couloigner, I.: Benefit of the angular texture signature for the separation of parking lots and roads on high resolution multispectral imagery. Pattern Recog. Lett. **27**(9), 937–946 (2006)
10. Tuncer, O.: Fully automatic road network extraction from satellite images. In: Proc. Recent Adv. Space Technol., pp. 708–714, June 2007
11. Steger, C.: An unbiased detector of curvilinear structures. IEEE Trans. Pattern Anal. Mach. Intell. **20**(2), 113–125 (1998)
12. Amo, M., Martinez, F., Torre, M.: Road extraction from aerial images using a region competition algorithm. IEEE Trans. Image Process. **15**(5), 1192–1201 (2006)
13. Geman, D., Jedynak, B.: An active testing model for tracking roads in satellite images. IEEE Trans. Pattern Anal. Mach. Intell. **18**(1), 1–14 (1996)
14. Gamba, P., Dell'Acqua, F., Lisini, G.: Improving urban road extraction in high-resolution images exploiting directional filtering, perceptual grouping, and simple topological concepts. IEEE Geosci. Remote Sens. Lett. **3**(3), 387–391 (2006)
15. Barzohar, M., Cooper, D.B.: Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. IEEE Trans. Pattern Anal. Mach. Intell. **18**(7), 707–721 (1996)
16. Mena, J.B.: State of the art on automatic road extraction for GIS update: A novel classification. Pattern Recognit. Lett. **24**(16), 3037–3058 (2003)
17. Fortier, M.F.A., Ziou, D., Armenakis, C., Wang, S.: Survey of work on road extraction in aerial and satellite images, Univ. Sherbrooke, Sherbrooke, QC, Canada, Tech. Rep. 241, vol. 24, no. 16, pp. 3037–3058 (2003)
18. Shi, W., Miao, Z., Johan, D.: An integrated method for urban main road centerline extraction from optical remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. **52**(6), 3359–3372 (2013)

19. Chaudhuri, K., Kushwaha, N.K., Samal, A.: Semi-automated road detection from high resolution satellite image by directional morphological enhancement and segmentation techniques. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **5**(5), 1538–1544 (2012)
20. Easley, G., Labate, D., Wang, Q.-L.: Sparse directional image representations using the discrete shearlet transform. Appl. Comput. Harmon. Anal. **25**(1), 25–46 (2008)
21. Jiao, L.-C.H., Hou, B., Wang, S., Liu, F.: Image Multiscale Geometric Analysis: Theory and Application. Xian Electronic Science, Technology Univ. Press, Xian (2008)
22. Miao, Q.G., Shi, C., Xu, P.F., Yang, M., Shi, Y.B.: A novel algorithm of image fusion using shearlets. Opt. Commun. **284**(6), 1540–1547 (2011)
23. Lim, W.-Q.: The discrete shearlet transform: A new directional transform and compactly supported shearlet frames. IEEE Trans. Image Process. **19**(5), 1166–1180 (2010)
24. Xu, P.F., Miao, Q.G., Shi, C., Zhang, J.Y., Li, W.S.: An edge detection algorithm based on the multi-direction shear transform. J. Visual Commun. Image Represent. **23**(5), 827–833 (2012)
25. Miao, Q.G., Xu, P.F., Liu, T.G., Yang, Y., Zhang, J.Y., Li, W.S.: Linear feature separation from topographic maps using energy density and shear transform. IEEE Trans. Image Process. **4**, 1548–1558 (2013)
26. Miao, Z., Shi, W., Zhang, H., Wang, X.: Road centerline extraction from high-resolution imagery based on shape features and multivariate adaptive regression splines. IEEE Geosci. Remote Sens. Lett. **10**(3), 583–587 (2013)
27. van Uitert, R., Bitter, I.: Subvoxel precise skeletons of volumetric data base on fast marching methods. Med. Phys. **34**(2), 627–638 (2007)

Sparse Representation with Regularization Term for Face Recognition

Jian Ji(✉), Huafeng Ji, and Mengqi Bai

Department of Computer Science & Technology, Xidian University, Xi'an 710071, China
jji@xidian.edu.cn

Abstract. In recent years there has been a growing interest in the study of sparse representation based classification (SRC) which has obtained great success in face recognition. However, SRC is overly dependent on the size of training samples while overlooking the correlation information that is critical to the real-world face recognition problems. Besides, some method considers the correlation information but overlooks the discriminating ability of sparsity. In this paper, we propose a new method called trace norm sparse representation based classification (TSRC) which introduces a regularization term in the SRC model and considers both sparsity and correlation. The TSRC method can benefits from both ℓ_1-norm and ℓ_2-norm, which is flexible and can obtain satisfactory results. Experimental results on 2 face databases clearly show that the proposed TSRC method outperforms many state-of-the-art face recognition methods.

Keywords: Face recognition · Trace norm · Sparse representation based classification · Sparsity and correlation

1 Introduction

Face recognition, as one of the most successful applications of image analysis and understanding, has recently received significant attention and adequate development, especially during the past decades.Nevertheless, due to the different interference of different conditions cause corruption and errors of different degrees, for example, various facial expression, pose and illumination conditions, the face image processing effect is not so ideal. Furthermore, when the feature space is not sufficient sample database and high dimension, the existence of these problems will meet more challenges in face recognition.

The conventional method of face recognition(sparse PCA [1], 2DPCA [2]) selects a limited subset or model from training samples, instead of the entire training set for image detection or signal classification and representation. So when the train sample space is small, the performances are not very good. These methods based on feature space, such as NN (Nearest Neighbor) and the support vector machine (SVM), when the image between different classes is very similar, will have a low recognition effect.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 10–20, 2015.
DOI: 10.1007/978-3-662-48570-5_2

Therefore, face recognition methods based sparse representation classification emerge as the times requirement [3-5].Sparse representation method is based on the hypothesis that the testing images are approximation in a low dimension subspace which is obtained by the training samples, and then can be represented by a small number of training samples. Sparse representation based classification [3] (SRC) seek sparse representation of a query image in an over-complete dictionary, and then obtain recognition performance through comparing the minimal sparse error to identity the query image class. SRC can be seen as a generalization of NN and NFS, but it can get better recognition performance [3]. SRC overemphasize sparsity of data while ignoring the correlation between the dictionaries, which often results in lack of information. Thus, when the training samples are highly correlated, SRC will produce unstable results. Some of the literature has shown the importance of correlation structure [6-8]. Zhang proposed the CRC method which made full use of the correlation data for face recognition and used the ℓ_2-norm model [9].

Only when the training sample is large, the SRC method shows a good recognition performance and it can't use the correlation data to obtain useful information. While the CRC method can get a good result by the correlation, but when the correlations of training samples are limited, it may not perform well. We propose a new face recognition method called the trace norm sparse representation classification (TSRC) which applies the trace norm as the regularization term into the dictionary. The trace norm can benefit from ℓ_1-norm and ℓ_2-norm, in other words it can take advantage of sparsity as well as data correlation. After we proved the feasibility of the regularization term and answered the minimization problem of the trace norm, we draw a lot of experiments in different face image databases, and compare the face recognition performance between different methods including SRC [3], SVM [10], NN, NFS [11] and LSRC [6].

2 Backgrounds of Sparse Representation

2.1 Generalized Sparse Representation

Denote the data set of training samples labeled with the i-th class as $A_i = [v_{i,1}, v_{i,2}, ..., v_{i,n_i}]$, Any new test sample y from the same class can be linearly expressed as:

$$y = \alpha_{i,1} v_{i,1} + \alpha_{i,2} v_{i,2} + \alpha_{i,n_i} v_{i,n_i} \tag{1}$$

where $\alpha_{\mathrm{i},j}$ are some scalars.

We define a new matrix A for the entire training set as the concatenation of the n training samples of all k object classes:

$$A = [A_1, A_2, ..., A_k] = [v_{1,1}, v_{1,2}, ..., v_{k,n_k}] \tag{2}$$

Then, the linear representation of y can be rewritten in terms of all training samples as:

$$y = Ax_0 \tag{3}$$

where $x_0 = [0,...,0,\alpha_{i,1},\alpha_{i,2},...,\alpha_{i,n_i},0,...,0]^T$ is a coefficient vector whose entries are zero except those associated with the i-th class.

The purpose of sparse representation is to estimate the main information of the test sample using non-zero coefficient as little as possible. In other words, we need to find the x_0 which has less non-zero coefficient and can be a good estimation of y with A.

2.2 Classification Based on Sparse Representation (SRC)

J. Wright et al. introduced the Sparse Representation based Classification (SRC) method which had applied to face recognition and pattern recognition [3]. The model is as follows,

$$(\ell^0): \hat{x}_1 = \arg\ \min ||x||_0 \quad s.t \quad Ax = y \tag{4}$$

$||x||_0$ is the ℓ_0-norm, defined as the number of non-zero entries in the vector x. The problem of ℓ_0-norm can be proved is NP hard, even if making approximately calculation is also very difficult [12]. Some paper reveals that if the solution x_0 sought is sparse enough, the ℓ_0-minimization problem (4) is equal to the solution to the following ℓ_1-minimization problem:

$$(\ell^1): \hat{x}_1 = \arg\ \min ||x||_1 \quad s.t \quad Ax = y \tag{5}$$

$||x||_1$ is the ℓ_1-norm, defined as $||x||_1 = \sum_i |x_i|$, that is the sum of the absolute values of all the entries. The model (3) can be explicitly modified to the flowing form account for small possible dense noise:

$$y = Ax_0 + z \tag{6}$$

where $||z||_2 < \varepsilon$, z is noise item , the sparse solution x_0 can be obtained by the following ℓ_1-minimization problem:

$$\hat{x}_1 = \arg\min ||x||_1 \quad s.t \quad ||Ax - y||_2 \le \varepsilon \tag{7}$$

$||x||_2$ is the ℓ_2-norm, defined as $||x||_2 = \sqrt{\sum_i x_i^2}$. Based on [13], when $||x_0||_0 < \rho m$ and $||z||_2 \le \varepsilon$, we can learn that there are constants ρ and ζ satisfied with

$$||\hat{x}_1 - x_0|| \le \zeta\varepsilon \tag{8}$$

So we can use the formulas (8) to examine the computed $\hat{x}_1$.

For x , $\delta_i(x)$ is a new vector whose only nonzero entries are the entries in x that are associated with class i . So we can approximate the given test sample y as $\hat{y}_i = A\delta_i(\hat{x}_1)$. We then classify y based on these approximations by assigning it to the object class that minimizes the residual between y and $\hat{y}_i$:

$$\min_i \ r_i(y) \| y - A\delta_i(\hat{x}_1) \|_2 \tag{9}$$

3 Sparse Representation with Regularization Term for Face Recognition

3.1 Why Did We Introduce the Regularization Term?

The SRC algorithm is under the assumption that image and training images are in very good agreement. The results show that when there are enough training samples which can cover all changes, y can be correctly expressed. Therefore, SRC may not obtain satisfactory results at the case where y are not aligned and dictionary contains a small amount of sample. At the same time, due to the sparsity, when the samples are highly correlated, SRC may have the problem of unstable. If the object and the query image are similar, the SRC method tends to choose a random object instead of choosing them all. This means that, SRC does not capture the relevant structure of the dictionary that plays crucial role in the face recognition [14].

Good performances of SRC comes from the collaborative representation of y is on all training samples [9]. CRC can make good use of the advantages of data correlation [9]. Therefore, in the CRC, the images is represented by an over complete dictionary which use ℓ_2 -norm rather than use ℓ_1 -norm to control coding vector. The object function of ℓ_2 -norm is as follows.

$$(\ell^2) : \hat{x}_2 = \arg\min ||x||_2 \quad s.t \quad Ax = y \tag{10}$$

Considering the noise problem, the equation can be changed into

$$(\ell^2) : \hat{x}_2 = \arg\min ||x||_2 \quad s.t \quad ||Ax\text{-}y||_2 \le \varepsilon \tag{11}$$

ℓ^2 made CRC obtain a stable results through the use of a more dense vector, but when the training samples were not highly correlated, the CRC would not be able to get good results.

Only when the training sample is large, the SRC method can show a good recognition performance and it can’t use the correlation data to obtain useful information. While the CRC method can get a good result by the correlation, but when the correlations of training samples are limited, it may not perform well. the trace norm classification method based on sparse representation (TSRC) overcome the disadvantages of SRC and CRC. For fully considering the sparsity and correlation, we combine structure of matrix A and coding coefficient x and introduce the trace norm inspired by [15]. Giving a matrix M, diag(x) indicates converting the matrix M into a vector in which the i-th entry is M_{ii} located in the diagonal of M. The ℓ_1-norm of M is defined as $||M||_1 = \sum_{ij} |m_{ij}|$ and the trace $||M||_*$ is regarded as the sum of all the singular values of the matrix M. Thus we get following linear representation model:

$$\min ||ADiag(x)||_* \quad s.t. \quad y = Ax \tag{12}$$

where, $||ADiag(x)||_*$ is a regularization term defined as $\Omega_A(x)$. With the regularization term, we will no longer ignore sparsity or correlation.

3.2 Extreme Value of the Regularization Term

There are two extreme cases for the trace norm regularization term which can be discussed as follows.

1). When the columns of matrix A are not related and A is an orthogonal matrix, that is $A^T A = I$. And then we can get

$$\begin{aligned} &|| ADiag(x) ||_* = Tr[(ADiag(x))^T (ADiag(x))]^{1/2} \\ &= Tr[(Diag(x))^T (Diag(x))]^{1/2} \\ &=|| x ||_1 \end{aligned} \tag{13}$$

Thereby, $\Omega_A(x)$ is equivalent to ℓ_1-norm. So we can change (12) into

$$\min ||x||_1 \qquad s.t. \quad y = Ax \tag{14}$$

2). In the case where the images of different subjects look similar to a_1, that is $A = a_1 E^T$ and $A^T A = EE^T$ (E is a vector of size n whose elements are one), we can express $\Omega_A(x)$ as:

$$\Omega_A(x) = ||a_1 x^T||_* = ||a_1||_2 ||x||_2 = ||x||_2 \tag{15}$$

Then (12) can be changed into

$$\min ||x||_2 \quad s.t. \quad y = Ax \tag{16}$$

Generally, the images in the dictionary are neither too independent of each other nor look the same. Our model is able to obtain the correlation structure of the training samples. So we can easily know that

$$||x||_2 \le \Omega_A(x) \le ||x||_1 \tag{17}$$

It is show that x obtained by $\Omega_A(x)$ is more sparse than the one obtain by ℓ_2, but is less sparse than the one obtained by ℓ_1. This means that we can take advantage of ℓ_1 and ℓ_2 .

3.3 Our Novel Sparse Representation Model (TSRC)

If the noise obeys Gauss distribution, the objective function can be turned into

$$\min||ADiag(x)||_* \quad s.t. \quad ||y\text{-}Ax||_2 \le \varepsilon \tag{18}$$

Instead if the occlusion follows the Laplace distribution, we consider the following problem:

$$\min||ADiag(x)||_* \quad s.t. \quad ||y\text{-}Ax||_1 \le \varepsilon \tag{19}$$

Problem (19) is more robust to occlusion and variations than problem (18) [16]. Problem (19) can be changed into the following problem:

$$\min||y - Ax||_1 + \lambda||ADiag(x)||_* \tag{20}$$

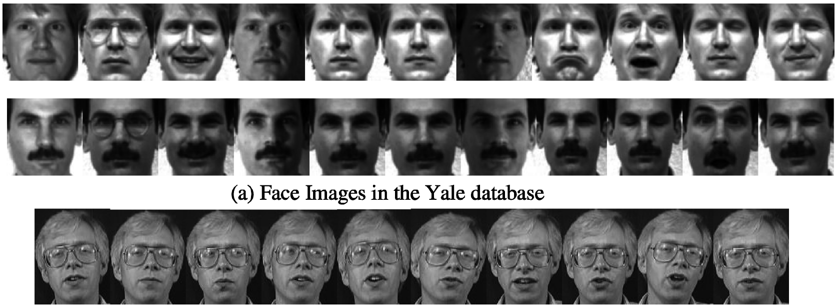
where $\lambda > 0$ is the regularization parameter. We adopt Alternating Direction Met hod (ADM) [17] to solve problem (20).
The final novel sparse representation model is expressed as follows:

$$\min||y - Ax||_1 + \lambda||ADiag(x)||_* + \eta||x||_2 \tag{21}$$

η is an optional parameter whose value depends on the distribution type of noise. The type of noise determines the value of the parameter η. If the noise obeys Gauss distribution, $\eta = 1$ and if the occlusion follows the Laplace distribution, we set the η to 0.

4 Experiment

This part we will demonstrate the recognition effect of TSRC in 2 face image databases in Fig. 1, respectively, to select two groups of images totally having 22 images on Yale face database (each group has 11 images) and one group ORL face database images totally having 10 images.



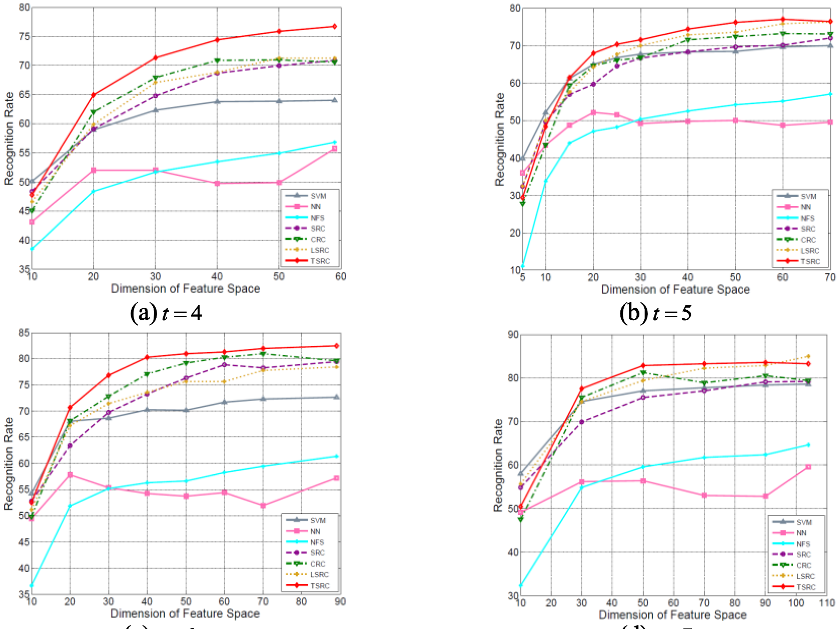
(a) Face Images in the Yale database

(b) Face Images in the ORL database

Fig. 1. Yale and ORL face databases

4.1 Experiments on Yale Face Database

The Yale face database is composed of 15 volunteers, 11 gray face images per person for a total of 165 different images of 32×32 pixel. These images with different expressions are obtained in different illumination. As we can see in Fig. 1(a) expressions of each individual are different. In this experiment, we randomly selected $t(=4,5,6,7)$ images of each individual as a training sample and the rest of the images as test images. For different t with different dimension of the feature space (in 10 increment), we record the average precision in Fig. 2 and the maximum average accuracy as well as the standard deviation corresponding to the value in the Table 1.



(a) $t = 4$ (b) $t = 5$

(c) $t = 6$ (d) $t = 7$

Fig. 2. In the Yale face database, recognition rate of various method under different t and dimension of feature space

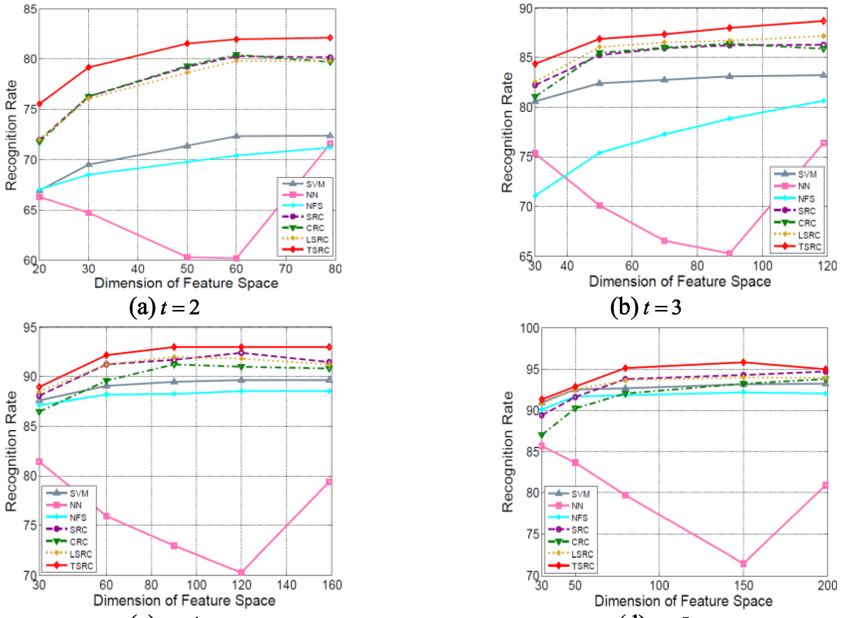
Table 1. Maximum average accuracy and standard deviation of different methods

Algorithms	$t = 4$	$t = 5$	$t = 6$	$t = 7$
SVM[10]	64.00±2.57(59)	69.89±5.48(70)	72.67±3.28(89)	78.50±6.73(104)
NN	55.71±4.65(59)	52.11±4.30(20)	57.87±4.92(20)	59.50±3.69(104)
NFS[11]	56.76±5.30(59)	57.00±4.74(70)	61.33±5.96(89)	64.50±5.21(104)
SRC[3]	70.86±4.56(59)	72.00±4.02(70)	79.47±3.68(89)	79.17±3.17(104)
CRC[9]	70.95±4.67(50)	73.11±4.79(60)	80.93±3.93(89)	81.17±3.93(50)
LSRC[6]	71.24±2.49(50)	76.22±3.93(70)	78.40±3.86(70)	85.00±5.56(104)
TSRC	74.56±4.71(59)	77.00±3.98(60)	81.31±2.67(89)	83.17±4.89(104)

We can see from the Fig. 2 and Table 1 that TSRC has better recognition rate than other methods at different t values with the change of feature space. When the t value is small (as $t = 4, 5$), the maximum value of recognition is generally less than the larger value of t (as $t = 6, 7$), which just corresponds to the case of SRC. Because TSRC can keep the dictionary correlation and sparsity, so when the t value is small, the good identification rate can manifest its advantages. Smaller t values mean that the number of training samples is small, TSRC can still get changes of the query images by choosing training samples with sufficient correlation so as to get better recognition rate. When t increasing, TSRC, SRC, CRC and LSRC have better recognition rate and when $t = 4$ and $t = 6$, SRC, CRC, LSRC have relatively similar curve. When $t = 5$ and $t = 7$, LSRC is superior to SRC and CRC's performance, this is due to the partial information of dictionary considered by LSRC. However, the methods are not good as TSRC proposed in this paper in addition to the extremely individual points because TSRC can accurately grasp the structural information of dictionary and enable it to better adapt to the query image.

4.2 Experiments on ORL Face Database

The ORL face database consists of 40 individuals and each individual has 10 gray images including different illumination, facial and detail changes, as Fig. 1 (b) shown. We select training samples of number t and the remaining are the query images in this experiment.The experimental results are shown in Fig. 3 and Table 2.



(a) $t = 2$

(b) $t = 3$

(c) $t = 4$

(d) $t = 5$

Fig. 3. In the ORL face database, recognition rate of various method under different t and dimension of feature space.

Table 2. Maximum average accuracy and standard deviation of different methods as well as the feature space dimension when the maximal accuracy is obtained.

Algorithms	$t = 2$	$t = 3$	$t = 4$	$t = 5$
SVM	73.38±4.00(79)	83.25±1.94(119)	89.67±1.90(120)	93.20±1.60(199)
NN	71.59±3.23(79)	79.36±2.37(119)	81.46±1.89(30)	85.70±2.37(30)
NFS	71.19±3.48(79)	80.64±1.70(119)	88.54±2.03(120)	92.20±1.72(120)
SRC	80.28±2.52(60)	86.29±1.58(119)	92.37±0.88(120)	94.70±1.44(199)
CRC	80.44±2.41(60)	86.39±2.07(90)	91.21±1.72(90)	93.75±2.12(199)
LSRC	79.81±2.46(60)	87.14±1.87(119)	92.00±1.22(90)	94.00±2.12(150)
TSRC	81.36±2.58(79)	88.52±1.96(119)	93.00±1.87(159)	95.69±2.09(150)

We can see from Fig. 3 and Table 2 that recognition rate of TSRC is higher than other methods. Comparison of several figures, the most obvious is that the recognition rate of NN method is the lowest and it is unstable. When the number of training is low ($t = 2,3$), SRC, CRC and LSRC have similar recognition rate. As can be seen from table 2, the recognition rate of TSRC shows a rise tendency as a whole, but the feature space dimension when the maximal accuracy is obtained is not rising. For example, when $t = 5$, the feature space dimension is 150 when the maximal accuracy is obtained which is small than other t, that is to say when the training samples are sufficient, the blindly increase of feature space dimension may not increase the recognition rate. However, TSRC have better recognition performance than other methods.

4.3 Summary

In general, SRC, CRC and LSRC have stable recognition rate in most cases. When the training sample size was small, CRC showed better recognition performance because it considered the correlation of data while sparsity showed lower effect. LSRC can get good recognition rate than SRC because the local information and sparsity of sample date were taken into account. But when the local information is not sufficient, TSRC can consider correlation and sparsity of the sample, so in most cases TSRC can get better recognition results than other methods. Therefore, the experiments proved that TSRC is a good method for face recognition.

5 Conclusions

We do have proved that the TSRC method have better recognition performance than other face recognition methods, such as NN, SVM, CRC and LSRC. It can benefit from sparsity and correlation. Specifically, TSRC can obtain comparable results to SRC when the dictionary is with low correlation, and performs as well as CRC when the data are with high correlation. TSRC can make good use of correlation between the query images and training samples, and then it can obtain relatively much information. LSRC only considers the limited information of few local samples in a small number of training samples. Experimental results on face database clearly show that the proposed TSRC method outperforms many state-of-the-art face recognition methods.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61101248) and the Fundamental Research Funds for the Central Universities of China (No. JB140315).

References

1. d'Aspremont, A., Ghaoui, L.E., Jordan, M., Lanckriet, G.: A Direct Formulation of Sparse PCA Using Semidefinite Programming. SIAM Rev. **49**, 434–448 (2007)
2. Zhang, D.Q., Zhou, Z.H.: Two-directional two-dimensional PCA for efficient face representation and recognition. Neurocomputing **69**, 224–331 (2005)
3. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. on Pattern Analysis and Machine Intelligence **31**(2), 210–227 (2009)
4. Pillai, J.K., Patel, V.M., Chellappa, R.: Sparsity inspired selection and recognition of iris images. In: Proc. IEEE Third International Conference on Biometrics: Theory, Applications and Systems, pp. 1–6 (2009)
5. Hang, X., Wu, F.-X.: Sparse representation for classification of tumors using gene expression data. Journal of Biomedicine and Biotechnology (2009). doi:10.1155/2009/403689
6. 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 (2010)
7. Lu, J., Tan, Y., Wang, G., Gao, Y.: Image-to-set face recognition using locality repulsion projections and sparse reconstruction-based similarity measure. IEEE Trans. on Circuits and Systems for Video Technology **23**(6), 1070–1080 (2013)
8. Wang, M., Gao, Y., Lu, K., Rui, Y.: View-based discriminative probabilistic modeling for 3d object retrieval and recognition. IEEE Trans. on Image Processing **22**(4), 1395–1407 (2013)
9. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: IEEE International Conference on Computer Vision, pp. 471–478 (2011)
10. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(27), 1–27 (2011)
11. Shan, S., Gao, W., Zhao, D.: Face identification from a single example image based on face-specific subspace (fss). In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 2125–2128 (2002)
12. Amaldi, E., Kann, V.: On the Approximability of Minimizing Nonzero Variables or Unsatisfied Relations in Linear Systems. Theoretical Computer Science **209**, 237–260 (1998)

13. Donoho, D.: For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ_2 -norm -Norm Near Solution Approximates the Sparest Solution. Comm. Pure and Applied Math. **59**(10), 907–934 (2006)
14. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2), 301–320 (2005)
15. Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., Yan, S.: Robust and efficient subspace segmentation via least squares regression. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 347–360. Springer, Heidelberg (2012)
16. Wright, J., Ganesh, A., Yang, A., et al.: Sparsity and Robustness in face recognition (2011). arXiv:1111.1014
17. Lin, Z., Chen, M., Wu, L., Ma, Y.: The augmented lagrange multiplier method for exact recovery of a corrupted low-rank matrices, UIUC Technical Report UILU-ENG-09-2215, Tech. Rep. (2009)

Salient Region Detection by Region Color Contrast and Connectivity Prior

Mei-Huan Chen[1], Yan Dou[1,2(✉)], and Shi-Hui Zhang[1,2]

[1] College of Information Science and Engineering, Yanshan University,
Qinhuangdao 066004, Hebei, China
douyan@ysu.edu.cn
[2] Hebei Province Key Laboratory of Computer Virtual Technology
and System Integration, Qinhuangdao 066004, China

Abstract. The visual salient regions detection is one of the fundamental problems in computer vision, so saliency estimation has become a valuable tool in image processing. In this paper, we propose a novel method to realize the calculation of saliency, using color contrast and connectivity prior (called CCP for short). There are three cues integrated to obtain high-quality map, including contrast, spatial distribution and high-level prior. We evaluate our approach on three standard benchmark datasets with other state-of-the-art approaches, the results show that the proposed method has the higher precision and recall, the final maps are more closed to the ground truth.

Keywords: Salient region detection · Contrast · Spatial distribution · Connectivity prior

1 Introduction

People can easily focus attention on the important parts or the salient regions in a scene. Salient object detection has been studied in physiology, neural systems, psychology and computer vision for a long time. It is motivated by the importance of saliency detection in applications such as image segmentation [1,2], object recognition [3], image retrieval [4,5], adaptive compression of images [6] and so on.

Nowadays, existing visual attention approaches mainly include two kinds, namely, fast, bottom-up, data driven saliency geodesic; and slower, top-down, task driven saliency geodesic. The former is popular, and some models are very successful in salient region detection [7-15], which almost focus on the contrast of low level image features. While the contrast information often helps produce good significant results, it always generates high values for the area which is not salient, especially for the regions with low contrast from the surrounding or connected heavily with the object.

This research is partly supported by the National Science Foundation, China (no. 61379065). Hebei Province Science and Technology Support Program, China (no. 13211801D). The Doctoral Foundation of Yanshan University (no. B540).

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 21–30, 2015.
DOI: 10.1007/978-3-662-48570-5_3

Inspired by the above discussion, in this paper, we propose a new algorithm of the contrast, i.e., region color contrast. Our method is based on the assumption that the salient region is not only assigned high contrast values, but also located near the image center, besides, warm colors are more pronounced.

The contrast may work well for low-level saliency calculation, but they are neither abundant nor high-accurate if used alone. So several approaches [16-26] exploited some high-level prior knowledge to help get more sufficient values. [19-21] made use of the boundary prior (or called connectivity prior). These articles simply thought the image boundary as background. But in fact, this is arbitrary and fragile when the region only slightly contacts the boundary. So this paper renewedly presents the connectivity prior, which states that the background is more heavily connected to the image boundary than the foreground.

Our first main contribution is the new region color contrast algorithm, which combine color difference with the center and warm superiority. Besides, using the color spatial distribution offsets the drawback produced by the region color contrast. The second contribution is the novel definition of connectivity prior, and integrated with other cues for saliency computation to obtain better results.

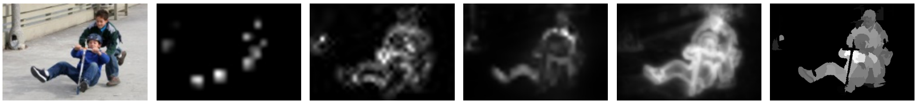


Fig. 1. The development process of the saliency detection, from left to right: source image, IT [7], MZ [8], HC [14], CA [5], CCP

2 Related Work

In this part, we briefly introduce the related work on image saliency detection. Many early works solved the problem of saliency detection with contrast feature. [27] thought that the saliency is decided by the center-surround contrast of low-level features. [7] defined image salient using a Difference of Gaussians approach. [14] obtained the saliency value from local contrast based on image segmentation. [15] combined local and global contrast with linearity method. From then, many approaches [5,8,28] were presented by researchers which united the local, region and global contrast. Most methods analyzed contrast feature from the difference of color or luminance perspective [14,17,29-31]. [29] used color contrast and distribution based on region level with adaptive saliency refinement approach to gain the results. [30] measured the contrast cues from multiple scales of image structure with better results. [31] introduced salient region detection adopting color uniqueness with focusness and objectness.

Later on, some prior knowledge are considered to enhance the effect, such as center prior [17,18], shape prior [23], background prior [19-21], context information [5,22,23], depth cues [24-26]. [17] believed the region more closed to the center of the image has the higher value. [23] put up shape prior which armed at extracting a closed contour covered the salient object, thus to protect the original feature of the goal. In [19], the article treated image boundary regions as background, and image patch's saliency was defined as the shortest-path distance to the image boundary. The context information was used in many approaches, [5] presented a detection algorithm based on the context of the dominant objects just as essential as the objects themselves. [26] proposed an object lying at a different depth level from the others will noticeably attract people's attention. Above high-level prior knowledge are always integrated into low-level features to make their methods possibly suitable for salient objects.

Many researches have been attempting to combine several different features or prior knowledge to pay their respective advantages. Among them, the most commonly used method is to simply unite the consequences from these features by using weighted averaging. And [32] took use of Conditional Random Fields, [23] utilized iterative Energy Minimization, [28] learned the weight by Support Vector Machine. Beyond that, there are a lot of novel methods, [16] fused the higher-level priors to a low-rank matrix recovery to compose salient maps. Also [12,13] turned to the frequency domain to solve the question.

In this paper, we propose a new model for the saliency region detection, firstly, we decompose an image into small regions, then obtain the region color contrast map with the new definition, take use of the spatial distribution to avoid the defects caused by contrast. Add connectivity prior to make the effect better. At last, the high quality map acquired by fusion of the above features.

3 Salient Region Detection by Region Color Contrast and Connectivity Prior

In the following we describe the details of our method, and we show the new definition of the region color contrast, color spatial distribution as well as the connectivity prior based on the SLIC segmentation.

3.1 The Region Obtained

For the image abstraction, we use the SLIC superpixels [33] to divide the image into perceptually uniform regions. The "soft" approach could segment an image, and keep them local, compact and edge aware. In the paper, N indicates the number of the divided regions. Superpixel result examples are shown in Fig.2 (b).

3.2 Region Color Contrast

Human usually pay more attention to those image regions that contrast strongly with their surroundings. The color difference (or called uniqueness), which is generally defined as the infrequence of a segment r_i given its color in CIELab compared to all other regions r_j. In this paper, we introduce the center prior $c(r_i)$ to bias the image center region with higher saliency value. The experience in [17] showed that almost people like to put the interested object to the center position in the picture. Besides, many psychologists found that warm colors such as red, yellow and orange are more pronounced, so we put warm colors superior $w(r_i)$ into region color contrast. The definition of the region color contrast $RCC(r_i)$ is following, and Fig.2 (c) gives the examples.

$$RCC(r_i) = w(r_i) \bullet c(r_i) \bullet \sum_{j=1}^{N} \log\left(1 + d\left(r_{ij}\right)\right) \tag{1}$$

$w(r_i)$ indicates the warm colors superior, we obtain a 2-D histogram distribution $H(F)$ in the CIELab apace from the labeled salient object, which similar to [16]. Analogously the background histogram distribution $H(B)$ is generated. For each certain color, the values from the two histograms are h_f and h_b. $w(r_i) = \exp\left(\left(h_f(\bar{c}) - h_b(\bar{c})\right)/\sigma_1^2\right)$, where the $\bar{c}$ indicates the mean color of the region r_i. We set $\sigma_1 = 0.25$ in our experience.

$c(r_i)$ indicates the center information, objects near the image center are more attractive to people. $c(r_i) = 1/\left(d(p^i, v)/\sigma_2^{\ 2}\right)$, where p^i is the center point of the region r_i, and υ is the center point of the image, $d(*)$ is the meaning of the Euclidean distance between the two points. The parameter σ_2 controls the final effect derived from the center information, we set $\sigma^2 = 3$ in our experience.

$d(r_{ij})$ indicates the difference of the region r_i and the surroundings in the CIELab. People easily perceive the colors in the CIELab space, because it makes up for the uneven distribution of the RGB color model. The region is more salient, if the value is larger than others. $d(r_{ij}) = \left\|\bar{c_i} - \bar{c_j}\right\|^2$, where $\bar{c_i}$, $\bar{c_j}$ are the average color value of the region r_i and r_j respectively.

3.3 Color Spatial Distribution

The experience shows the elements are more salient when they are located in a stationary image region rather than evenly distributed over the whole image. So it testify that color spatial distribution is so vital to the saliency detection. Conceptually, we define the color spatial distribution measure for a region r_i using the $CSD(r_i)$, the factor indicates a compact object should be deemed more salient than spatially dispersive elements, Fig.2 (d) gives the examples.

$$CSD(r_i) = \sum_{j=1}^{N} \log\left(1 + col\left(r_{ij}\right) \bullet \left\|p_i - p_j\right\|^2\right) \tag{2}$$

In equation (2), $col(r_{ij}) = \exp\left(\left(-\left\|\bar{c_i} - \bar{c_j}\right\|^2\right)/\sigma_3^2\right)$, describes the similarity of the color c_i and color c_j in the region r_i and r_j. p_i and p_j are the positions of region r_i and r_j severally. The parameter σ_3 controls the color sensitivity of the region distribution, we use $\sigma_3 = 19$ in all our experiment.

3.4 Connectivity Prior

Previous work mostly put sight on the salient object, but the problem remains challenging and huge behavioral discrepancies. Inspired by [19], we define a new method of the connectivity prior directed as background information.

Based on the segmentation of 3.1, we structure an undirected weighted graph by contiguous superpixels (r_i, r_j), and the weight is assigned by the Euclidean distance between their average colors in the CIELab color space. Then the shortest path of the region r_i and r_j on the graph is $\varphi_m(r_i, r_j) = \min_{r_i=r_1, r_2 \ldots, r_N, = r_j} \sum_{i=1}^{N} \varphi(r_i, r_{i+1})$.

Next we could obtain the all area related with the region r_i, namely $\eta_{area}(r_i)$, and the length along the boundary $\eta_{boud}(r_i)$, then the ratio $Rat(r_i)$ of the $\eta_{boud}(r_i)$ and $\eta_{area}(r_i)$ denotes the connecting strength of the region r_i and the boundary. Where $\kappa(*)$ is 1, if the superpixel is on the image boundary and 0 otherwise, Bon is the image boundary

$$Rat(r_i) = \frac{\eta_{bond}(r_i)}{\eta_{area}(r_i)} = \frac{\sum_{j=1}^{N} \exp(-\varphi_m(r_i, r_j)) \bullet \kappa(r_i \in Bon)}{\sum_{j=1}^{N} \exp(-\varphi_m(r_i, r_j))} \tag{3}$$

Lastly, define the connectivity prior value $CP(r_i)$ as a new weighting cue. It is close to 1 when $Rat(r_i)$ is small, and close to 0 when it is large. That is to say, if the $Rat(r_i)$ is large, indicating the certain area has a high possibility to be background, and has a lower possibility to be foreground. The definition is $CP(r_i) = \exp \frac{-Rat(r_i)^2}{2\sigma_4^2}$, empirically set $\sigma_4 = 1$, Fig.2 (e) gives the examples.

3.5 Combination of Features

In fact, we consider the all features as the independent portion, so if we use a reliable measure to combine them, the effect may be more remarkable. We start by normalizing the all features to the range [0...1]. Then we define our final saliency map *Sal* as: $Sal = CP * \exp(RCC * CSD)$. Fig.2 (f) gives the examples.

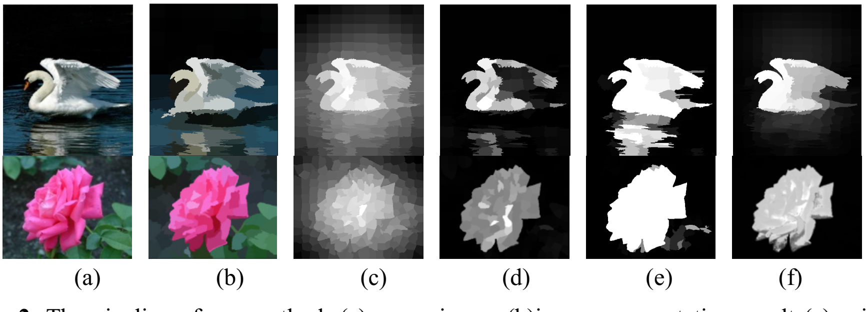


Fig. 2. The pipeline of our method: (a)source image (b)image segmentation result (c)region color contrast (d)color spatial distribution (e)connectivity prior (f)the final map

4 Experiments

We compare our method (CCP) with several state-of-the-art saliency approaches, such as IT[7], AC[11], SR[10], FT[12], RC[14], HC[14], CA[5], LR[16], SF[15], GC[34], UFO[31], HS[35], PD[36]. To evaluate these methods, the results obtained either from the original authors or running the authors' publicly available source code.

4.1 Data Sets and Evaluation Methods

Utilize three standard benchmark datasets: ASD-1000[12], SOD[37] and SED2[38]. Most images in the ASD database have single salient object and there are strong contrast between foreground and background, which used widely now. The SOD database includes 300 images based on the Berkeley segmentation dataset, which is more challenging with multiple objects in more complicated backgrounds. SED2 has 100 images containing two salient objects with different sizes and locations.

For performance evaluation, we adopt two methods as [12,14]. The first protocol is precision and recall rate. We compare every binary masks segmented by the threshold in the range [0...255] with ground truth, then obtain the curve. The second protocol is F-Measure, which is computed as: $F = \left(\left(\beta^2+1\right)P*R\right)/\left(\beta^2 P+R\right)$ (P is precision, R is recall), we set $\beta^2 = 0.3$, similar to [12,14], whose test images are segmented by the adaptive threshold.

4.2 Quantitative Comparison

Fig.3 (a,b) shows the precision-recall curves of the above approaches on the ASD-1000 dataset. The average precision, recall and F-measure using adaptive threshold segmentation is shown in Fig.3 (c). From the figures, we can clearly find that our method acquires the higher precision, recall and F-measures than the others. Besides, Fig.4 (a,b) demonstrates the precision-recall curves on the SOD and SED2 datasets. The experience manifests proposed method could detect salient regions keeping a high accuracy.

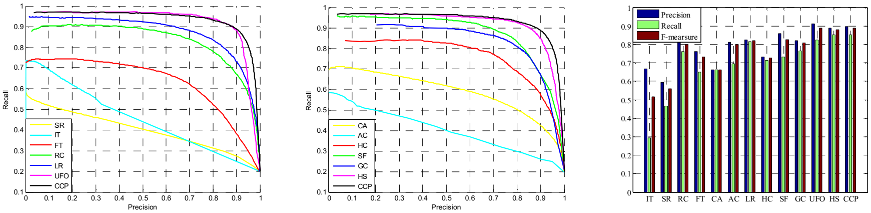
4.3 Qualitative Comparison

We show visual comparison of some approaches in Fig.5. Note that the test images of the top, middle and bottom two rows are from ASD-1000, SOD-300 and SED2-100, respectively. It is worth pointing out the results produced by our method are more closed to ground truth even in the cluttered background. For example, in the third and fourth rows, our proposed method can reasonably highlight the salient regions and suppress the background.

4.4 Subsidiary Details

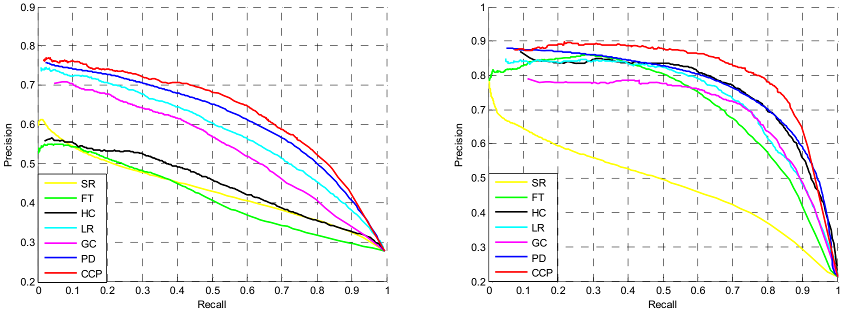
In the above content, we described the theory and experience of the new method. There are two details will be discussed here. The first is the influence of the center prior (center for short) and warm color prior (warm for short) to the region color contrast. Then we will use the precision-recall curve (Fig.6) on ASD-1000 dataset to indicate the fact, "coldif" indicates the color difference for convenience.

The second detail is the combination mode of the three cues in the part 3.5, namely, RCC (region color contrast), CSD (color spatial distribution) and CP (connectivity prior). We also use the *PR* curve (Fig.7) to evaluate the several combinations of our algorithm on the ASD-1000.



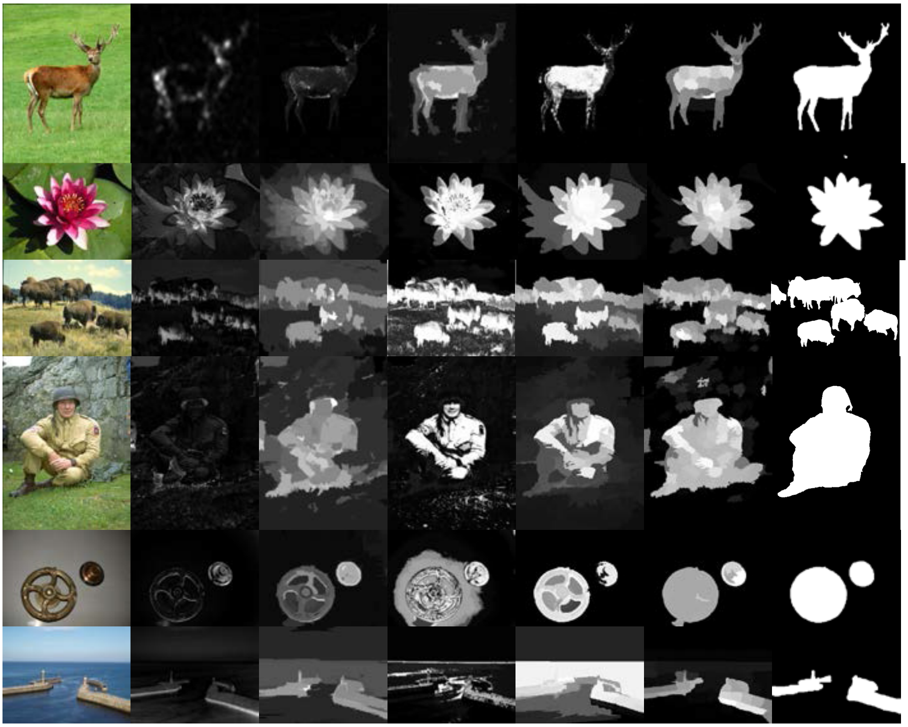
(a) (b) (c)

Fig. 3. The evaluations for our method with other approaches on the ASD-1000 dataset: (a) Precision-recall curves for some methods; (b) Precision-recall curves for the other methods; (c) Average precision, recall and F-measure



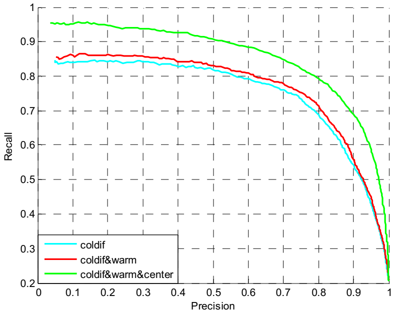
(a) (b)

Fig. 4. The evaluations for our method with other approaches on the SOD and SED2 dataset: (a) Precision-recall curves on the BSD dataset; (b) Precision-recall curves on the SED2 dataset;



(a)input image (b)FT (c) RC (f) GC (e) HS (f) CCP (g)ground truth

Fig. 5. Visual comparison of existing approaches with our method



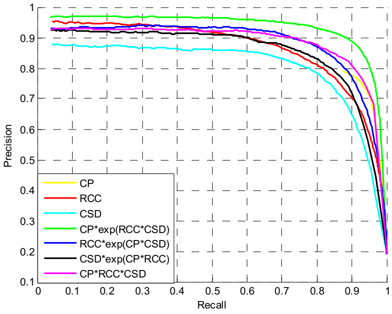


Fig. 6. The importance of the two cues for the region color contrast

Fig. 7. Evaluation of several integrations of our method on the ASD-1000 dataset

5 Conclusions

In this paper, we present a saliency detection based on contrast and boundary prior. While contrast has been used by other works, we define it from two sides, region color contrast and color spatial distribution. Besides, we take the center prior and warm color superiority into account. For the boundary prior estimation, we propose a novel method, which weigh the connectivity strength of region and boundary. More importantly, we combine the three cues in an effective way that leads to the outstanding performance when compared with the state-of-the-arts on the public dataset. But this method only considers single scale, which may be lead to the salient results inaccuracy. For future work, we plan to exploit multi-scale image abstraction to make the saliency detection better applied to the actual.

References

1. Han, J., Ngan, K., Li, M., Zhang, H.: Unsupervised Extraction of Visual Attention Objects in Color Images. IEEE Trans. Circuits Syst. Video Technol. **16**(1), 141–145 (2006)
2. Ko, B., Nam, J.: Object-of-interest Image Segmentation Based on Human Attention and Semantic Region Clustering. J. Opt. Soc. Am. A Opt. Image Sci. Vis. **23**(10), 2462–2470 (2006)
3. Rutishauser, U., Walther, D., Koch, C., Perona. P.: Is Bottom-Up Attention Useful for Object Recognition? In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 37–44 (2004)
4. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: Internet Image Montage. ACM Trans. Graph **28**(5), 1–10 (2009)
5. Goferman, S., Manor, L., Tal, A.: Context-Aware Saliency Detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 1915–1926 (2012)
6. Christopoulos, C., Skodras, A., Ebrahimi, T.: The JPEG2000 Still Image Coding System: an Overview. IEEE Trans. Consumer Elec. **46**(4), 1103–1127 (2002)
7. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
8. Ma, Y.F., Zhang, H.J.: Contrast-based image attention analysis by using fuzzy growing. In: ACM Multimedia, pp. 374–381 (2003)
9. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Annual Conf. Neural Information Processing Systems, Canada, pp. 545–552 (2006)
10. Hou, X.D., Zhang, L.: Saliency detection: a spectral residual approach. In: IEEE Conf. Computer Vision Pattern Recognition, USA, pp. 1–8 (2007)
11. Achanta, R., Estrada, F.J., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008)
12. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 1597–1604 (2009)
13. Li, J., Levine, M.D., An, X., He, H.: Saliency detection based on frequency and spatial domain analysis. In: Computer Vision-BMVC (2011)
14. Cheng, M., Zhang, G., Mitra, N., Huang, X., Hu, S.: Global contrast based salient region detection. In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 409–416 (2011)
15. Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung, A.: Saliency filters: contrast based filtering for salient region detection. In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 733–740 (2012)
16. Wu, Y., Shen, X.: A unified approach to salient object detection via low rank matrix recovery. In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 853–860 (2012)

17. Fu, K., Gong, C., Yang, J.: Saliency object detection via color contrast and color distribution. In: IEEE Conf. Asian Conference on Computer Vision, Korea, pp. 111–122 (2012)
18. Kannan, R., Ghinea, G., Swaminathan, S.: Salient region detection using patch level and region level image abstractions. Signal Processing Letters, IEEE **22**(6), 686–690 (2015)
19. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: European Conf. Computer Vision, Italy, pp. 29–42 (2012)
20. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: a discriminative regional feature integration approach. In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 2083–2090 (2013)
21. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 3166–3173 (2013)
22. Wang, L., Xue, J., Zheng, N., Hua, G.: Automatic salient object extraction with contextual cue. In: IEEE Conf. Computer Vision (ICCV), Barcelona, pp. 2214–2219 (2011)
23. Jiang, H., Wang, J., Yuan, Z.: Automatic salient object segmentation based on context and shape prior. In: British Machine Vision Conf., vol. 3, no. 4, p. 7 (2011)
24. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: influence of depth cues on visual saliency. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 101–115. Springer, Heidelberg (2012)
25. Jiang, R., Crookes, D.: Deep salience: visual salience modeling via deep belief propagation. In: AAAI, pp. 2773–2779 (2014)
26. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD salient object detection: a benchmark and algorithms. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part III. LNCS, vol. 8691, pp. 92–109. Springer, Heidelberg (2014)
27. Koch, C., Ullman, S.: Shifts in Selective Visual Attention: towards the Underlying Neural Circuitry. Human Neurobiology **4**(4), 219–227 (1985)
28. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE Conf. Computer Vision (ICCV), Japan, pp. 2106–2113 (2009)
29. Zhang, X., Hu, K., Wang, L., Zhang, X.L., Wang, Y.G.: Salient Region Detection Based on Color Uniqueness and Color Spatial Distribution. IEICE Transactions on Information and Systems **97**(7), 1933–1936 (2014)
30. Duan, L., Kong, L.F.: Salient Region Detection with Hierarchical Image Abstraction. Journal of Information Science and Engineering **31**, 861–878 (2015)
31. Jiang, P., Ling, H., Yu, J., et al.: Salient region detection by UFO: uniqueness, focusness and objectness. In: IEEE Conf. Computer Vision (ICCV), pp. 1976–1983 (2013)
32. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 353–367 (2011)
33. Achanta, R., Smith, K., Lucchi, A.: SLIC Superpixels. Technical report (2010)
34. Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction. In: IEEE Conf. Computer Vision (ICCV), pp. 1529–1536 (2013)
35. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 1155–1162 (2013)
36. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 1139–1146 (2013)
37. Movahedi, V., Elder, J.H.: Design and perceptual validation of performance measures for salient object segmentation. In: IEEE Conf. Computer Vision and Pattern Recognition, USA, pp. 49–56 (2010)
38. Alpert, S., Galun, M., Brandt, A., et al.: Image segmentation by probabilistic bottom-up aggregation and cue integration. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(2), 315–327 (2012)

Simultaneously Retargeting and Super-Resolution for Stereoscopic Video

Kai Kang[1], Jing Zhang[1], Yang Cao[1], and Zeng-Fu Wang[1,2](✉)

[1] Department of Automation, University of Science and Technology of China,
HeiFei, China
{xzkk,zjwinner}@mail.ustc.edu.cn, {forrest,zfwang}@ustc.edu.cn

[2] Institute of Intelligent Machines, Chinese Academy of Sciences, HeiFei, China

Abstract. This paper presents a novel approach that is able to resize stereoscopic video to fit various display environments with different aspect-ratios, while preserving the prominent content, keeping temporally consistent, adapting depth, as well as increasing the resolution. Our proposed approach can address retargeting and super-resolution problems simultaneously via replacing the down-sampling matrix appearing in super-resolution algorithm with a novel one, named as content-aware-sampling matrix, derived from retargeting method. The new matrix can sample the image into any resolution while preserving its important information as much as possible. Our approach can be roughly subdivided into three steps. In the first step, we calculate the overall saliency map for a shot, while considering the conspicuous information from still image and the motion information from video. In the second step, given the target resolution, we compute the retargeting parameters by a global optimization and formulate them into a matrix. Finally, we substitute the matrix into the objective function used for super-resolution, and optimize it iteratively to achieve high visual quality outcome. The experimental results based on user studies verify the effectiveness of our approach.

Keywords: Stereoscopic video · Retargeting · Super-resolution

1 Introduction

Stereoscopic contents, such as still images and videos, extend visual communication to the third dimension by presenting two parallel views of the observed scenery. The fascinating 3D view experience has received much attention and the popularity of 3D entertainment has been significantly increased. In recent years, many researchers have made remarkable progress in 3D capture and display technology. More and more commercial products like 3D cinemas, televisions, smart phones and PDAs have come into our lives. Unfortunately, most of them have different resolutions and aspect-ratios. Fig. 1 presents a typical case when we expect stereoscopic contents to be viewed on a variety of display devices other than originally intended. As can be seen, the butterfly is stretched and the quality is degraded due to interpolation method. It is imperative to take some

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 31–40, 2015.
DOI: 10.1007/978-3-662-48570-5_4

measures to ameliorate visual experience. One is retargeting the image's aspect-ratio while protecting the important regions from being severely distorted, and the other is predicting unknown pixels from current observations to enhance details. Obviously, our work contains two classical problems, retargeting and super-resolution.

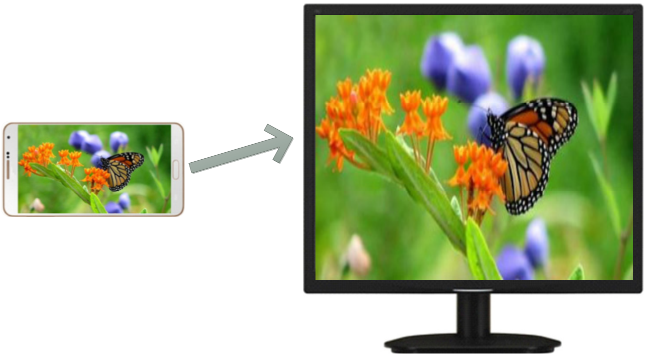


Fig. 1. A typical case when we put the low-resolution image that is suitable to phone's screen on a television. The result displayed on the television is generated by uniformly scaling along vertical dimension. Note the butterfly is stretched and the quality is degraded due to interpolation method. In this case, the phone's resolution is 1028 × 720 and the television's is 1028 × 1024.

In the past decades, a lot of retargeting algorithms with good performance have been devised. They can be classified into three categories, cropping, seam carving and warping. Very recently, Niu et al. [14] propose an aesthetics-based method which firstly automatically crops the periphery pixels of the input stereoscopic photo and uniformly scale it to fit various display devices while preserving its aesthetic value. Avidan et al. [1] develop a seam carving method which can greedily remove or insert horizontal or vertical seams, the paths of pixels, passing through the less important regions in the image. Subsequently, they also extend seam carving to retarget 2D video [15] while taking the temporal coherency into account via duplicating or deleting 2D seam manifolds from 3D space-time volumes instead of 1D seams. Afterward, Utsugi et al. [17] present a seam carving-based method to retarget stereoscopic image by fusing stereo matching results into the framework of seam carving and selecting appropriate type of seams to virtually manipulate the depths of objects in the scene. Lately, Guthier et al. [6] apply seam carving to stereoscopic video retargeting. However, seam carving can not avoid bringing serious discontinuity artifacts, what's worse, the artifacts are magnified for videos. On the contrary, Wang et al. [19] propose a kind of continuous method based on warping, which places a rectangular grid mesh onto the image then computes a new geometry for the mesh, such that the regions with high importance are scaled uniformly at the expense of spreading larger distortion to the other regions. This warping-based method has been extended to stereoscopic image retargeting [3] and stereoscopic video retargeting [9]. Chang et al. utilize rectangular grid mesh to simultaneously retarget a binocular image and adjusts depth by a sparse set of correspondences embedded in mesh without estimating depth map or dense correspondences. In addition, they pay more attention on how to adapt the depth to make comfortable visual experience. Recently, Kopf et al. propose warping-based method for stereoscopic video, which utilizes deformed pathlines to preserve the temporal coherence.

Super-resolution, a classical and challenging problem, aims at recovering the visually pleasing high-resolution image from one or more low-resolution input images. The existing methods can be roughly divided into three classes, interpolation methods [10], multi-frame methods [4], and example-based methods [8,16,20,21,25]. It is note that example-based methods have become the mainstream, as they have achieved outstanding results. In terms of video super-resolution, Liu et al. [12] use Bayesian theory to devise the state of the art approach to video super-resolution by estimating the underlying motion, blur kernel and noise level simultaneously. Recent years, many researchers have shifted their focus to mix-resolution image or video super-resolution, they utilize high-frequency information of the full-resolution view to up-sample the corresponding low-resolution view according to the correspondences indicated by the associated disparity map [5,22–24].

To our knowledge, no work has been reported on simultaneously solving retargeting and super-resolution problems. In fact, most of retargeting methods adopt simple interpolation methods, which are based on piecewise smooth assumption, to estimate the unknown pixels. As a consequence, the interpolation process deteriorates the quality of results. Particularly, when the resolutions before and after retargeting have a large size difference, the deteriorated effects become more and more noticeable. In this paper, we incorporate super-resolution algorithm into retargeting method by proposing a novel sampling matrix to achieve the good visual quality. The retargeting results of uniformly scaling, bilinear interpolation based method [3] and our method are shown in Fig. 2. To evaluate the performance of our approach, we have done subjective experiment on four stereoscopic videos[1]. And our experimental results demonstrate the effectiveness of the method.

This paper is organized as follows. Section 2 demonstrates how we simultaneously deal with retargeting and super-resolution problems. The experimental results are presented in Section 3. In the end, Section 4 present the conclusion of this paper.

2 Algorithm

In this section, we first explain how we implement the retargeting method for stereoscopic video. Next, we illustrate the modified model for stereoscopic video

[1] http://sp.cs.tut.fi/mobile3dtv/stereo-video/

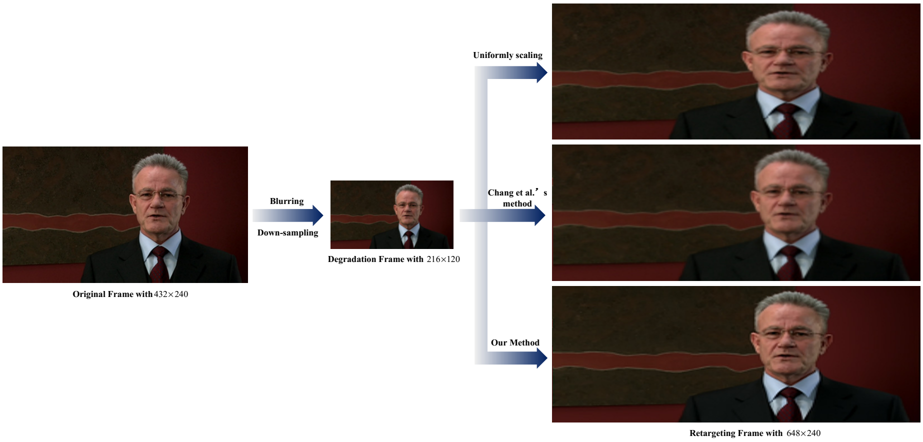

Uniformly scaling
Blurring
Down-sampling
Chang et al.' s method
Our Method
Original Frame with 432×240
Degradation Frame with 216×120
Retargeting Frame with 648×240

Fig. 2. The retargeting results of uniformly scaling, bilinear interpolation based method and our method. The presented results is based on the 70-th left view frame of video 'bullinger'. The original frame resolution is 432 × 240. We blur and down-sample it to generate the degradation version with 216 × 120. We test our method by retargeting the degradation version to high-resolution with 648 × 240. We can easily observe that the speaker's face has been stretched in uniformly scaling result, and the Chang et al.'s results is blurring. On the contrary, our method can produce high-quality results with sharp edges while does not distort the speaker area.

super-resolution. Finally, we demonstrate how to generate the content-aware-sampling matrix and fuse retargeting and super-resolution into a unified problem.

2.1 Stereoscopic Video Retargeting

For stereoscopic vide retargeting, we also utilize warping-based method, similar to [9]'s work. We extend [3]'s method to stereoscopic video retargeting. Different from Kopf et al.'s work, we calculate a uniform retargeting parameter for a shot, as we assume that there is no artificial camera motion and no severe movements of foreground objects in a shot. We have observed that no matter how much we weight the temporal consistency constraint there are still many noticeable flickering artifacts in the results. Considering the temporal consistency is the crucial fact for enjoyable viewing experiences, we prefer to retarget a shot uniformly at the expense of algorithm's flexibility.

Before introducing stereoscopic video retargeting, we illustrate how we obtain the uniform saliency map for a shot. Specifically, we exploit the graph-based method, a bottom-up spatial attention model, proposed by Harel et al. [7] to get i-th frame's 2D saliency map S_{2D}^{i}. Next, we adopt Liu's code [11] to estimate the optical flow fields. We treat velocity's magnitude as motion saliency value, then the i-th frame's motion saliency map S_{m}^{i} can be defined as:

$$S_{m}^{i}\left(p\right) = norm\left(\left\|\mathbf{o}\left(p\right)\right\|_{2}\right), \tag{1}$$

where $\mathbf{o}(p)$ represents the velocity at pixel p. And $norm(\bullet)$ is designed to normalize the saliency value between 0 to 1, besides, S_{2D}^{i} has been normalized. Next we smooth S_{2D}^{i} and S_{m}^{i} by

$$S_{smth}^{i} = \frac{1}{|Neibor(i)|} \sum_{j \in Neibor(i)} S^{j}, \tag{2}$$

where $Neibor(i)$ means the neighbor frames' indexes for the i-th frame, and $|Neibor(i)|$ is the number of neighbor frames. Then, we obtain each uniform saliency map by

$$S_{u} = \max_{i} \left(S_{smth}^{i} \right). \tag{3}$$

The overall uniform saliency map S_{u}^{comb} is obtained in a linear combination way

$$S_{u}^{comb} = \alpha S_{u}^{m} + (1-\alpha) S_{u}^{2D}, \tag{4}$$

where S_{u}^{m} and S_{u}^{2D} denote the uniform saliency maps of motion and 2D saliency. And α controls the trade-offs between motion and 2D saliency value. In our implement, we set α to 0.5.

Given the saliency map, we place grid mesh represented by $M = (V, E, Q)$ on each view of image respectively. Let $V = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_{end}\}$ and V' denote vertices' positions before and after retargeting. Note the left view vertex $\mathbf{v}_{i}^{L}$ corresponds to the right view vertex $\mathbf{v}_{i}^{R}$ with the same index i. Then we measure the importance of each quad $q \in Q$ by averaging its inside pixels' saliency value. $E(q)$ represents the edges set of quad $q \in Q$, and each edge can be denoted as $(\mathbf{v}_i, \mathbf{v}_j)$ where both $\mathbf{v}_i$ and $\mathbf{v}_j$ belong to quad $q \in Q$. The new mesh determined by output vertices's positions is obtained by minimizing the following energy function

$$\Psi = \lambda_q \left(\Psi_{q}^{L} + \Psi_{q}^{R} \right) + \lambda_l \left(\Psi_{l}^{L} + \Psi_{l}^{R} \right) + \lambda_a \Psi_a + \lambda_c \Psi_c, \tag{5}$$

where the upper right scripts indicate which view the energy belongs to. Similar to [19], Ψ_q and Ψ_l are **quad deformation** and **grid line bending** respectively. Like [3], we also exploit **alignment energy** Ψ_a and **disparity consistency energy** Ψ_c. Different from [3], we treat spare optical flow fields [11] between two views as the matched features instead of SIFT [13]. Although the SIFT is more accurate, optical flow is more stable than SIFT and can provide dense correspondence which can be flexibly sampled into sparse matched features. Please refer to [3,19] for more details.

2.2 Stereoscopic Video Super-Resolution

Our stereoscopic video super-resolution algorithm is based on multi-frame methods which take advantage of the sub-pixel displacements among the observations. Given a unknown high resolution (HR) frame I and a set of low resolution (LR) observations $Y = \{Y_1, Y_2, ..., Y_N\}$, the acquisition process of observations can be formulated as:

$$Y_k = DF_kHI + V, \; k = 1, 2, ..., N. \tag{6}$$

Note the unknown HR frame I, LR observation Y_k, as well as noise V are rearranged in column lexicographic order in the pixel domain. Suppose the HR frame's resolution is $rP \times rQ$ and each of the LR frame's is $P \times Q$, where r is the down-sampling factor, the sizes of I and Y_k are $(rP \times rQ) \times 1$ and $(P \times Q) \times 1$ respectively. The blurring matrix H, $(rP \times rQ) \times (rP \times rQ)$, is used to describe atmospheric, camera lens', or sensors' effects. The motion matrix F_k, $(rP \times rQ) \times (rP \times rQ)$, maps reference frame to the k-th frame. The down-sampling matrix D, $(P \times Q) \times (rP \times rQ)$, follows the sensor array sampling process. We assume that the blurring effect is approximated by point spread function and independent white Gaussian noise is added to the degraded frame.

As super-resolution is a kind of inverse problem, It is difficult to estimate the real solution. The reason is that when the number of observations is fewer than r^2, the problem becomes under-determined. In this case, there are an infinite number of solutions. When more than or equal to r^2 frames are available, the problem becomes square or over-determined. Although this kind of problem seems to have meaningful solution, the solution is still not stable. It is because that a little bit of noise will lead to large perturbations in the final solution. Most of super-resolution algorithm add image prior to this inverse problem to make the inverse problem more stable. In this paper, we adopt l_1-norm image prior [18], an approximation of total variation (TV) prior, due to its edge-preserving and piecewise-smoothing property. Then, the unknown frame I can be obtained by

$$\hat{I} = \arg\min_{I} \left[\sum_{k=1}^{N} \|Y_k - DF_kHI\|_2^2 \right] + \lambda_{l_1} \sum_i (|\varDelta_i^x I| + |\varDelta_i^y I|), \tag{7}$$

where $\varDelta_i^x$ and $\varDelta_i^y$ denote the horizontal and vertical first order differences at pixel i respectively, and λ_{l_1} is the regularization parameter, which is used to weight the first term (data term) against the second term (regularization term). Since both data and regularization terms are convex, we utilize the steepest descend method to gradually approach to the global optimization.

To our knowledge, it is a challenge to estimate high-quality HR motion from low-quality observations, and the quality of estimated motion concerns the performance of super-resolution algorithm. Since it is easy to gain the high-quality LR motion by estimating optical flow, we decide to put F_k in front of D in Eq.(7). Then, the size of F_k is $(P \times Q) \times (P \times Q)$. To make the estimation more robust, we exploit the accuracy of optical flow to weight data term in Eq.(7)

$$\hat{I} = \arg\min_{I} \left[\sum_{k=1}^{N} \mathbf{A}_k \|Y_k - F_kDHI\|_2^2 \right] + \lambda_{l_1} \sum_i (|\varDelta_i^x I| + |\varDelta_i^y I|), \tag{8}$$

where $\mathbf{A}_k$ denotes the accuracy weight matrix which contains the accuracy of the estimate motion from reference LR frame to k-th LR frame. It is a diagonal matrix whose diagonal elements have negative exponential relationship with the accuracy of optical flow.

It is worth to mention that stereoscopic video provides more reliable observations compared with monocular video. And experiments have confirmed that the more reliable observations have improved quality of outcomes.

2.3 Simultaneously Retargeting and Super-Resolution for Stereoscopic Video

This section illustrates how we realize retargeting and super-resolution simultaneously. Our work is dedicated to overcoming the blurring effects introduced by interpolation adopted by conventional retargeting methods, and extending current super-resolution methods to increase resolution to any size without distorting salient regions. We achieve these goals by replacing the down-sampling matrix D in Eq.(8) with a novel sampling matrix, named as content-aware-sampling matrix, denoted as R, which can be used to sample input image to arbitrary resolution. The Eq.(8) can be rewritten as

$$\hat{I} = \arg\min_{I} \left[\sum_{k=1}^{N} \mathbf{A}_k \left\| Y_k - F_k R H I \right\|_2^2 \right] + \lambda_{l_1} \sum_i \left(|\Delta_i^x I| + |\Delta_i^y I| \right). \tag{9}$$

Suppose we resize the LR observation from $P \times Q$ to $r_1 P \times r_2 Q$, where r_1 and r_2 are horizontal and vertical scaling factors respectively, the size of matrix R is $(P \times Q) \times (r_1 P \times r_2 Q)$. Since the factors r_1 and r_2 are independent, the output's aspect-ratio is arbitrary. We apply the method mentioned in section 2.1 to build up the vertices' warping relations between original domain and retargeting domain. As explained in section 2.1, the retargeting domain's vertices $V' = \{\mathbf{v}'_0, \mathbf{v}'_1, ...\mathbf{v}'_{end}\}$ are determined by a global optimization. The warping mapping $\Gamma(q)$ for each quad q can be computed by its vertices' positions. We assume that each quad undergoes an affine transformation. Then, the warping mapping can be obtained in a least-squares way. The affine transformation can be expressed as

$$\tilde{\mathbf{e}} = \Gamma(q)\,\mathbf{e}. \tag{10}$$

Since warping mapping is invertible, we can compute mapping from original domain to retargeting domain or from retargeting domain to original domain. In this work, we need the latter one. Note the augmented vector $\mathbf{e}$ and $\tilde{\mathbf{e}}$ in Eq. (10) represent the pixel positions in retargeting domain and original domain respectively. Generally, the new positions after mapping are generally non-integer, hence there is no pixel value that can be directly assigned to them. Like many retargeting methods, we adopt bilinear interpolation method to estimate an appropriate pixel value. Similar to the formulation of motion matrix F_k, we can formulate a sparse matrix R that describes a linear relationship between original domain $\tilde{I}$ and retargeting domain I

$$\tilde{I} = RI. \tag{11}$$

This linear relationship makes it possible to embed retargeting method in super-resolution framework, as illustrated in Eq.(9). To make it more clear, $\tilde{I}$ indicates the LR observation and I is the blurred version of HR unknown estimation. The matrix R is used to sample the HR resolution frame to LR observation one, which performs similar functions to down-sampling matrix D. Since matrix R stems from retargeting method which takes important information in account, we call the matrix R as content-aware-sampling matrix.

3 Experiment

In this section, we validate the potential of the proposed algorithm by processing four stereoscopic videos [2]. We have implemented our system on a PC with Intel CPU 2.80GH and RAM 4.00GB in MATLAB environment. In the retargeting part, we initialize quad with 20 × 20 pixels, and randomly abstract four high-quality optical flow features for each quad as the matched features. In the super-resolution part, we utilize first and last two frames as well as the corresponding frames on the other view to estimate current frame's high-resolution version. As we deal with color video in this paper, we estimate each color channel's high-resolution version separately.

Table 1. Parameters of the input videos

video	book arrival	bullinger	door flowers	leaving laptop
original size	512×384	432×240	512×384	512×384
degradation size	256×192	216×120	256×192	256×192
retargeting size	768×384	648×240	768×384	768×384
number of frame	100	100	150	100

Table 2. The average score of eight viewers' ratings

video	retargeting	super-resolution	overall
book arrival	2.75	2.75	2.875
bullinger	2.375	2.5	2.5
door flowers	2.375	2.25	2.375
leaving laptop	2.625	2.75	2.625

The testing videos' parameters are presented in table 1. In the experiment, we resize the degradation videos obtained by sequentially blurring and down-sampling original ones to high-resolution version but with different aspect-ratio. After resizing, the width of all videos has been scaled one-and-a-half times more than height. Then, original, degradation and retargeting group samples for each video are available. Besides, we add another group by uniformly scaling the degradation group. Next, we put them on LCD 3D display with 1360 × 768 resolution successively. Note we add black pixels to the videos to fit the display resolution. We invite eight viewers who are totally naive to our experiment to rate the results. The scores among 1 to 3 means worse, fine, well respectively. We ask the viewers to rate the retargeting score by comparing our results with the uniformly scaling results, as well as the super-resolution score by comparing with the original and degradation versions. Finally, we ask them rate the overall score in terms of visual experience. Table 2 presents the average score of eight viewers' ratings. From table, we find that the overall score is higher than other rows and the other two score are also relatively higher. This point verifies that the scores are valid. Since the scores are over 2 meaning more than fine, we can draw a conclusion that our method is effective.

4 Conclusion

In this paper, we propose a novel sampling matrix, inspired by warping based retargeting algorithms, to sample image to any resolution while considering its contents. Since we deal with stereoscopic videos, some modifications have been made to saliency detection and super-resolution methods. Experimental results on the four stereoscopic videos show that our method can increase (or decrease) and resize the resolution simultaneously without distorting prominent features.

References

1. Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. In: ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2007, vol. 26 (2007)
2. Brust, H., Tech, G., Muller, K.: Report on generation of mixed spatial resolution stereo data base. Tech. rep., MOBILE3DTV project (2009)
3. Chang, C.H., Liang, C.K., Chuang, Y.Y.: Content-aware display adaptation and interactive editing for stereoscopic images. IEEE Transactions on Multimedia **13**, 589–601 (2011)
4. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Transactions on Image Processing **13**, 1327–1344 (2004)
5. Garcia, D.C., Dorea, C., de Queiroz, R.L.: Super resolution for multiview images using depth information. IEEE Transactions on Circuits and Systems for Video Technology **22**, 1249–1256 (2012)
6. Guthier, B., Kiess, J., Kopf, S., Effelsberg, W.: Seam carving for stereoscopic video. In: IEEE IVMSP Workshop (2013)
7. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems (2007)
8. He, L., Qi, H., Zaretzki, R.: Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
9. Kopf, S., Guthier, B., Hipp, C., Kiess, J., Effelsberg, W.: Warping-based video retargeting for stereoscopic video. In: IEEE International Conference on Image Processing (ICIP) (2014)
10. Li, X., Orchard, M.T.: New edge-directed interpolation. IEEE Transactions on Image Processing **10**, 1521–1527 (2001)
11. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. Ph.D. thesis, Massachusetts Institute of Technology (2009)
12. Liu, C., Sun, D.: A bayesian approach to adaptive video super resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**, 91–110 (2004)

14. Niu, Y., Liu, F., Feng, W.C., Jin, H.: Aesthetics-based stereoscopic photo cropping for heterogeneous displays. IEEE Transactions on Multimedia **14**, 783–796 (2012)
15. Rubinstein, M., Shamir, A., Avidan, S.: Improved seam carving for video retargeting. In: ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2008, vol. 27 (2008)
16. Timofte, R., Smet, V.D., Gool, L.V.: Anchored neighborhood regression for fast example-based super-resolution. In: IEEE International Conference on Computer Vision (ICCV) (2013)
17. Utsugi, K., Shibahara, T., Koike, T., Takahashi, K., Naemura, T.: Seam carving for stereo images. In: 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON) (2010)
18. Villena, S., Vega, M., Molina, R., Katsaggelos, A.K.: Bayesian super-resolution image reconstruction using an l1 prior. In: Proceedings of 6th International Symposium on Image and Signal Processing and Analysis (2009)
19. Wang, Y.S., Tai, C.L., Sorkine, O., Lee, T.Y.: Optimized scale-and-stretch for image resizing. In: ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia 2008, vol. 27 (2008)
20. Yang, C.Y., Yang, M.H.: Fast direct super-resolution by simple functions. In: IEEE International Conference on Computer Vision (ICCV) (2013)
21. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Transactions on Image Processing **19**, 2861–2873 (2010)
22. Zhang, J., Cao, Y., Wang, Z.: A simultaneous method for 3d video super-resolution and high-quality depth estimation. In: IEEE International Conference on Image Processing (ICIP) (2013)
23. Zhang, J., Cao, Y., Zha, Z.J., Zheng, Z., Chen, C.W., Wang, Z.: A unified scheme for super-resolution and depth estimation from asymmetric stereoscopic video. IEEE Transactions on Circuits and Systems for Video Technology (2014). doi:10.1109/TCSVT.2014.2367356
24. Zhang, J., Cao, Y., Zheng, Z., Chen, C., Wang, Z.: A new closed loop method of super-resolution for multi-view images. Machine Vision and Applications **25**, 1685–1695 (2014)
25. Zhu, Y., Zhang, Y., Yuille, A.L.: Single image super-resolution using deformable patches. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)

Local Variation Joint Representation for Face Recognition with Single Sample per Person

Meng Yang(✉), Tiancheng Song, Shiqi Yu, and Linlin Shen

College of Computer Science and Software Engineering,
Shenzhen University, Shenzhen, China
yangmengpolyu@msn.com

Abstract. Sparse representation based classification (SRC) was originally applied to multiple-training-sample face recognition with promising performance. Recently SRC has been extended to face recognition with single sample per person by using variations extracted from a generic training set as an additional common dictionary. However, the extended SRC ignored to learn a better variation dictionary and to use local region information of face images. To address this issue, we propose a local variation joint representation (LVJR) method, which learns a variation dictionary and does joint and local collaborative representation for a query image. The learned variation dictionary was required to do similar representation for the same-type facial variations, while the joint and local collaborative representation could effectively use local information of face images. Experiments on the large-scale CMU Multi-PIE and AR databases demonstrate that the proposed LVJR method achieves better results compared with the existing solutions to the single sample per person problem.

Keywords: Local variation · Joint representation · Face recognition · Single sample per person

1 Introduction

As one of the most visible applications in computer vision and pattern recognition, face recognition (FR) has been receiving significant attention in the community [17]. In practical FR scenarios such as face identification/verification in uncontrolled or less controlled environment [6, 16], there are many problems which have attracted much attention of researchers. For instance, face recognition with single sample per person is one of the most important FR problems. In the scenarios (e.g., law enforcement, e-passport, driver license, etc), there is only a single training face image per person. This makes the problem of FR particularly hard since very limited information is provided to predict the variations in the query sample. How to achieve high FR performance in the case of single training sample per person (SSPP) is an important and challenging problems in FR.

The performance of FR would be greatly affected by the limited number of training samples per person [26]. First, many discriminant subspace and manifold learning algorithms (e.g., LDA and its variants [15]) cannot be directly applied to FR with SSPP.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 41–50, 2015.
DOI: 10.1007/978-3-662-48570-5_5

Second, sparse representation based classification (SRC) [12], cannot be easily applied to the problem of SSPP, either, since multiple training samples per person are needed to well reconstruct the query face. As reviewed in [26], many specially designed FR methods have been developed. According to the availability of an additional generic training set, the FR methods for SSPP can be divided into two categories: methods without using a generic training set, and methods with generic learning.

The SSPP methods without generic learning often extract robust local features (e.g., gradient orientation [10] and local binary pattern [1]), generate additional virtual training samples (e.g., via singular value decomposition [25], geometric transform and photometric changes [27]), or perform image partitioning (e.g., local patch based LDA [23], self-organizing maps of local patches [22], and multi-manifold learning from local patches [8]). Although these methods have reported improved FR results, they ignored to introduce additional variation information into the single-sample gallery set. Meanwhile, local feature extraction and discriminative learning from local patches can be sensitive to image variations (e.g., extreme illumination and expression), while the new information introduced by virtual training sample generation can be rather limited.

Opposite to the first category of FR with SSPP, methods with generic learning try to borrow new and useful information (e.g., generic intra-class variation) from a generic training set. An intrinsic reason is the fact that face image variations for different subjects share much similarity. Since a generic training set could be easily collected, it has been widely employed in [21, 20, 9] to extract discriminant information for FR with SSPP. For instance, the expression subspace and pose-invariant subspace were learned from a collected generic training set to solve the expression-invariant [21] and pose-invariant [9] FR problems, respectively. Deng *et al.* [3] extended the SRC method to FR with SSPP. The so-called Extended SRC (ESRC) computes the intra-class variation in a generic training set and then uses the generic variation matrix to code the difference between the query and gallery samples. Following ESRC, Zhu *et al.* [16] proposed a block-based method, in which the weight of each block is iteratively updated to reduce occlusion affect on the final coding.

Dictionary learning has been extensively studied in image processing and computer vision [14, 24], etc. However, most of the dictionary learning methods for pattern classification are conducted on the gallery set with multiple samples per class. How to better learn the variation dictionary is still an open question. Recently, Yang *et al.* [13] proposed sparse variation dictionary learning (SVDL) method to learn a variation dictionary and then project the variation dictionary to the space of gallery images.

Although much improvement has been reported, there are several issues remained the generic training based methods for FR with SSPP. First, the variation matrix can be very big and redundant since many subjects in the generic training set are involved. This will increase the computational burden of the final FR algorithm. Second, SVDL required that each subject in the generic training set should include the same number of variations, which may not be available in practical application. Third, ESRC and SVDL both use holistic features which may not effective for facial variation. Although the method proposed by Zhu *et al.* [16] uses local information, the iterative reweighted procedure would need more computation.

To solve the above mentioned problems, we propose to a local variation joint representation (LVJR) method for FR with SSPP. In order to better exploit different types of variation information, we learn a compact variation dictionary with powerful variation representation ability. In addition, a novel joint and local collaborative representation model was also proposed for the representation and classification of a query image. Extensive experiments have been conducted on the large-scale face databases with various variations, including illumination, expression, pose, session, and occlusion, etc. The experimental results show that the proposed LVJR achieves much better performance than state-of-the-art methods for FR with SSPP.

Section 2 presents a brief review of related work. Section 3 gives the proposed LVJR model. Section 4 describes the optimization procedure of LVJR. Section 5 conducts experiments and Section 6 concludes the paper.

2 Brief Review of Related Work

Recently, sparse representation based classification (SRC) [12], has achieved very promising results, which have led to many following works [3, 11]. However, SRC cannot be directly applied to FR with only a single sample per person (SSPP). To address this issue, Deng *et al.* [3] proposed to integrate the intra-class variation matrix extracted from a generic training set to represent the testing sample. Since our work is developed based on ESRC, we give a review of ESRC here.

Denote the intra-class variation matrix of generic training set by $\boldsymbol{V}=[\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n]$, where $\boldsymbol{V}_i$ is the i^{th}-type variation matrix, and each column of $\boldsymbol{V}_i$ is the difference between a i^{th}-type variation sample and a reference. Let $\boldsymbol{G}= [\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_c]$ and $\boldsymbol{y}$ be the gallery set with a single sample per person and the testing sample, respectively. The procedures of ESRC [3] are described as follows.

1. Sparsely code $\boldsymbol{y}$ on the matrix $[\boldsymbol{G}\ \boldsymbol{V}]$ via l_1-norm minimization :

$$\left[\hat{\boldsymbol{\rho}};\hat{\boldsymbol{\beta}}\right]=\arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left\|\boldsymbol{y}-[\boldsymbol{G}\ \boldsymbol{V}][\boldsymbol{\rho};\boldsymbol{\beta}]\right\|_2^2+\lambda\left\|[\boldsymbol{\rho};\boldsymbol{\beta}]\right\|_1 \tag{1}$$

   where λ is a scalar constant, $\boldsymbol{\rho}$ is the coding coefficient associated with $\boldsymbol{G}$, and $\boldsymbol{\beta}$ is the coding coefficient associated with $\boldsymbol{V}$.
2. Classify $\boldsymbol{y}$ via

$$\text{identity}(\boldsymbol{y})=\arg\min_i\left\|\boldsymbol{y}-\boldsymbol{g}_i\hat{\rho}_i-\boldsymbol{V}\hat{\boldsymbol{\beta}}\right\|_2 \tag{2}$$

   where $\hat{\boldsymbol{\rho}}=\left[\hat{\rho}_1;\hat{\rho}_2;\cdots;\hat{\rho}_c\right]$, and $\hat{\rho}_i$ is the coefficient associated with class *i*.

ESRC has shown interesting results [3], but it doesn’t learn a variance dictionary. Recently, Yang *et al.* [13] proposed a variation dictionary learning approach while it needs a strict requirement (e.g., each subject should have the same number of variations) on the generic training set and ignores to use local region information.

3 Local Variation Joint Representation (LVJR)

In this section, we proposed a local variation joint representation (LVJR) method for FR with SSPP. In LVJR, a variation dictionary learning with joint representation method was proposed for constructing a variation dictionary and a joint and local collaborative representation model was proposed for classification.

3.1 Variation Dictionary Learning with Joint Representation

Face images from different subjects have a big inter-class similarity. In FR with SSPP, we also assume that the face images from different subjects would share similar variations. This kind of assumption has been applied to FR [21][9] with improved results.

For a type of variation matrix of a local region, ***V***, we want to learn a variation dictionary ***D*** so that joint representation of these variations could be conducted on ***D***. Here joint representation requires that the coding coefficients of the variations in the same category should be similar. The proposed variation dictionary learning model could be written as

$$\min_{\boldsymbol{D},\boldsymbol{A}} \left\|\boldsymbol{V}-\boldsymbol{D}\boldsymbol{A}\right\|_2^2+\gamma\left\|\boldsymbol{A}\right\|_{2,1} \tag{3}$$

where $\|.\|_{2,1}$ is defined as $\|\boldsymbol{A}\|_{2,1}=\sum_k\|\boldsymbol{a}_k\|_2$, $\boldsymbol{a}_k$ is the k-th row vector of the coefficient matrix ***A***, γ are a scalar variable. The mixed-norm $\|.\|_{2,1}$ requires the between-row sparisity by using l_1-norm and regularizes the variables in each row vector via l_2-norm, which could make the variation in the same category (e.g., illumination with certain direction, certain type of expression) have similar coding vectors.

3.2 Joint and Local Collaborative Representation

Based on the learned variation dictionary we could develop a joint and local representation model to effectively exploit the local information. Let $\boldsymbol{y}=[\boldsymbol{y}^1,\boldsymbol{y}^2,\ldots,\boldsymbol{y}^K]$, where $\boldsymbol{y}^k$ is the k-th local region of ***y***. Similarly, the variation matrix of a generic training set could also be divided into K local regions, and each local region could learn a variation dictionary, $\boldsymbol{D}^k$.

In the joint and local representation phase, we want the coding coefficients of different local regions should be similar because these local regions come from the same query image. In order to efficiently solve the joint representation, we adopt l_2-norm to regularize the coding coefficients inspired by [7]. The proposed joint and local collaborative representation model could be written as

$$\min_{\boldsymbol{\alpha}^k} \sum_{k=1}^{K}\left(\left\|\boldsymbol{y}^k-\left[\boldsymbol{G}^k\ \boldsymbol{D}^k\right]\boldsymbol{\alpha}^k\right\|_F^2+\lambda\left\|\boldsymbol{\alpha}^k\right\|_2^2+\mu\left\|\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}}\right\|_F^2\right) \tag{4}$$

where $\boldsymbol{\alpha}^k=\left[\boldsymbol{\rho}^k;\boldsymbol{\beta}^k\right]$ is the coding coefficient for k-th local region, $\boldsymbol{\rho}^k$ is the coding sub-coefficient vector associated to the gallery set, $\boldsymbol{G}^k$, and $\boldsymbol{\beta}^k$ is the coding sub-coefficient vector associated to the variation dictionary, $\boldsymbol{D}^k$. Here $\bar{\boldsymbol{\alpha}}$ is the mean vector of all $\boldsymbol{\alpha}^k$.

When we solve Eq.(4), the classification could be conducted via

$$\text{identity} = \arg\min_i \left\{ \sum\nolimits_{k=1}^{K} \omega_k \left\| \boldsymbol{y}^k - \boldsymbol{g}_i \boldsymbol{\rho}_i^k - \boldsymbol{D}^k \boldsymbol{\beta}^k \right\|_2 \right\} \tag{5}$$

where $\omega_k = \exp\left(- \left\| \boldsymbol{y}^k - \boldsymbol{G}\boldsymbol{\rho}^k - \boldsymbol{D}^k \boldsymbol{\beta}^k \right\|_2^2 \Big/ 2\sigma^2 \right)$, $\boldsymbol{\rho}_i^k$ is the coding sub-coefficient associated to the i-th gallery image, and $\sigma^2 = \sum\nolimits_{k=1}^{K} \left\| \boldsymbol{y}^k - \boldsymbol{G}\boldsymbol{\rho}^k - \boldsymbol{D}^k \boldsymbol{\beta}^k \right\|_2^2 \Big/ K$.

4 Solving Algorithm of JLVR

4.1 Solving Variation Dictionary Learning

The model of variation dictionary learning with joint representation could be efficiently solved by alternatively updating the dictionary $\boldsymbol{D}$ and coding coefficient $\boldsymbol{A}$.

When the dictionary, $\boldsymbol{D}$, is fixed, Eq.(3) changes to

$$\min_{\boldsymbol{A}} \left\| \boldsymbol{V} - \boldsymbol{D}\boldsymbol{A} \right\|_2^2 + \gamma \left\| \boldsymbol{A} \right\|_{2,1} \tag{6}$$

which could be efficiently solved by the Iterative Projection Method [5]. Denote $\boldsymbol{\Lambda}=\boldsymbol{A}^{(t)}\text{-}(\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{A}^{(t)}\text{-}\boldsymbol{D}^T\boldsymbol{V})/\sigma$, the solution could be written as

$$\boldsymbol{A}^{(t+1)}[k]=\boldsymbol{\Lambda}[k]\cdot\text{Max}(0,1\text{-}\lambda/(2\sigma\|\boldsymbol{\Lambda}[k]\|_2)) \tag{7}$$

where σ is a scalar parameter in [37], Max(.) is a maximal operator, $\boldsymbol{A}^{(t+1)}[k]$ and $\boldsymbol{\Lambda}[k]$ are the k-th row vector of $\boldsymbol{A}^{(t+1)}$ and $\boldsymbol{\Lambda}$ in the t+1 iteration, respectively.

When the coding coefficient, $\boldsymbol{A}$, is fixed, Eq.(3) changes to

$$\min_{\boldsymbol{D}} \left\| \boldsymbol{V} - \boldsymbol{D}\boldsymbol{A} \right\|_2^2 \tag{8}$$

which could be efficient solved atom by atom via the metaface learning [4].

4.2 Solving Joint and Local Collaborative Representation

The proposed joint and local collaborative representation model, Eq.(4), could be efficiently solved. For each local region, the coding coefficient could be derived

$$\boldsymbol{\alpha}^k = \boldsymbol{\alpha}^{k,0} + \tau \boldsymbol{P}^k \bar{\boldsymbol{\alpha}} \tag{9}$$

where $\boldsymbol{\alpha}^{k,0} = \boldsymbol{P}^k \left[\boldsymbol{G}\ \boldsymbol{D}^k \right]^T \boldsymbol{y}^k$, and $\boldsymbol{P}^k = \left(\left[\boldsymbol{G}\ \boldsymbol{D}^k \right]^T \left[\boldsymbol{G}\ \boldsymbol{D}^k \right] + \lambda \boldsymbol{I} \right)^{-1}$.

Based on $\bar{\boldsymbol{\alpha}} = \sum\nolimits_{k=1}^{K} \boldsymbol{\alpha}^k \big/ K$ .By summing $\boldsymbol{\alpha}^k$, we could get

$$K\bar{\boldsymbol{\alpha}} = \sum\nolimits_{k=1}^{K} \boldsymbol{\alpha}^k = \sum\nolimits_{k=1}^{K} \boldsymbol{\alpha}^{k,0} + \tau \sum\nolimits_{k=1}^{K} \boldsymbol{P}^k \bar{\boldsymbol{\alpha}} \tag{10}$$

And then we could derive

$$\bar{\boldsymbol{\alpha}} == \left(I - \tau \big/ K \sum\nolimits_{k=1}^{K} \boldsymbol{P}^k \right)^{-1} \sum\nolimits_{k=1}^{K} \boldsymbol{\alpha}^{k,0} \big/ K \tag{11}$$

Based on Eq.(11) and Eq.(9), we could get an analytical solution of $\boldsymbol{\alpha}^k$.

5 Experiments

In this section, we perform FR with SSPP on benchmark face databases, including large-scale CMU Multiple PIE [19] and AR [2], to demonstrate the performance of SVDL. We first discuss the parameter setting in Section 5.1; in Section 5.2 we test the robustness of LVJR to various variations on CMU Multi-PIE; in Section 5.3, we evaluate LVJR on the AR database.

We compare the proposed LVJR with several state-of-the-art methods on FR with STSPP, including ESRC [3], ESRC-KSVD (the variation dictionary is learned via KSVD[24]), Adaptive Generic Learning (AGL) for Fisherfaces [18], and Discriminative Multi-Manifold Analysis (DMMA) [8], Sparse Variation Dictionary Learning (SVDL) [13], and some baseline classifiers such as SRC [12], Nearest Subspace (NS) and Support Vector machine (SVM). It should be noted that NS is reduced to Nearest Neighbor (NN) in the case of FR with STSPP. Among these methods, NN, SVM, SRC and DMMA do not use a generic training set, while ESRC, AGL, SVDL and JVJR need a generic training set.

5.1 Parameter Setting

There are three regularization parameters, γ, λ and μ, in LVJR. γregularizes the variation dictionary learning, while λ and μ controls the l_2-norm regularization and similarity of coding coefficients in the joint and local collaborative representation. If no specific instruction, we fix γ=λ =μ=0.005, and initialize dictionary atom number as 400.

5.2 Evaluation to Various Variation on CMU-PIE Dataset

We test the robustness of all the competing methods by using the large-scale CMU Multi-PIE database [19], whose images were captured in four sessions with simultaneous variations of pose, expression, and illumination. For each subject in each session, there are 20 illuminations with indices from 0 to 19 per pose per expression. Among the 249 subjects in Session 1, the first 100 subjects were used for gallery training, with the remaining subjects for generic training. For the gallery set, we used the single frontal image with illumination 7 and neutral expression. The image is cropped to 100×82. Here LVJR divided a face image into 2×2 local regions.

1) Illumination Variation: as [13], we use all the frontal face images with neutral expression in Sessions 2, 3, and 4 for testing. The generic training set is composed of all the frontal face images with neutral expression in Session 1. Table 1 lists the recognition rates in the three sessions by the competing methods.

From Table 1, we can see that LVJR achieves the best results in all cases, and SVDL performs the second best, followed by ESRC. That shows a learned variation dictionary could generate a better performance. SRC does not get good result since the single training sample of each class has very low representation ability. DMMA is the best method without using generic training set; nonetheless, its recognition rates are not high since the illumination variation cannot be well learned from the gallery set via multi-manifold learning.

Table 1. Face recognition rates on Multi-PIE database with illumination variations.

Session	Session 2	Session 3	Session 4
NN	45.3%	40.2%	43.7%
SVM	45.3%	40.2%	43.7%
SRC [12]	52.4%	46.7%	49.5%
DMMA [8]	63.2%	55.4%	60.4%
AGL [18]	84.9%	79.4%	78.3%
ESRC [3]	92.6%	84.9%	86.7%
ESRC-KSVD	92.7%	84.9%	86.7%
SVDL	94.8%	87.7%	91.0%
LVJR	**96.0%**	**90.9%**	**92.1%**

2) Expression and Illumination Variations: as [13], the testing samples include the frontal face images with smile in Session 1, smile in Session 3, and surprise in Session 2. In each test, the images in the generic training set include all the frontal face images with the corresponding expression and the frontal face image with illumination 7 and neutral expression in Session 1. The recognition rates of all competing methods are listed in Table 2.

Table 2. Face recognition rates on Multi-PIE database with expression and illumination variations.

Expression	Smi-S1	Sim-S3	Sur-S2
NN	46.9%	28.8%	18.0%
SVM	46.9%	28.8%	18.0%
SRC [12]	49.6%	28.1%	20.4%
DMMA[8]	58.2%	31.5%	22.0%
AGL [18]	84.9%	39.3%	31.3%
ESRC [3]	81.6%	50.5%	49.6%
ESRC-KSVD	85.0%	50.4%	51.2%
SVDL	88.8%	58.6%	54.7%
LVJR	**93.7%**	**63.9%**	**67.6%**

We can see that LVJR outperforms all the other methods in all the three tests, with at least 4.9%, 5.3%, and 12.9% improvements over the second best, SVDL. That validates that the local information explored in our proposed LVJR is very helpful for final recognition. In addition, all the methods achieve the best results when Smi-S1 is used for testing because the training set is also from Session 1. Again, the methods using generic training set usually have better performance than the ones without using generic training set.

3) Pose, Illumination and Expression Variations: As [13], the testing samples include face images with pose ‘05_0’ in session 2 and pose ‘04_1’ in session 3, and face images with pose ‘04_1’ and smile expression in Session 3 (please refer to Figs. 1(b)~(d) for examples). In each test, the images in the generic training set include all the face images with the corresponding expression and pose, and the frontal face image with illumination 7 and neutral expression in Session 1. The recognition rates of all competing methods are listed in Table 3.

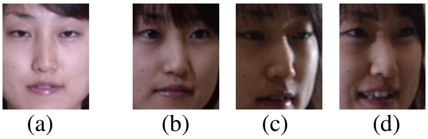


Fig. 1. Face images with pose variations in different sessions. (a) shows the single gallery sample; (b), (c) and (d) show the testing samples with pose, illumination and expression variations in Sessions 2 and 3, respectively.

Table 3. Face recognition rates on Multi-PIE database with pose, expression and illumination variations.

Pose	P05_0-S2	P04_1-S3	Smi-P04_1-S3
NN	26.0%	8.7%	12.0%
SVM	26.0%	8.7%	12.0%
SRC [12]	25.0%	7.3%	10.3%
DMMA[8]	27.1%	5.3%	11.0%
AGL [18]	66.7%	24.9%	23.9%
ESRC [3]	63.9%	31.8%	26.9%
ESRC-KSVD	67.1%	29.9%	25.6%
SVDL	77.8%	38.3%	34.4%
LVJR	**80.4%**	**40.0%**	**35.4%**

From Table 3, we can see that LVJR is still the best methods although the recognition rates of all methods are not high for big pose variation. This experiment also validate that the joint representation of local information on the learned variation dictionary could advance the recognition accuracy. We also run the experiments by learning a variation dictionary without requiring the representation of variation in the same category should be similar, of which the results (e.g., 79.7%, 31.9% and 33.1%) are lower than LVJR.

5.3 Evaluation on Various Variation AR Database

We then conduct FR with SSPP on the AR database [2]. As [16], a subset of AR contains two-session data of 50 male and 50 female subjects (each person has 26 pictures with the normalized size as 165×120) are included in the experiments. For each subjects there are two sessions and for variations (e.g., expression, illumination, disguise, and disguise+illumination). Here for each subject, the neutral face image without disguise and illumination in Session 1 is used as a gallery image. And the first 80 subjects are used to construct the gallery set and query set, with the remaining subjects for a generic training set. Here the face images are divided into 7×7 local regions and $\gamma$and λ are set as 0.001 and 0.05, respectively.

The recognition rates of the competing methods for query images from Session 1 are listed in Table 4. LVJR gets much better performance than all the other methods. For instance, compared to SVDL, the improvement for the variation of illumination and disguise is nearly 9%, which shows that local information could be effectively exploited by the proposed LVJR.

Table 4. Recognition accuracy (%) on AR database (Session1)

Pose	Illumination	expression	Disguise	Illumination+disguise
SVM	55.8	90.4	43.1	29.4
SRC [12]	80.8	85.4	55.6	25.3
DMMA[8]	92.1	81.4	46.9	30.9
AGL [18]	93.3	77.9	70.0	53.8
ESRC [3]	99.6	85.0	83.1	68.6
SVDL	98.3	86.3	86.3	79.4
LVJR	**100**	**94.6**	**93.1**	**88.4**

We also compare the proposed LVJR with Local generic representation (LRG) [16] in Session 1 of AR dataset. Following the experimental setting of [16], the face images are divided into 4×4 local regions. The accuracy and average running time on the same machine are listed in Table 5. LVJR is 25 times faster than LGR but with similar accuracy.

Table 5. Recognition accuracy (%) on AR database (Session1) of LGR and LVJR

Variation	Illumination	Expression	disguise
LGR	**100** (0.53second)	97.9(0.52second)	**98.8**(0.53second)
LVJR	**100(0.02 second)**	**98.8(0.02second)**	**98.8(0.02second)**

6 Conclusion

In this paper, we proposed a local variation joint representation method, which learns a variation dictionary with joint representation and does a joint and local collaborative representation. The learned variation dictionary could well exploit the variance information in the generic training set while with a small size. And the joint and local collaborative representation could fully use the local information of face images. The extensive experiments with various face variations demonstrated the superiority of LVJR to state-of-the-art methods for face recognition with SSPP.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China under Grant no. 61402289, National Science Foundation of Guangdong Province (2014A030313558), and Shenzhen Scientific Research and Development Funding Program under Grant JCYJ20140509172609171.

Reference

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)
2. Martinez, A., Benavente, R.: The AR face database. Technical Report 24, CVC (1998)
3. Deng, W.H., Hu, J.N., Guo, J.: Extended src: undersampled face recognition via intra-class variant dictionary. IEEE PAMI **34**(9), 1864–1870 (2012)
4. Yang, M., Zhang, L., Yang, J., Zhang, D.: Metaface learning for sparse representation based face recognition. In: ICIP (2010)

5. Rosasco, L., Verri, A., Santoro, M., Mosci, S., Villa, S.: Iterative Projection Methods for Structured Sparsity Regularization. MIT Technical Reports, MIT-CSAIL-TR-2009-050, CBCL-282
6. Wolf, L., Hassner, T., Taigman, Y.: Effective face recognition by combining multiple descriptors and learned background statistics. IEEE PAMI **33**(10), 1978–1990 (2011)
7. Zhang, L., Yang, M., Feng, X.C.: Sparse representation or collaborative representation: which helps face recognition? In: Proc. ICCV (2011)
8. Lu, J.W., Tan, Y.P., Wang, G.: Discriminative multi-manifold analysis for face recognition from a single training sample per person. IEEE PAMI **35**(1), 39–51 (2013)
9. Li, A.N., Shan, S.G., Gao, W.: Coupled bias–variance tradeoff for cross-pose face recognition. IEEE TIP **21**(1), 305–315 (2012)
10. Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: Subspace learning from image gradient orientations. IEEE PAMI **34**(12), 2454–2466 (2012)
11. Wagner, A., Wright, J., Ganesh, A., Zhou, Z.H., Mobahi, H., Ma, Y.: Towards a practical face recognition system: robust alignment and illumination by sparse representation. IEEE PAMI **34**(2), 372–386 (2012)
12. Wright, J., Yang, A.Y., Sastry, S.S., Ma, Y.: Robust face recogntion via sparse representation. IEEE PAMI **31**(2), 210–227 (2009)
13. Yang, M., Van Gool, L., Zhang, L.: Sparse variation dictionary learning for face recognition with a single training sample per person. In: Proc. ICCV (2013)
14. Rubinstein, R., Bruckstein, A., Elad, M.: Dictionary learning for sparse representation modeling. Proceeding of the IEEE **98**(6), 1045–1057 (2010)
15. Belhumeur, P.N., Hespanha, J., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE PAMI **19**(7), 711–720 (1997)
16. Zhu, P., Yang, M., Zhang, L., Lee, I.-Y.: Local generic representation for face recognition with single sample per person. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 34–50. Springer, Heidelberg (2015)
17. Zhao, W., Chellppa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Survey **35**(4), 399–458 (2003)
18. Su, Y., Shan, S., Chen, X., Gao, W.: Adaptive generic learning for face recognition from a single sample per person. In: CVPR (2010)
19. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing **28**, 807–813 (2010)
20. Wang, J., Plataniotis, K., Lu, J., Venetsanopoulos, A.: On solving the face recognition problem with one training sample per subject. Pattern Recognition **39**, 1746–1762 (2006)
21. Mohammadzade, H., Hatzinakos, D.: Expression subspace projection for face recognition from single sample per person. IEEE Affective Computing **4**(1), 69–82 (2012)
22. Tan, X., Chen, S., Zhou, Z., Zhang, F.: Recognizing partially occluded expression variant faces from single training image per person with SOM and soft k-NN ensemble. IEEE NN **16**(4), 875–886 (2005)
23. Chen, S., Liu, J., Zhou, Z.: Making FLDA applicable to face recognition with one sample per person. Pattern Recognition **37**(7), 1553–1555 (2004)
24. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE SP **54**(11), 4311–4322
25. Zhang, D., Chen, S., Zhou, Z.: A new face recognition method based on SVD perturbation for single example image per person. Applied Mathematics and Computation **163**(2), 895–907 (2005)
26. Tan, X., Chen, S., Zhou, Z., Zhang, F.: Face recognition from a single image per person: A survey. Pattern Recognition **39**, 1725–1745 (2006)
27. Shan, S., Cao, B., Gao, W., Zhao, D.: Extended fisherface for face recognition from a single example image per person. ISCAS **2**, 81–84 (2002)

Multi-view Sparse Embedding Analysis Based Image Feature Extraction and Classification

Yangping Zhu[1], Xiaoyuan Jing[1,2](✉), Qing Wang[3], Fei Wu[4], Hui Feng[2], and Shanshan Wu[1]

[1] College of Automation, Nanjing University of Posts and Telecommunications,
Nanjing 210023, China
zhuyangping_1991@163.com, jingxy_2000@126.com

[2] State Key Laboratory of Software Engineering, Wuhan University,
Wuhan 430079, China

[3] College of Computer, Nanjing University of Posts and Telecommunications,
Nanjing 210003, China
happywq2009@126.com

[4] College of Communication and Information Engineering,
Nanjing University of Posts and Telecommunications, Nanjing 210003, China
wufei_8888g@126.com

Abstract. Multi-view feature extraction is an attractive research topic in computer vision domain, since it can well reveal the inherent property of images. Most existing multi-view feature extraction methods focus on investigating the intra-view or inter-view correlation. However, they fail to consider the sparse reconstruction relationship and the discriminant correlation in multi-view data, simultaneously. In this paper, we propose a novel multi-view feature extraction approach named Multi-view Sparse Embedding Analysis (MSEA). MSEA not only explores the sparse reconstruction relationship that hides in multi-view data, but also considers discriminant correlation by maximizing the within-class correlation and simultaneously minimizing the between-class correlation from intra-view. Moreover, we add orthogonal constraints of embedding matrices to remove the redundancy among views. To tackle the linearly inseparable problem in original feature space, we further provide a kernelized extension of MSEA called KMSEA. The experimental results on two datasets demonstrate the proposed approaches outperform several state-of-the-art related methods.

Keywords: Sparse embedding analysis · Multi-view · Discriminant correlation · Orthogonal constraints

1 Introduction

In computer vision domain, many applications are usually involved with different views of data. With respect to feature extraction, multi-view features can

Y. Zhu and X. Jing—The work described in this paper was fully supported by the National Natural Science Foundation of China under Project No. 61272273, the Research Project of Nanjing University of Posts and Telecommunications under Project No. XJKY14016, and the Postgraduate Scientific Research and Innovation Plan of Jiangsu Province Universities under Project No. CXLX13-465.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 51–60, 2015.
DOI: 10.1007/978-3-662-48570-5_6

well reveal the inherent property of data. Multi-view feature extraction aims to exploit different characteristics or views of data, which is an attractive and important research direction [1,2].

Existing supervised multi-view extraction methods can be roughly categorized into two types. ***(1) Shared subspace learning based methods.*** They focus on learning a common shared subspace, in which the correlation among multiple views can be well revealed. Mostly they are based on canonical correlation analysis (CCA) [3], which is a vital multi-view extraction technique, since it can well utilize the inter-view correlation. Other shared subspace learning based methods include discriminant analysis of canonical correlations (DCC) [4], multiple discriminant CCA (MDCCA) [5], multi-view discriminant analysis (MvDA) [6], intra-view and inter-view supervised correlation analysis (I^{2}SCA) [7], etc. ***(2) Transfer learning and dictionary learning based methods.*** They focus on incorporating the transfer learning or dictionary learning techniques into the multi-view feature extraction process. Transfer learning can alleviate the distribution differences among different views. And dictionary learning holds favorable reconstruction capability for multi-view features. Based on them, transfer component analysis (TCA) [8] and uncorrelated multi-view fisher discrimination dictionary learning (UMDDL) [9] are presented.

Although there exist much effort on multi-view extraction, existing methods almost fail to preserve the sparse reconstruction relationship and simultaneously consider the discriminant correlation in multi-view data. In this paper, we propose a novel multi-view feature extraction approach named Multi-view Sparse Embedding Analysis (MSEA). The contributions are summarized as follows:

1. We incorporate the sparse embedding analysis and learn a shared dictionary for multiple views, such that the sparse reconstruction relationship in multi-view data can be well preserved. Moreover, we consider the discriminative correlation by maximizing the within-class correlation and simultaneously minimizing the between-class correlation from intra-view. Since there exist much redundancy in multi-view features, we add the orthogonal constraints into the objective function, such that the redundant information among views can be effectively reduced.
2. We further provide a kernelized extension of MSEA, that is, KMSEA, to tackle the linearly inseparable problem in the original feature space.

The rest of this paper is organized as follows. In Section 2, we briefly review the related work. In Section 3, we describe the proposed MSEA approach and its kernelized extension KMSEA. Experimental results and analysis are provided in Section 4, and conclusion is drawn in Section 5.

2 Related Work

In this section, we briefly review the related methods, which are generally divided into following two types.

Shared subspace learning based methods mainly try to learn a common shared subspace for multiple views. Discriminant analysis of canonical correlations (DCC) [4] maximizes the within-class correlation and minimizes the between-class correlation for two sets of variables. Multiple discriminant CCA (MDCCA) [5] was designed for multiple views in the comparison with DCC. Kan et al. [6] presented a Multi-view discriminant analysis (MvDA) method, which maximizes between-class variations and minimizes within-class variations of the learning common space from both intra-view and inter-view. Intra-view and inter-view supervised correlation analysis (I^{2}SCA) [7] simultaneously extracts the discriminatingly correlated features from both inter-view and intra-view.

Transfer learning and dictionary learning based methods are mainly based on the transfer learning and dictionary learning techniques. Transfer component analysis (TCA) [8] attempts towards learning a few transfer components across domains by using maximum mean miscrepancy strategy. In the subspace spanned by these transfer components, data properties are preserved and data distributions in different domains are close to each other. Uncorrelated Multi-view Fisher Discrimination Dictionary Learning (UMDDL) [9] learns the multiple structural and discriminant dictionaries, which can well reconstruct the multi-view data.

3 Proposed Approach

3.1 Multi-view Sparse Embedding Analysis (MSEA)

Multi-view features can reveal the inherent property of data. Although these features come from different views, there exist some useful latent shared information, e.g., sparse structure, in the multi-view data [9]. How to effectively exploit this kind of latent sparse structure is vital for improving the performance of multi-view feature extraction. In this paper, we attempt towards incorporating the sparse embedding analysis into multi-view feature extraction. We learn a shared dictionary and multiple embedding matrices, which can make inherent sparse structure still be preserved in the projected multi-view features. The scheme of our MSEA is illustrated in Fig. 1.

The entire objective function of our MSEA contains three parts: sparse embedding analysis, intra-view discriminant correlation, and orthogonal constraints of embedding matrices. Then we describe these three parts in detail.

(1) Sparse Embedding Analysis. Let Y_i denote the i^{th} view of samples, and assume that they have been normalized, that is, $\hat{Y}_i \hat{Y}_i^T = 1, i = 1, 2, ..., N$, where N is the number of view. We try to learn multiple embedding matrices, with each corresponding to one view. The target is to project the original feature of multi-view samples into a shared subspace and help learn the shared dictionary. Then, this part of the objective function is defined as follows:

$$\begin{aligned} &\langle D, W, X \rangle \arg \min_{D,W,X} \sum_{i=1}^{N} \left\| W_i \hat{Y}_i - DX \right\|_F^2, \\ &s.t. \ \|x_j\|_0 \leq T_0, \ \forall j \end{aligned} \tag{1}$$

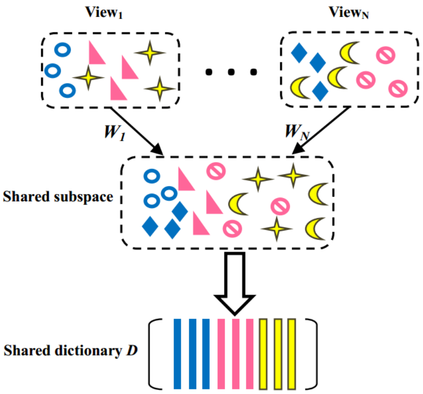


Fig. 1. Illustration of the Scheme of MSEA.

where W_i is the i^{th} embedding corresponding to the i^{th} view. The dimensionality of W_i is $p \times n$, where n is the dimensionality of original samples feature and p is the dimensionality of feature after embedding. D is the shared dictionary, and X is the sparse representation coefficients.

(2) Intra-view Discriminant Correlation Maximization. Recently, some study shows that the correlation information in views is significant in the feature extraction [7],[9]. To make the extracted features hold favorable discriminability, our MSEA tries to incorporate the intra-view discriminant correlation into the objective function. This target is to maximize the within-class correlation and minimize the between-class correlation from intra-view, simultaneously, that is,

$$\langle W \rangle = \arg\max_{W_i} \sum_{i=1}^{N} \left(C^i \right). \tag{2}$$

The discriminant correlation mentioned above can be defined as $C^i = C_w^i - \beta C_b^i$, where C_w^i is the intra-view within-class correlation and C_b^i is the intra-view between-class correlation of the i^{th} view. $\beta > 0$ is a tunable parameter that indicates the relative significance of C_w^i versus C_b^i. Specifically, C_w^i and C_b^i are defined as

$$\begin{aligned} C_w^i &= \frac{\left[1/\sum\limits_{p=1}^{c} n_p^2 \right] \sum\limits_{p=1}^{c} \sum\limits_{r=1}^{n_p} \sum\limits_{t=1}^{n_p} {\hat{y}_{pr}^i}^T W_i^T W_i \hat{y}_{pt}^i}{\sqrt{\frac{1}{n} \sum\limits_{p=1}^{c} \sum\limits_{r=1}^{n_p} \left(y_{pr}^i - \bar{y}^i \right)^T W_i^T \left(y_{pr}^i - \bar{y}^i \right) W_i} \sqrt{\frac{1}{n} \sum\limits_{p=1}^{c} \sum\limits_{t=1}^{n_p} \left(y_{pt}^i - \bar{y}^i \right)^T W_i^T \left(y_{pt}^i - \bar{y}^i \right) W_i}}, \\ &= \frac{n \bullet tr \left\{ \sum\limits_{p=1}^{c} \sum\limits_{r=1}^{n_p} \sum\limits_{t=1}^{n_p} W_i \hat{y}_{pt}^i \hat{y}_{pr}^{iT} W_i^T \right\}}{\left(\sum\limits_{p=1}^{c} n_p^2 \right) \sqrt{\hat{Y}_i^T W_i^T W_i \hat{Y}_i} \sqrt{\hat{Y}_i^T W_i^T W_i \hat{Y}_i}} = \frac{n \bullet tr \left\{ W_i \hat{Y}_i A \hat{Y}_i^T W_i^T \right\}}{\left(\sum\limits_{p=1}^{c} n_p^2 \right) \hat{Y}_i^T W_i^T W_i \hat{Y}_i} \end{aligned}$$

$$C_b^i = \frac{\left[1/\left(n^2 - \sum\limits_{p=1}^{c} n_p^2\right)\right] \sum\limits_{p=1}^{c} \sum\limits_{\substack{q=1\\q\neq p}}^{c} \sum\limits_{r=1}^{n_p} \sum\limits_{t=1}^{n_q} {\hat{y}_{pr}^i}^T W_i^T W_i \hat{y}_{qt}^i}{\sqrt{\frac{1}{n} \sum\limits_{p=1}^{c} \sum\limits_{r=1}^{n_p} \left(y_{pr}^i - \bar{y}^i\right)^T W_i^T \left(y_{pr}^i - \bar{y}^i\right) W_i} \sqrt{\frac{1}{n} \sum\limits_{q=1}^{c} \sum\limits_{t=1}^{n_p} \left(y_{qt}^i - \bar{y}^i\right)^T W_i^T \left(y_{qt}^i - \bar{y}^i\right) W_i}},$$

$$= \frac{n \bullet tr\left\{\sum\limits_{p=1}^{c} \sum\limits_{r=1}^{n_p} \sum\limits_{t=1}^{n_p} W_i \hat{y}_{pt}^i \hat{y}_{pr}^{iT} W_i^T\right\}}{\left(n^2 - \sum\limits_{p=1}^{c} n_p^2\right) \sqrt{\hat{Y}_i^T W_i^T W_i \hat{Y}_i} \sqrt{\hat{Y}_i^T W_i^T W_i \hat{Y}_i}} = -\frac{n \bullet tr\left\{W_i \hat{Y}_i A \hat{Y}_i^T W_i^T\right\}}{\left(n^2 - \sum\limits_{p=1}^{c} n_p^2\right) \hat{Y}_i^T W_i^T W_i \hat{Y}_i}$$

where $A = diag\left(E_{n_1}, E_{n_2}, ..., E_{n_c}\right)$ denotes a $n \times n$ symmetric, positive semi-definite, blocked diagonal matrix. E_{n_k} is a $n_k \times n_k$ matrix with all elements equalling to 1. Since A is a positive semi-definite matrix, we let $A = HH^T$ and obtain a more brief representation of this part in objective function:

$$\langle W \rangle = \arg\min_{\tilde{W}} \gamma \left\| \tilde{W} \tilde{Y} H \right\|_F^2, \tag{3}$$

where $\gamma = \left(\frac{n}{\sum\limits_{p=1}^{c} n_p^2} + \frac{n\beta}{n^2 - \sum\limits_{p=1}^{c} n_p^2}\right)$, $\tilde{W} = [W_1, ..., W_N]$, $\tilde{Y} = \begin{pmatrix} \hat{Y}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{Y}_N \end{pmatrix}$.

(3) Orthogonal Constraints of Embedding Matrices. Although multi-view data reveal different characteristics of data, there also exist some redundant information among those views describing the same object. Therefore, to remove this kind of redundant information, we add the orthogonal constraints of above learned embedding matrices, that is,

$$W_i {W_i}^T = I, i = 1, 2, ..., N. \tag{4}$$

By combining the Formula (1), (3) and (4), the entire objective function of our MSEA is defined as follows:

$$\begin{aligned} &\langle D, W, X \rangle = \arg \min_{D, \tilde{W}, X} \left\| \tilde{W} \tilde{Y} - DX \right\|_F^2 - \gamma \left\| \tilde{W} \tilde{Y} H \right\|_F^2. \\ &s.t. W_i {W_i}^T = I, i = 1, 2, ..., N, and \|x_j\|_0 \leq T_0, \forall j \end{aligned} \tag{5}$$

3.2 The Optimization of MSEA

There is no theoretical guarantee that our objective function in Formula (5) is jointly convex to (D, W, X). However, it is convex with respect to each of D, W, X when the others are fixed. Hence, this objective function can be solved based on the idea of divide-and-conquer. Before we conduct the iterative solution for MSEA, we try to simplify the optimization problem in Formula (5) by using a optimization trick in the literature [15]. We introduce two matrices, $Q \in n \times p$

and $B \in n \times p$, where n is the size of original samples feature and p is the size of feature after embedding. Then, the embedding matrices W_i can be represented by $W_i = {Q_i}^T {Y_i}^T$ and the shared dictionary D can be represented by $D = \tilde{W}\tilde{Y}B$. With this optimization trick and after some manipulations, the original optimization problem in Formula (5) can be simplified as follows:

$$\begin{aligned} &\arg\min_{\tilde{Q},X,B} \left\|\tilde{Q}^T K - \tilde{Q}^T KBX\right\|_F^2 - \gamma \left\|\tilde{Q}^T KH\right\|_F^2, \\ &s.t. {Q_i}^T K_i Q_i = I, i = 1, 2, ..., N, \left\|x_j\right\|_0 \leq T_0, \forall j \end{aligned} \tag{6}$$

where $K_i = \hat{Y}_i^T \hat{Y}_i$, $\tilde{Q} = [Q_1, ..., Q_N]$, and $K = \begin{pmatrix} K_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K_N \end{pmatrix}$. The above Formula also can be tackled by the divide-and-conquer strategy. We divide the objective function in Formula (6) into three sub-problems:

(1) Updating $\boldsymbol{X}$**.** We update the sparse representation coefficients X by fixing the matrix B and $\tilde{Q}$. The objective function can be simplified as follows:

$$\begin{aligned} &\langle X \rangle = \arg\min_X \left\|\tilde{Q}^T K - \tilde{Q}^T KBX\right\|_F^2. \\ &s.t. \left\|x_j\right\|_0 \leq T_0, \forall j \end{aligned} \tag{7}$$

This is a typical sparse representation problem, which has been effectively solved by method of optimal directions (MOD) [10]. We directly utilize the MOD algorithm to update X.

(2) Updating $\boldsymbol{B}$**.** We update the matrix B by fixing X and $\tilde{Q}$. The objective function can be simplified as follows:

$$\langle B \rangle = \arg\min_B \left\|\tilde{Q}^T K - \tilde{Q}^T KBX\right\|_F^2. \tag{8}$$

We let $L(B) = \left\|\tilde{Q}^T K - \tilde{Q}^T KBX\right\|_F^2$, and take the derivative of $L(B)$ with respect to B. By setting the derivative result being equal to zero, we obtain:

$$B = X^T \left(XX^T\right)^{-1}. \tag{9}$$

(3) Updating $\boldsymbol{\tilde{Q}}$**.** We update the matrix $\tilde{Q}$ by fixing X and B. First, we conduct singular value decomposition (SVD) on K, and further let $U = S^{\frac{1}{2}} V^T ((I - BX)(I - BX)^T - A) V S^{\frac{1}{2}}$, $G_i = S^{\frac{1}{2}} V^T Q_i$. Then the objective function with respect to matrix $\tilde{Q}$ can be reformulated as follows:

$$\begin{aligned} &\langle G_i \rangle = \arg\min_G \ tr\left({G_i}^T U G_i\right) \\ &s.t. \ \ {G_i}^T G_i = I \end{aligned} . \tag{10}$$

The matrices G_i can be solved by using the generalized eigen-value decomposition. The eigen-vectors of U are made up of the matrices G_i. Once we obtain G_i, we can calculate the matrices Q based on $G_i = S^{\frac{1}{2}} V^T Q_i$.

The entire optimization of our approach is summarized in Algorithm 1.

```
Algorithm 1. MSEA
Step 1: Randomly initialize the $\tilde{Q}$, $B$ and $X$;
Step 2: while $j < m$ (max iteration number) do:
    2.1 Updating the matrix $X$ in Formula (7)with MOD algorithm;
    2.2 Updating the matrix $B$ , by using the Formula (9);
    2.3 Updating the matrix $\tilde{Q}$, by using $Q_i = VS^{-\frac{1}{2}} G_i$,
    where $G_i$ is the eigen-vectors of $U$ in Formula (10);
Step 3: Output the sparse representation coefficients $X$, the embedding
matrices $W_i = {Q_i}^T {Y_i}^T$ and the shared dictionary $D = \tilde{W}\tilde{Y}B$.
```

3.3 Classification Strategy of MSEA

Given the learned D, W_i, and X, we design an effective classification strategy for MSEA. Specifically, we first project the testing samples into a novel feature space by using the multi-view embedding matrices W_i. Then, we employ the shared dictionary D to represent the projected features of samples, that is,

$$x = \arg\min_{x} \left\{ \|y - Dx\|_2^2 + \gamma \|x\|_1 \right\},$$

where x is the sparse representation coefficients. We classify the testing samples according to identity $(y) = \arg\min_{i} \{e_i\}$, where $e_i = \|y - D_i \alpha_i\|_2$ is the representation error of each class, and $\alpha_i = [\alpha_1, \alpha_2, \cdots, \alpha_c]^T$ is the sparse representation coefficients of each class. We classify the testing samples into the class with the smallest reconstruction error.

3.4 Kernelized MSEA

To tackle the linearly inseparable problem in the original feature space, we extend a kernelized extension of MSEA called KMSEA by using kernel trick. Kernel trick has shown its effectiveness in some methods [11, 12]. We first perform the kernel mapping for samples and then realize the MSEA in the mapped space.

Assume that $\phi : R^d \rightarrow F$ denotes a nonlinear mapping from the low-dimensional feature space to high-dimensional feature space. Then the mapping process from the sample set Y to space F can be represented as $Y \rightarrow \phi(Y)$. The objective function of KMSEA is defined as

$$\langle D, W, X \rangle = \arg \min_{D, \tilde{W}, X} \left\| \tilde{W} \phi\left(\tilde{Y}\right) - DX \right\|_F^2 - \gamma \left\| \tilde{W} \phi\left(\tilde{Y}\right) H \right\|_F^2. \tag{11}$$
$$s.t. W_i {W_i}^T = I, i = 1, 2, ..., N, and \|x_j\|_0 \leq T_0, \forall j$$

The optimization of KMSEA is similar to that of MSEA. We similarly introduce two matrices, $Q \in n \times p$ and $B \in n \times p$, and then the embedding matrices W_i and shared dictionary D can be represented by $W_i = {Q_i}^T \phi\left({Y_i}^T\right)$ and $D = \tilde{W}\phi\left(\tilde{Y}\right)B$, respectively. Substituting W_i and D into the Formula (11), we employ the kernel trick and then the objective function of KMSEA can be reformulated as

$$\begin{gathered} \arg\min_{D,\tilde{W},X} \left\|\tilde{Q}^T\tilde{K} - \tilde{Q}^T\tilde{K}BX\right\|_F^2 - \gamma\left\|\tilde{Q}^T\tilde{K}H\right\|_F^2 \quad , \\ s.t. {Q_i}^T\tilde{K}_iQ_i = I, i = 1, 2, ..., N, and \|x_j\|_0 \le T_0, \ \forall j \end{gathered} \tag{12}$$

where $\tilde{K}_i = \phi\left(\tilde{Y}_i\right)^T\phi\left(\tilde{Y}_i\right)$ is the RBF kernel trick. $\tilde{K} = \begin{pmatrix} \tilde{K}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tilde{K}_N \end{pmatrix}$, $\tilde{Q} = [Q_1, ..., Q_N]$. Similar to the MSEA, the KMSEA also can be solved by using the divide-and-conquer strategy. Its optimization process is similar to Algorithm 1.

4 Experiments

In this section, we evaluate our two approaches MSEA and KMSEA. We choose three state-of-the-art multi-view feature extraction methods, including the**TCA**[8], **UMDDL**[9], and **I^{2}SCA** [7], as the compared methods. We validate the effectiveness of our approaches through two aspects: the mean recognition rate and the sample distribution figure.

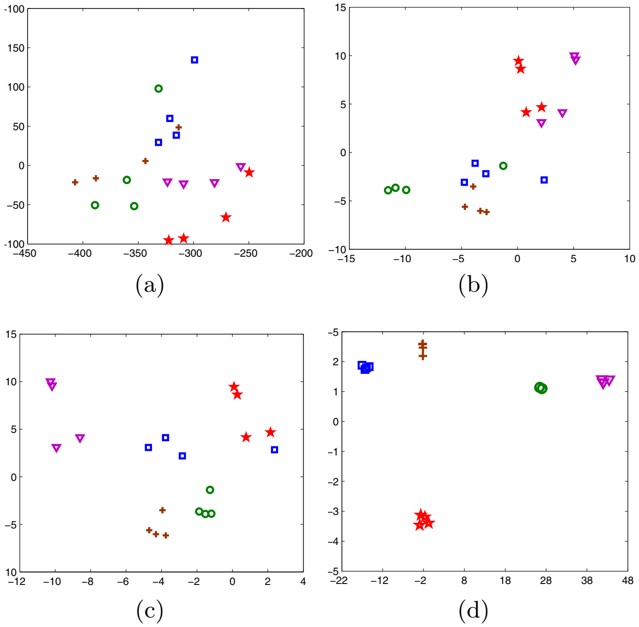
The experiments are conducted on two widely-used multi-view datasets. Multiple feature dataset (MFD) [13] contains 10 classes of handwritten numerals. These digit characters are represented in terms of six views of feature sets. In the experiment, we randomly choose 100 samples per class as the training set and the remaining 100 samples as the testing set. Multi-PIE dataset [14] contains various views, illumination and expressions variations. We choose its subset containing 1632 samples from 68 classes in 5 poses (C05, C07, C09, C27, C29). We randomly select 5 samples per class as the training samples and the remaining as the testing set.

Table 1 shows the average recognition rates and the standard deviation of 20 random runs for all methods on MFD and Multi-PIE datasets. We can observe that both MSEA and KMSEA outperform the compared methods on two datasets. Moreover, KMSEA obtains better performance than MSEA.

Table 1. Average recognition rate (± standard deviation) on two datasets.

Datasets	TCA	UMMDL	I^{2}SCA	MSEA	KMSEA
MFD	91.87±3.67	92.07±4.67	92.11 ±3.97	92.22± 4.23	**92.84±4.93**
Multi-PIE	91.87±4.56	92.53±4.02	92.91±3.46	93.51±3.38	**93.92±3.27**

In order to analyze the separabilities of all methods, we provide the distribution of samples with two principal features extracted from 5 different views by using all related methods on Multi-PIE dataset. Here, we employ the PCA transform to obtain two principal features. Note that since UMDDL is a dictionary learning method, not a feature extraction method, we cannot provide its sample distribution figure.



(a)

(b)

(c)

(d)

Fig. 2. Sample distributions of methods on Multi-PIE dataset of 20 samples in the feature space. (a): TCA; (b): I^{2}SCA; (c): MSEA; (d): KMSEA.

Fig. 2 shows the distribution of two principal features of 20 samples (from 5 different persons and 4 samples per person) extracted on the Multi-PIE dataset. The markers with different shapes and colors stand for 5 different persons. It shows that the proposed approaches achieve preferable separabilities in comparison with other methods. As for the MFD dataset, we obtain the similar results. Due to the limited space, we don't provide the results in detail here.

5 Conclusion

In this paper, we propose a novel multi-view feature extraction approach named MSEA. It not only can preserve the sparse reconstruction information, but also can consider the discriminative correlation in multi-view data. To remove the redundancy among views, we add orthogonal constraints of embedding matrices.

Furthermore, we provide a kernelized extension KMSEA to tackle the linearly inseparable problem. Experiments demonstrate that the MSEA and KMSEA outperform several state-of-the-art related methods with respect to the recognition rate and separabilities on sample distribution figures.

References

1. Guo, Y. H.: Convex subspace representation learning from multi-view data. In: AAAI Conference on Artificial Intelligence, pp. 387–393 (2013)
2. Long, B., Yu, P.S., Zhang, Z.M.: A general model for multiple view unsupervised learning. In: Proceedings of the SIAM International Conference on Data Mining, pp. 822–833 (2008)
3. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical Correlation analysis: An Overview with Application to Learning Methods. Neural Computation **16**(12), 2639–2664 (2004)
4. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(6), 1005–1018 (2007)
5. Gao, L., Qi, L., Chen, E.Q., Guan, L.: Discriminative multiple canonical correlation analysis for multi-feature information fusion. In: IEEE International Symposium on Multimedia, pp. 36–43 (2012)
6. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.: Multi-view discriminant analysis. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 808–821. Springer, Heidelberg (2012)
7. Jing, X.Y., Hu, R.M., Zhu, Y.P., Wu, S.S., Liang, C., Yang, J.Y.: Intra-view and inter-view supervised correlation analysis for multi-view feature learning. In: AAAI Conference on Artificial Intelligence, pp. 589–602 (2014)
8. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain Adaptation via Transfer Component Analysis. IEEE Transactions on Neural Networks **22**(2), 199–210 (2011)
9. Jing, X.Y., Hu, R.M., Wu, F., Liang, C., Yang, J.Y.: Uncorrelated multi-view Fisher discrimination dictionary learning for recognition. In: AAAI Conference on Artificial Intelligence, pp. 470–474 (2014)
10. Engan, K., Aase, S.O., Husoy, J.H.: Method of Optimal Directions for Frame Design. IEEE International Conference on Acoustics, Speech and Signal Process **5**(1), 2443–2446 (1999)
11. Sun, T.K., Chen, S.C., Jin, Z., Yang, J.Y.: Kernelized discriminative canonical correlation analysis. In: International Conference on Wavelet Analysis and Pattern Recognition, pp. 1283–1287 (2007)
12. Bach, F.R., Jordan, M.I.: Kernel Independent Component Analysis. Journal of Machine Learning Research **3**, 1–48 (2002)
13. Yuan, Y.H., Sun, Q.S., Zhou, Q., Xia, D.S.: A Novel Multiset Integrated Canonical Correlation Analysis Framework and Its Application in Feature Fusion. Pattern Recognition **44**(5), 1031–1040 (2010)
14. Cai, D., He, X., Han, J., Zhang, H.J.: Orthogonal Laplacianfaces for Face Recognition. IEEE Transactions on Image Processing **15**(11), 3608–3614 (2006)
15. Nguyen, H.V., Patel, V.M., Nasrabadi, N.M., Chellappa, R.: Sparse embedding: a framework for sparsity promoting dimensionality reduction. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 414–427. Springer, Heidelberg (2012)

Automatic Image Semantic Annotation Based on the Tourism Domain Ontological Knowledge Base

Pengfei Zhang[1], Junping Du[1(✉)], Dan Fan[1], and Yipeng Zhou[2]

[1] Beijing Key Laboratory of Intelligent Telecommunication Software
and Multimedia, School of Computer Science,
Beijing University of Posts and Telecommunications, Beijing 100876, China
{fly2015_zhang,komaconss}@163.com, junpingdu@126.com
[2] School of Computer and Information Engineering,
Beijing Technology and Business University, Beijing 100048, China
yipengzhou@163.com

Abstract. In this paper, we proposed a method of automatic image semantic annotation based on the tourism domain's ontological knowledge base. We need to do other things based on the traditional semantic annotation method. Firstly, we need to acquire the names of the scenic spots thorough image classification. Then we have to build ontological knowledge base on tourism domain and consider the annotation words and the names of the scenic spots as reasoning conditions. At last, we can use the ontological knowledge base to ratiocinate so as to enhance the accuracy of image annotation, and what's more, to associate annotation words with the name of scenic spot so that we can make annotation words more specific.

Keywords: Image annotation · Knowledge base · Tourism ontology · Image classification

1 Introduction

In recent years, with the development of cross media technology, the travel information we got from the internet is not only text information but also contains different types of data. It greatly enriched the source of knowledge and expands the perspective we understand about the tourism information. However, the content of the tourism images is so complicated that the primary problem of image semantic understanding is image semantic analysis. Since the 1980s, the study on the semantic of cross media data has began. Although the technology of text mining based on natural language understanding has made a great achievement, we still face unprecedented difficulties about text mining technology because of the limited feature we can mine [1]. Similarly, semantic learning and recognition of image is currently faced with the problem that how to cross the semantic gap [2-3]. Following is the basic methods of image semantic analysis and automatic annotation [4-5]. The first method is based on the content of cross media data [6]. The second method makes full use of the text information associated with visual data and transforms the problem of visual data into the

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 61–69, 2015.
DOI: 10.1007/978-3-662-48570-5_7

problem of text. For example, the Plsa-Words algorithm proposed by Monay belongs to the second method. The third method is automatic image annotation by fusing semantic topics [7]. All these methods depend only on visual data or text data or only can obtain basic elements of the picture and cannot obtain the content what we want.

However, due to the complexity of the tourism image data and there are rich semantic contents contained in images that the traditional annotation method cannot analyze the specific content in the image. So in this paper we propose the method for image annotation tourism based on the tourism domain's ontological knowledge base and do further keywords filtering on this basis. First, SVM-based image classification method was used to obtain the names of scenic spots. Then, build ontological knowledge base according to the information we got from scenic spots [8]. Finally, in order to obtained the specific content keywords and complete the secondary image annotation ,we consider the scenic name and label words as reasoning conditions and reasoning based on the prior knowledge base on the basis of image annotation. This method can be used to extract the specific content contained in the image and has very good effect on analyzing the semantic content of cross media data.

2 Frame Design of Automatic Image Annotation

In this paper, we propose a method of automatic image semantic annotation based on the tourism domain's ontology knowledge base to realize tourism image annotation. The method mainly consists of three parts: SVM-based classification [9], automatic image annotation by fusing semantic topics, and construction and reasoning of tourism domain's prior knowledge base. We consider the result of image classification as inference conditions and make up for the deficiency of the method that uses automatic image annotation by fusing semantic topics. That method can only analyze the obvious content contained in the images but cannot relate to the scenic spot. Good results can be obtained by knowledge base inference and the results we obtained will relate to scenic spot.

As shown in Fig.1, it's the frame of the method that uses automatic image semantic annotation based on the tourism domain's ontological knowledge base.

Following is the basic process of the method:

(1) Annotation for each image with the method that uses automatic image annotation by fusing semantic topics and we can get the keywords correctly represent the basic content in the image.

(2) Classified images according to the names of scenic spots by SVM-based image classification and consider the result we got as one of the inference conditions.

(3) Build ontological knowledge base on travel which contains the name of the scenic spot, location of the scenic spot, features of the scenic spot, entertainments activities in the spot and other information.

(4) Reasoning according to the names of scenic spots we obtained by image classification and the keywords we got by image annotation. By using knowledge base to ratiocinate, we can obtain the keywords which are related to the scenic spot and can express the specific content of the image.

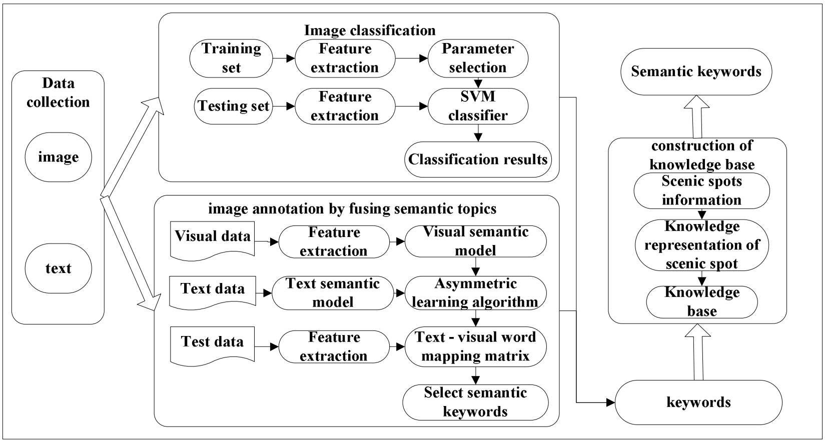


Fig. 1. Frame design of automatic image annotation

3 The Image Annotation Algorithm Based on Reasoning According to Ontological Knowledge

3.1 Automatic Image Annotations by Fusing Semantic Topics

In this paper we model the visual data and text data by probability latent semantic analysis (PLSA) and then we fuse the results of PLSA in which the visual data and text data share the same potential space. The way we fuse the model of visual data and text data is that fusing different distributions of themes we obtained by PLSA with a weight for each image and get a new kind of distribution of themes. The fusion weight of each model is determined by the contribution of the image content which is determined by the entropy of the distribution of visual words.

Suppose that the topic number of visual data is m and the topic number of text data is n, then the model after fusion contains k topics, and k=m+n. Using s and t to express the two topics of PLSA model, and then the topics distribution of visual data and text data could be expressed as $P_v(s|d)$ and $P_w(t|d)$. We can get two topics distribution $P_v(s|d_i)$ and $P_w(t|d_i)$ of every image by fusing the two PLSA models. And the topics distribution $P(z|d_i)$ after fusion was determined by the following formula:

$$P(z_k \mid d_i) = \begin{cases} \alpha_{vi} . P_v(s_k \mid d_i), k = 1,2,...,m \\ \alpha_{wi} . P_w(t_{k-m} \mid d_i), k = m+1, m+2, ..., m+n \end{cases} \tag{1}$$

Here, α_{vi} and α_{wi} represent the weights of visual data and text data respectively in the image d_i, and the weights can be calculated by the following empirical formula.

The experimental results show good annotation effect when the entropy of the visual words distribution is less than 3 or more than 6, however the effect is not always good when the entropy is between 3 and 6. This is because the images usually contains complex contents, and we cannot fully learn its complexity just rely on the entropy and empirical formula , that is to say, we are unable to determine the most reasonable weight of visual and textual modal data.

Description of the algorithm:

Suppose that there is a training set named D={(d_1,c_1),…,(d_N, c_N)} that contains the images and texts, and let T_D={d_1,…,d_N} denote the training set of images, and let L={w_1,…,w_L} denote the vocabulary list. So the images were included in the training set like $d_i \in T_D$ and the texts were included in vocabulary list like $c_i \subset L(i \in 1,…,N)$. In addition, we have to suppose that there is a testing set named T_T and $T_T \cap T_D = \emptyset$, $d_{new} \in T_T$.

Following is the description of the training algorithm which is used to model the data included the training set D and learns the association between image and text.

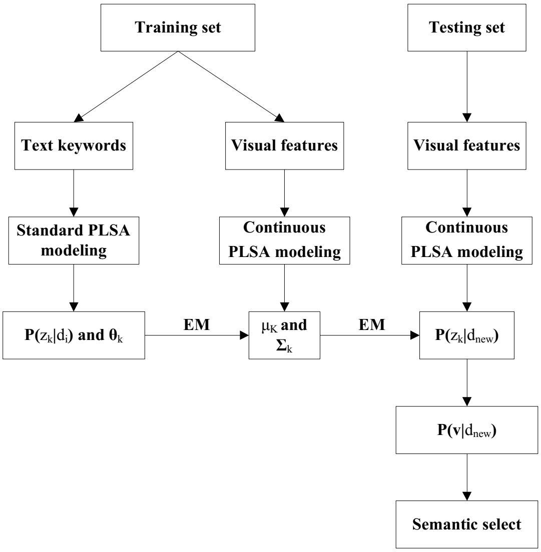


Fig. 2. Automatic image annotation algorithms by fusing semantic topics

(1) Extract the visual features from each image $d_i \in T_D$ and $d_{new} \in T_T$, quantized the features in order to denote the visual words with v(d_i). Similarly, process the text c_d associated with the image d_i and denote the text words with w(d_i).

(2) Fuse the results of the two PLSA models which are respectively based on the denotation of visual words v(d_i) and the denotation of text words w(d_i) and we can get the following results: $P_v(v|s)$, $P_v(s|d)$ and $P_w(w|t)$,$P_w(t|d)$.

(3) In order to measure the importance of the data in visual modality and textual modality, we introduce the fusion parameters α_{vi} and α_{wi}. Calculate the fusion parameters by empirical formula and we can get the result P(z|d_i) after fusing the topic distribution $P_v(s|d_i)$ and $P_w(t|d_i)$ by the formula (1).

(4) According to the fused topic distribution P(z|d_i),calculate the final training results P(v|z) and P(w|z) by using EM algorithm.

(5) Calculate the topic distributions $P(z|d_{new})$ of images with EM algorithm by using the denote of visual words $v(d_{new})$ of image d_{ne} and the parameter $P(v|z)$ which is the result of training algorithm.

(6) Calculate the posterior probability of every keyword in the vocabulary list L by the following formula.

$$P(w \mid d_{new}) = \sum_{k=1}^{k} P(w \mid z_k) P(z_k \mid d_{new}) \tag{2}$$

(7) Select several keywords with maximum posteriori probability to annotate the image d_{new}.

3.2 The Construction and Reasoning of Tourism Ontological Knowledge Base

In this paper, we achieve the further reasoning annotation of image annotation results by constructing tourism repository and make the results more associated with scenery spots and more accurate as well.

At first, we need to define the class structure of the ontology of tourism. In order to meet the requirements of this article, we only need to consider the traveling and entertainment in the tourism which is usually include eating, accommodation, transportation, traveling, shopping, and entertainment. And we also need to know the human and geography knowledge about the scenic spots. So we build four concept classes including scenic spots, scenic characteristics, geographic location, and characteristic activity. Among them, scenic spots usually include the Summer Palace, the Palace Museum, the Great Wall, the Temple of Heaven, the Olympic park and JiuZhaiGou. Scenic spot features include mountain, water, tree, sky, palace, construction, bridges, ships, tourists. Different scenic spots have certain difference. Characteristic activity includes some festival celebration activities and sports activities hold by some scenic spots, etc. The activities may also have obvious difference in different scenic spots.

Secondly, we need define the class attribute of tourism ontology. In order to complete the work in this paper, we need to define all the attributes of different ontology classes. The affiliation among scenic spots: for example “Kunming Lake” is sub attraction of “the Summer Palace”, the relationship between “Kunming Lake” and “the Summer Palace” is affiliation. The “locate in” relationship between scenic spots and geographical location: “the Summer Palace” is located in the “Haidian District of Beijing city”. The containment relationship between scenic spots and their features: for example, the features of the Summer Palace are mountains and water, “the Summer Palace” and “mountains” have the containment relationship. With the above relationships, we labeled the tourism ontological knowledge base manually and then we can complete the subsequent reasoning work based on these relationships.

There are two inputs of the algorithm: one is the test image, and the other one is annotation words list which contains 10 keywords that have larger posterior probability.

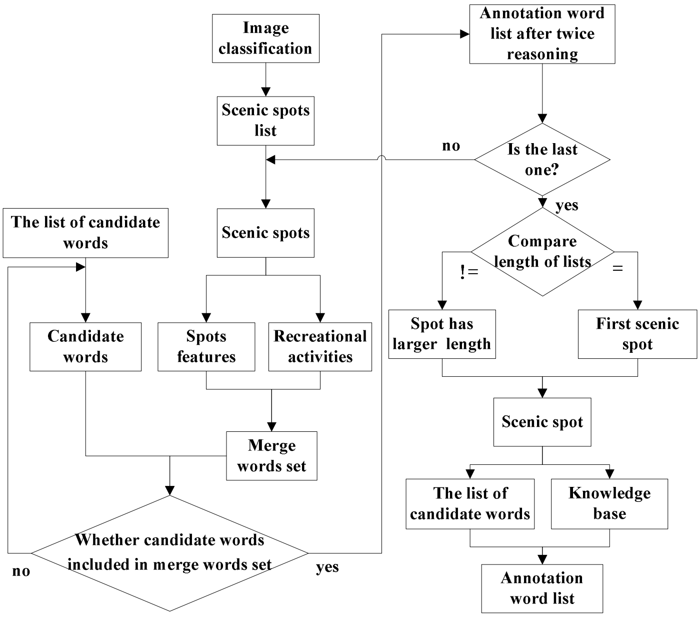


Fig. 3. The reasoning process

The algorithm process is as follows:

(1) Use image classification according to the scenic spots by SVM-based method and select two categories with larger membership degree of class label. Consider these two categories as a scenic spot list.

(2) Choose one spot sequentially from the scenic spot list and query spot features and recreational activities from the knowledge base according to the spot name.

(3) Traverse the candidate annotation word list and judge if a candidate annotation word is included in the spot features or recreational activities, if it is, add the word to the new word list ,and if not, read the next word.

(4) Judge whether the spot is the last one of the scenic spots list or not, if it is, execute step (5), and if not, return to step (2).

(5) Compare the length of the annotated word lists which were obtained according to different spots. If the length of all annotated word lists are equal, the first spot in the scenic spots list is the result, or we should choose the spot has longer annotated word list as the result.

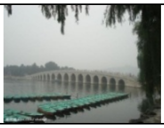
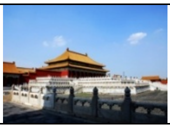
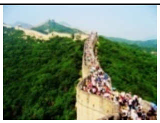
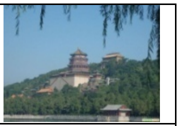
(6) Choose the spots corresponding to the candidate annotation word list and combine with the relation between different spots, the relation between spots and the features of all the spots and the relation between spots and recreational activities. Then ratiocinate to acquire the final annotation results.

4 Experimental Result Analysis

The experimental data is collected from Baidu tourism and Chanyouji blogs which include the travel data of the Summer Palace, the Great Wall, the Imperial Palace, the Temple of Heaven, the Olympic Park and so on, as well include the texts and pictures of their sub-attractions. The data is divided into a training set and a testing set. The former set data contains 3500 pictures of the Summer Palace, the Imperial Palace and their sub-attractions, while the latter data set contains 700 images which try to cover all the data included in the former data as much as possible.

We need to set the number of latent topics in the process of using PLSA to establish the main model, which determines the capacity of a PLSA model. If the topic number is too small, the PLSA model we obtained can't fully express the internal training of the data; on the contrary, if the topic number is too large, the efficiency of the system will be greatly reduced. What's more, with the increasing number of the unknown parameters of the PLSA model, the possibility of over fitting will increase. After the validation of experiments, this experiment used 100 latent topics to learn text modal data information and 120 potential subjects to learn the visual modal data. After the fusion of information, there are 220 potential themes obtained.

Table 1. Experimental Results of the two times image annotation algorithm

Figures				
Once annotation	bridge, boat, tree, water	building, sky, palace, mountain	Great Wall, tourists, mountain, water	mountain, Tower of Buddhist Incense, sky, water
Twice annotation	Seventeen Arches Bridge, Summer Palace, Kunming Lake, sky	sky, palace, Imperial Palace, Piled Elegance Hill	Great Wall, tree, sky, tourists	Tower of Buddhist Incense, Summer Palace, Longevity Hill, sky

From table 1 we concluded that after the second mark, the annotation results that we got seems to be more concrete, and it is associated with specific features in scenic spots and easier to understand. In terms of performance of the annotation, it can be measured with the accuracy of the annotation. For a given semantic keywords w_q, the accuracy was denoted as precision = B/A, that A means the number of all figures marked with w_q automatically; B indicates the number of images tagged with w_q correctly. In order to know whether the method we proposed is better than other methods, we annotate the test dataset with Plsa-Words algorithm. The accuracy comparison of the result for images annotation is showed in Fig. 4, which is fused with semantic theme and obtained from the tourism ontology knowledge inference. Table 2 shows the accuracy comparison of three annotation algorithms, where the one-time annotation represents automatic image annotation algorithm by fusing semantic topics and the two-time annotation represents the reasoning process based on knowledge after the one-time annotation.

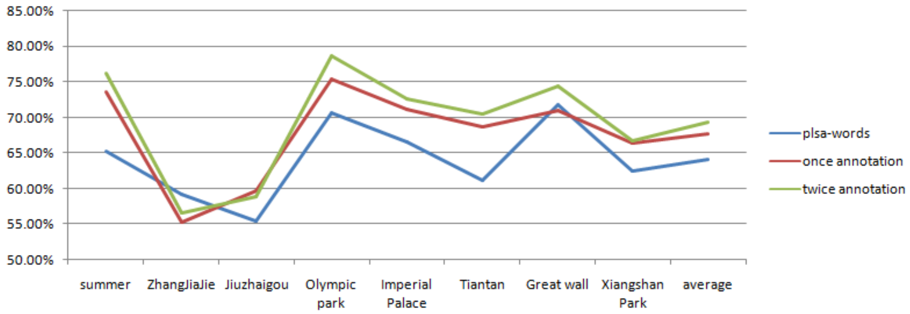


Fig. 4. The accuracy rate comparison of three annotation algorithms

Table 2. Contrast between the accuracy of three annotation algorithms

Scenic spots	Summer Palace	Zhang-JiaJie	JiuZhai-Gou	Olympic Park	Imperial Palace	Tian-Tan	Great Wall	Xiang-shan Park
Plsa-words	0.653	0.591	0.554	0.706	0.665	0.612	0.717	0.625
one-time	0.735	0.553	0.597	0.754	0.711	0.687	0.709	0.664
Two-time	0.762	0.566	0.589	0.787	0.726	0.705	0.744	0.667

With the inference method based on knowledge base, the average accuracy of image annotation is improved from 67.63% for one-time annotation to 69.33% for two-time annotation. Obviously, it’s better than 64.04% for Plsa-Words algorithm. However, from the graph we can see that the accuracy of all scenic spots is improved except ZhangJiaJie and JiuZhaiGou. Through analysis, the reason may relates to two points: on the one hand, these two scenic spots are both Natural scenic spot and the scenery elements are complex and contain many kinds of scenery which may lead to multiple results, thus affect the annotation results; on the other hand, the style of scenic spots has a great influence on the annotation results. For example, the different sub attractions have similar style in JiuZhaiGou and it is also very difficult to distinguish even for human. So it will inevitably have a certain impact on the annotation results.

5 Conclusions

Image annotation techniques are the key technology of image semantic analysis. And it plays an important role at the time when our mobile Internet has a high speed of development. Though there is a great progress about image annotation techniques in the world, there are still some limitations. In this study, we put forward an image annotation method relies on reasoning mechanism based on the tourism ontology repository which is on the basis of the image annotation method that fuses the semantic theme. This method combines the characteristics of tourism data, and has some innovations about the image annotation method the predecessor put up with. At last, we improve the accuracy of image annotation by using this method. And at the same time, by using the tourism ontology repository, the tagging results can combine with specific scenery spots instead of independent general content elements of image. However the accuracy of the results by using this method still has some room to improve. We can obtain more efficient experimental results by building a more complete training set, detailing the classification of the training set, expanding the tourism repository and improving the reasoning model.

Acknowledgments. This work was supported by the National Basic Research Program of China (973 Program) 2012CB821200 (2012CB821206), the National Natural Science Foundation of China (No. 61320106006), Beijing Excellent Talent Founding Project (2013D005003000009).

References

1. Liu, J.: Semantic Analysis and Classification of Food Safety Emergencies Cross Media Information. Beijing University of Posts and Telecommunications (2013)
2. Tang, J., Zha, Z., Tao, D., Chua, T.S.: Semantic-Gap-Oriented Active Learning for Multilabel Image Annotation. IEEE Transactions on Image Processing 21(4) (2012)
3. Bahmanyar, R., Datcu, M.: Measuring the semantic gap based on a communication channel model. In: Image Processing (ICIP) (2013)
4. Bakalem, M., Benblidia, N., Ait-Aoudia, S.A.: Comparative image auto-annotation. Signal Processing and Information Technology (ISSPIT) (2013)
5. Bao, H., Xu, G., Feng, S., Xu, D.: Advances in the technology of automatic image annotation. Computer Science (2011)
6. Li, X.: Technology research and analysis of massive image semantic retrieval. School of computer science and technology Zhejiang University (2009)
7. Li, Z., Si, Z., Liu, X.: Image semantic annotation method of modeling continuous visual features. Journal of Computer Aided Design & Computer Graphics (2010)
8. Heiyanthuduwage, S.R., Schwitter, R., Orgun, M.A.: Towards an OWL 2 profile for defining learning ontologies. In: Advanced Learning Technologies (ICALT) (2014)
9. Saxena, U., Goyal, A.: Content-based image classification using PSO-SVM in fuzzy topological space. In: Computer and Communication Technology (ICCCT) (2013)

Triple Online Boosting Training for Fast Object Detection

Ning Sun(✉), Feng Jiang, Yuze Shan, Jixin Liu, Liu Liu, and Xiaofei Li

Engineering Research Center of Wideband Wireless Communication Technology,
Ministry of Education, Nanjing University of Posts and Telecommunications,
Nanjing 210003, China
sunning@njupt.edu.cn

Abstract. In this paper, we present an online co-training method called Triple Online Boosting Training (TOBT). TOBT comprehensively combines the advantages of online boosting and tri-training to accomplish the function of online learning and sample validation at the same time. With the help of the novel feature extraction scheme named Fast Feature Pyramid (FFP) reported recently, we develop a real-time online method for multi-scale object detection. This method is proposed for detecting all-sized instances of a certain class in entire image, which is different from other online detectors for tracking purposes. Various experiments based no benchmark datasets and real videos demonstrate the efficacy of the proposed method in the respect of processing speed and stability for changing object appearance and scenarios.

Keywords: Object detection · Fast feature pyramid · Online boosting · Tri-training

1 Introduction

Sliding windows detectors are the most popular and powerful approaches in the field of object detection. This kind of detector is widely applied in the various environment, especially in the dynamical background scene, such as mobile or vehicle-mounted video system [1]. There are two main challenges of sliding windows based object detector involving real-scene automotive application. The first problem is computational cost for extracting the heterogeneous features to build multi-scale feature pyramid. The second one is the adaptive ability about variability of target appearance due to the change of illumination and different pose.

In decades, many research efforts have been devoted to improve the performance of object detection in the respect of processing speed [2, 3]. For instance, the groundbreaking work is the method of Viola and Jones [4], which develops a real-time (about 15 fps) face detector by boosting Haar like feature and integral image representation. After that, many successful sliding windows based methods are proposed for detecting multifarious target in various scenes, such as HOG, DPM and ChnFtrs and so on. However, these methods are unable to meet the need of real-time application due to heavy computational load. Recently, Dollár et al [5]

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 70–79, 2015.
DOI: 10.1007/978-3-662-48570-5_8

presents an ingenious speed-up scheme named Fast Feature Pyramid (FFP) for multi-scale feature extraction in 2009. Therefore, a more polished version of FFP was published in literature [6]. The key idea of FFP is that finely sampled pyramids may be obtained inexpensively by extrapolation from coarsely sampled ones, which prominently decreases the time-consuming of feature extraction procedure in the multi-scale object detection.

In order to deal with the problem about the generalization and stability of detector in real scene, the effective way is online learning mechanism. There are a lot of works focused on the online learning algorithm [7-10], most of which are successfully implemented in tracking application [11]. When we exploit continuous learning mechanisms in single task of object detection, no prior knowledge about size and position of targets help us to restrict the scope of sliding windows. The issue of automatic validation for image samples has to be addressed. Co-training [12] is a semi-supervised learning paradigm which trains two or more learners respectively from different views and lets the learners label some unlabeled examples for each other. Most previous theoretical analyses on co-training are based on the assumption that each of the views is sufficient to correctly predict the label. However, this assumption can hardly be met in real applications due to feature corruption or various feature noise [13-16]. Tri-training proposed by Zhou et al [14] in 2005 neither requires the instance space described with sufficient and redundant views nor does it put any constraints on the supervised learning algorithm.

The key contributions of this work can be summarized as follows:

1. We present novel online semi-supervised leaning algorithm called Triple Online Boosting Training (TBOT), which simultaneously achieves online learning of new features and sample images identification.
2. A fast multi-scale online object detection framework is developed through an asynchronous interactive mechanism with TOBT algorithm and fast feature pyramid scheme.
3. We design several experiments about pedestrian detection and tank detection based on the benchmark and real videos to evaluate performance of the proposed method. Experiments demonstrate that the proposed method achieves better performance than some previous object detection methods.

The remainder of this paper is organized as follows. We give overview of the proposed method and a more description of TBOT in Section 2. Experimental results on four different data sets compared to existing approaches are given in Section 3. Conclusions are presented in the last section.

2 The Proposed Method

2.1 Overview of Our Method

As shown in Fig. 1, the proposed method mainly consists of two modules: (1) fast detection (FD), (2) online verification and training (OVT). The detector in FD module is a binary classifier updated by module of OVT, which exploits the fast feature pyramid scheme to accelerate the procedure of multi-scale feature extraction to about five times the speed of traditional ones [2]. And, the image results regarded as object by fast detector are saved and delivered to the module of OVT. In OVT, we present a co-training algorithm called Triple Online Boosting Training (TOBT) to accomplish the function of online learning and sample validation at the same time. In TOBT, the image patches are input as ambiguous samples, and three pre-trained online boosting classifiers are initialized to start the tri-training scheme. Then, one ambiguous sample is labeled for a classifier if the other two classifiers agree on the labeling in each round. The iterative process continues until the three classifiers do not change. Finally, we build a cascaded classifier as fast detector using the weak hypotheses of the best one of the above-mentioned classifiers.

It should be pointed out that the operation of FD module and the OVT module is not synchronous. This is mainly due to the obvious different between the computational cost of fast detector and the one of another. And, it is not necessary that frequently updated the fast detector in a short time.

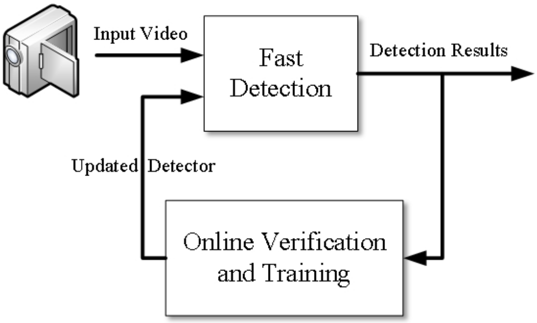


Fig. 1. The flowchart of the proposed method

2.2 The Fast Detector

The speed-up of the fast detector in our method is depended on the scheme of Dollár's Fast Feature Pyramid. When we built feature pyramid by this scheme, only feature data of one scale per octave is required to be computed precisely. And, the feature data of rest scale can be approximated by the ones of intermediate scales with minor loss in accuracy according to the exponential power law. In the respect of feature representation, we also follow the Aggregated Channel Features (ACF) in [6], which has been proven effective in general object detection method and suitable for FFP scheme.

2.3 Triple Online Boosting Training

The online learning is the pivotal function of our method. Meanwhile, it is very important to the performance of online learning that design a reliable verification module to identify the input image samples. We build a semi-supervised leaning algorithm named Triple Online Boosting Training (TOBT) to simultaneously achieve online learning and sample images identification, which comprehensively combine the benefits of online boosting and tri-training.

Table 1. Pseudo code of Triple Online Boosting Training

```
when TOBT is triggered by new input ambiguous image samples
Input: pre-trained classifiers $H_i^P, i = 1,2,3$; ambiguous image samples $A$ from fast detector; initial parameters $e_i^{'} = 0.5$, $l_i^{'} = 0$
Verification and online training phase:
  repeat until none of $H_i^P, i = 1,2,3$ changes
    for $i = 1$ $to$ $3$ do
      $L_i = NULL$; $Flag_i^u = FALSE$; $e_i = MeasureError(H_j \& H_k), (j,k \neq i)$
      if $(e_i < e_i^{'})$ then for every $x \in A$ do
                  if $(H_j(x) == H_k(x), (j,k \neq i))$ then $L_i = L_i \cup \{x, H_j(x)\}$
                  end for
      if $(l_i^{'} == 0)$ then $l_i^{'} = \lfloor e_i/(e_i^{'} - e_i) + 1 \rfloor$
      if $(l_i^{'} < |L_i|)$ then if $(e_i|L_i| < e_i^{'} l_i^{'})$ then $Flag_i^u = TURE$
                    else if $(l_i^{'} > e_i/(e_i^{'} - e_i))$ then
                           $L_i = SubSample\ (L_i, \lceil e_i^{'} l_i^{'}/e_i - 1 \rceil)$, $Flag_i^u = TURE$
      end for
      for $i = 1$ $to$ $3$ do
        if $(Flag_i^u == TURE)$ then $H_i = OnlineBoosting(L_i)$, $e_i^{'} = e_i$, $l_i^{'} = |L_i|$
      end for
  end repeat and choose
      $H_c = \arg\max(sum(H_i(x_m) == Lable(L_i(x_m)))), i = 1,2,3$
Building cascaded detector phase:
  $H_s = sort(H_c)$; $L = 0$;
  repeat until $H_s == NULL$
    $\omega = 0$
    for $h_n \in H_s$ do
      if $\omega > T$ then $\omega = 0$; $L = L + 1$; break
      else $\omega = \omega + \alpha_n$; $H_s = H_s - h_n$; $H_L = H_L + h_n$
    end for
  end repeat
  output $H_{final} = \bigcup_{m=1}^{L} H_m$
```

In the following, we describe the TOBT algorithm in detail. The TOBT is composed of two parts: (1) verification and online phase, (2) building cascaded detector phase. The pseudo code of TOBT is shown in Table 1. First of all, three classifiers, which are previously trained using online boosting [7] on three different sub-set of one training data set, and the training samples of the detection resulted from FD module in a certain period of time, are both prepared for starting the TOBT algorithm. In the verification and online phase, we mostly follow the scheme of Zhou's [14] tri-training expect that we use the validated image set L_i and online boosting algorithm to update corresponding classifier instead of using the original label sample set plus set L_i and any offline learning. After that, the best performance classifier H_c is chosen for building cascaded detector. At the beginning of building cascaded detector phase, the week hypothesis $h_n \in H_c$ is sorted in descending order based on the weight $\alpha_n = \ln((1-\varepsilon_n)/\varepsilon_n)$ of h_n, where ε_n is the error of h_n. Then, first layer of cascaded is combined with the best part of week hypotheses in H_s. The number of week hypothesis in each layer is limited by a pre-defined factor T. The process is repeated by filling the next layers with the remaining week hypothesis. After the levels are completed, their concatenation forms the cascade detector and updates the classifier in fast detection module.

3 Experiments and Discuses

We design several experiments on benchmark dataset and real scene video to test the performance of the proposed approach and compare the results with other the-start-of-are object detection methods: HOG [17] and ACF. The implementation of all test methods are based on C++ and OpenCV library except for the training of ACF, which uses directly the Matlab code of Dollár's toolbox[18]. And all experiments execute on the image processing server with Intel Xeon E7 4820 CPU and 32GB memory.

3.1 Person Detection

In this section, three detectors are training to detect persons for testing. Experiments are run on image sequence S3-T7-A View4 in PETS2006 and a real scene video named LIBRARY captured from the D1 resolution camera installed in the library of our campus (Fig.2). All three detectors are trained on the INIRA dataset [17].

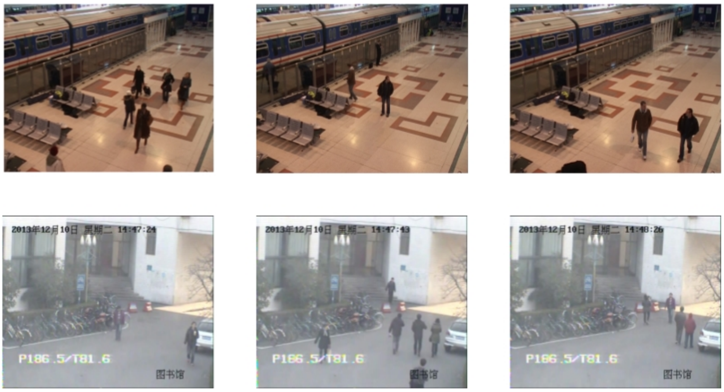

2013年12月10日 星期二 14:47:24
P186.5/T81.6
图书馆
2013年12月10日 星期二 14:47:43
P186.5/T81.6
图书馆
2013年12月10日 星期二 14:48:26
P186.5/T81.6
图书馆

Fig. 2. Test video in the pedestrian detection experiments. Top: test video PETS2006. Bottom: test video LIBRARY

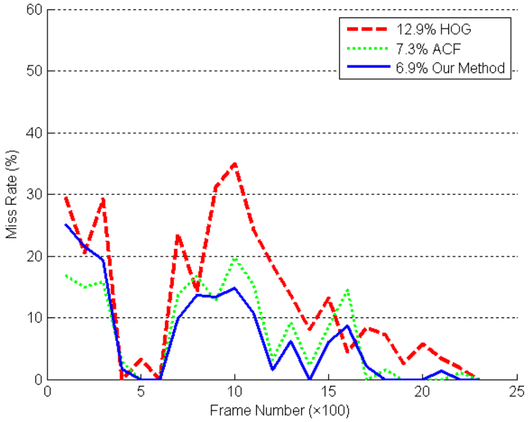


Fig. 3. Detection results on video PETS2006

Three detectors are run on the PETS2006 test video under the condition of one false positive per frame. According to our method, we trigger the TOBT algorithm to update fast detector every 100 frames. So, the average of miss rate per 100 frames was calculated to evaluate the accuracy of three detectors. As shown in the Fig.3, the average miss rate (See the legend in the top left corner of Fig.3) of our method is superior to one of other methods in the mostly period of test video, although the performance of our method is worse than that of ACF and HOG in few trigger periods at the beginning of video PETS2006. It is proved that the TOBT algorithm proposed in our method can continuously learn the new feature of pedestrian in the test video to improve the discriminability of the fast detector. Secondly, ACF based detectors (Our methods and ACF) achieve the outperforming detection results in the comparison with the HOG detectors, especially in the present of object partial occlusion. This result demonstrates that the multi-channel features are more discriminative that HOG feature, which is consistent with the study in literature [2].

The second test of pedestrian detection is applied on the real video called LIBRARY, which is consists of 1458 frames with the size of 720*576. The proposed method is also compared with detectors based on HOG and ACF under the same condition as the first test. In Fig.4, it can be found that the curves follow the similar trend as Fig.3, and the reduction of miss rate by the proposed method is more distinct than one in the first test. It is also shown that the TOBT algorithm can effectively improve the adaptability of detector for various application environments. Another similarity trend shown in Fig.3 and Fig.4 is that the miss rate of our method is higher than that in ACF in the first one or two round. This is caused by the different depth choice of decision tree in our method and ACF detector. We use one level decision tree (decision stump) in Online-boosting algorithm but two level depth decision tree in ACF. This kind of choice makes the proposed method faster in detection processing, which can be found in Fig.5.

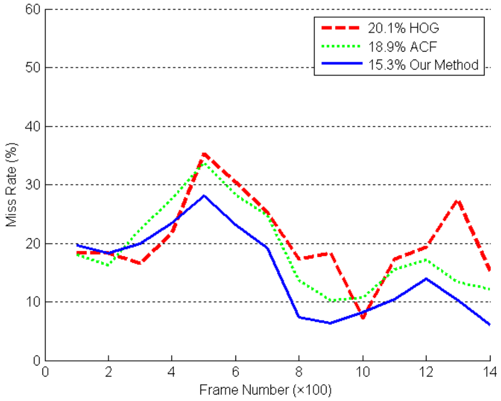


Fig. 4. Detection results on video LIBRARY

Moreover, we record the run-time of three detectors in the above-mentioned two tests. In Fig.5, we plot average miss rate versus runtime for three detectors. With D1 resolution video, the proposed method can achieve the detection speed at 28.5 fps, faster than the one of ACF at 22.2 fps and the one of HOG at 3.9 fps. It is mainly due to the reason that the weak hypothesis of the proposed method is the decision stump classifier, which is more efficient than the two level of decision tree of ACF detector.

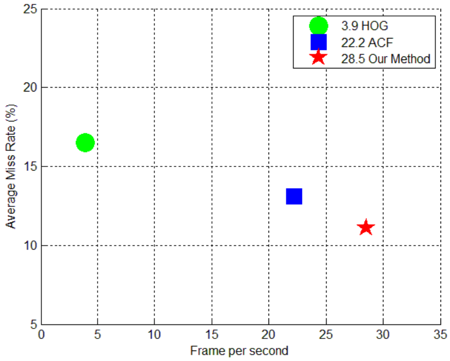


Fig. 5. Time versus miss rate of three detector

3.2 Tank Detection

After the experiments of person detection, we apply our method to detect tanks in the video, which is something more challenging task. Different from pedestrian detection, it is usually hard to gather the sufficient image samples of military equipment for training, just like main battle tank, belonging to enemy. The same as the above section, we compare the proposed method with HOG and ACF. The training samples and test video are both collected from internet. The training set is consist of 1370 image of tanks around the world except for the German Leopard 2 tank and 1500 negative samples randomly bootstrapped from landscape images (Fig.6). The test data is a 2089 frames video of Leopard 2 tank named TANK with D1 resolution (Fig.7). This experimental setting can assess the ability of online learning of the proposed method more effectively.

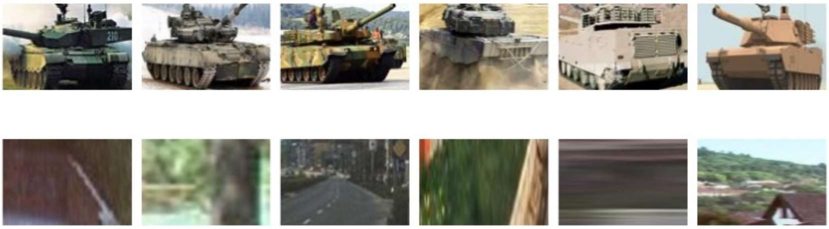


Fig. 6. The positive and negative samples of tank detection. Top: positive samples. Bottom: negative samples.

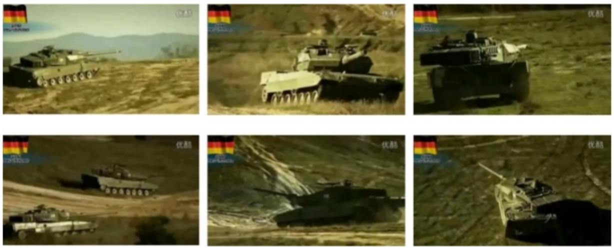


Fig. 7. The test video of tank detection

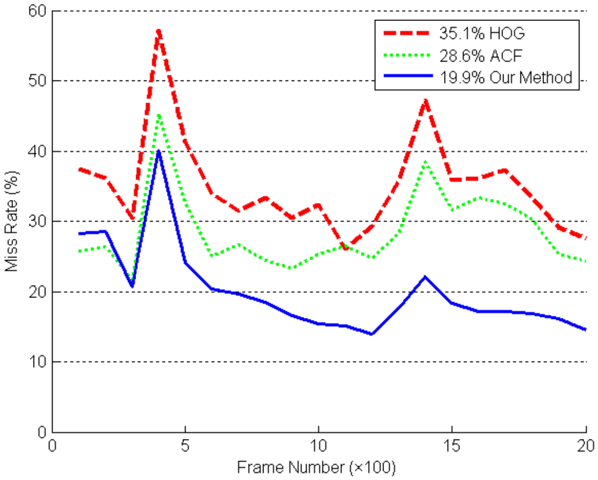


Fig. 8. Detection results on video TANK

Similar as the result of pedestrian detection experiments, the proposed method rapidly achieves the best accuracy of detection after three updating operation. And, the average miss rate of our method is lower than one of ACF above 30% (Fig.8). In the respect of run-time, the fps of three detectors keeps unchanged because that computational cost of detection is invariable with the sliding windows scheme under the same resolution video.

4 Conclusions

In this paper, the Triple Online Boosting Training (TOBT) algorithm combined with online learning and co-train is proposed for incrementally learning new features from autonomously invalidated image patches of detection results. Through an asynchronous interactive mechanism with TOBT algorithm and fast feature pyramid scheme, we build a real-time online method for universal object detection. Then, we design several experiments of pedestrian detection and tank detection based on the benchmark and real videos to evaluate performance of the proposed method and the other two the-state-of-art detectors. The comparison of experimental results prove the effectiveness of the proposed method in the field of accuracy and run-time.

Acknowledgments. This work was supported by the National Nature Science Foundation of China (61471206, 61401220) and Natural Science Foundation of Jiangsu province (BK20141428, BK20140884 and BK20140896).

References

1. Geronimo, D., Lõpez, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1239–1258 (2010)
2. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: An Evaluation of the State of the Art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2012)
3. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8926, pp. 613–627. Springer, Heidelberg (2015)
4. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 511–518 (2001)
5. Dollár, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. In: Proc. British Machine Vision Conf. (2010)
6. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast Feature Pyramids for Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1532–1545 (2014)
7. Grabner, H., Bischof, H.: On-line boosting and vision. In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, pp. 260–267 (2006)
8. Chang, W., Cho, C.: Online Boosting for Vehicle Detection. IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics **40**(3), 892–902 (2010)
9. Qi, Z., Xu, Y., Wang, L., Song, Y.: Online multiple instance boosting for object detection. Neurocomputing **74**, 1769–1775 (2011)

10. Visentini, I., Snidaro, L., Foresti, G.: Cascaded online boosting. Journal of Real-time image processing **5**(4), 245–257 (2010)
11. Wu, Y., Lim, J., Yang, M.-H.: Online object tracking: a benchmark. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2411–2418 (2013)
12. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proc. the 11th Annual Conference on Computational Learning Theory, pp. 92–100 (1998)
13. Goldman, S., Zhou, Y.: Enhancing supervised learning with unlabeled data. In: Proc. the 17th International Conference on Machine Learning (2000)
14. Zhou, Z., Li, M.: Tri-Training: exploiting unlabeled data using three classifiers. IEEE Trans. Knowledge and Data Engineering **17**(11), 1529–1541 (2005)
15. Xu, J., He, H., Man, H.: DCPE co-training for classification. Neurocomputing **86**, 75–85 (2012)
16. Li, M., Zhou, Z.: Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Trans. Systems, Man and Cybernetics **37**(6) (2007)
17. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2005)
18. Piotr's Computer Vision Matlab Toolbox. http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html

Structured Regularized Robust Coding for Face Recognition

Meng Yang(✉), Tiancheng Song, Feng Liu, and Linlin Shen

College of Computer Science and Software Engineering,
Shenzhen University, Shenzhen, China
yang.meng@szu.edu.cn

Abstract. The sparse representation based classifier (SRC) has been successfully applied to robust face recognition (FR) with various variations. To achieve much stronger robustness to facial occlusion, recently regularized robust coding (RRC) was proposed by designing a new robust representation residual term. Although RRC has achieved the leading performance, it ignores the structured information (i.e., spatial consistence) embedded in the occluded pixels. In this paper, we proposed a novel structured regularized robust coding (SRRC) framework, in which the spatial consistence of occluded pixels was exploited by pixel weight learning (PWL) model. Efficient algorithms were also proposed to fastly learn the pixel's weight and accurately recover the occluded area. The experiments on face recognition in several representative datasets clearly show the advantage of the proposed SRRC in accuracy and efficiency.

Keywords: Structure regularized · Robust coding · Face recognition

1 Introduction

Face recognition (FR) has been extensively studied in the past two decades [5], and many representative methods, such as Eigenfaces [6], Fisherfaces [6], LBP [7], have been proposed. In order to deal with facial occlusion, Eigenimages [8-9], probabilistic local approaches [10] and Markov random fields [19] were proposed for FR with occlusion. Although much progress have been made, robust FR to occlusion/disguise is still a challenging issue due to the variations of occlusion such as different categories of disguises, and the unknown intensity of occluded pixels.

Recently, sparse coding [1] and deep learning [17][18] have been widely applied to face recognition. Although deep learning has shown very promising accuracies, it still has some limitations, such as requirements of large amounts of training samples and super computational machines, and lacks of strong theoretical analysis and specific model for face recognition with various occlusions.

A successful work applying sparse coding to robust face recognition is sparse representation based classifier (SRC) [1], which was proposed for robust face recognition, producing very promising performance in FR with occlusion. By coding a query image $\boldsymbol{y}$ as a sparse linear combination of all the training samples via Eq. (1), SRC classifies $\boldsymbol{y}$ by searching for the class that produces the minimal reconstruction error.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 80–89, 2015.
DOI: 10.1007/978-3-662-48570-5_9

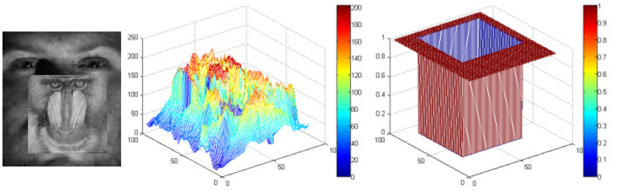
$$\min_{\boldsymbol{\alpha}} \left\|\boldsymbol{y}-\boldsymbol{X\alpha}\right\|_2^2+\lambda\left\|\boldsymbol{\alpha}\right\|_1 \quad (1)$$

where $\|.\|_1$ is the sparse l_1-norm and each column vector in $\boldsymbol{X}$ is a training sample. In order to make SRC robust to facial occlusion, an identity matrix $\boldsymbol{I}$ was introduced as a dictionary to code the outlier pixels (e.g., occluded pixels):

$$\min_{\boldsymbol{\alpha},\boldsymbol{e}} \left\|\boldsymbol{y}-\boldsymbol{X\alpha}-\boldsymbol{Ie}\right\|_2^2+\lambda\left\|\boldsymbol{\alpha}\right\|_1+\lambda\left\|\boldsymbol{e}\right\|_1 \quad (2)$$

By solving Eq. (2), SRC shows good robustness to face occlusions such as block occlusion and disguise. A theoretical support for the success of SRC may be that only a small part of pixels are occluded in most cases (So it is reasonable to require the representation residual $\boldsymbol{e}$ sparse).

It is easy to see in Eq. (2) that the representation residual, i.e., $\boldsymbol{e}$, is regularized by l_1-norm, which may not be optimal when the representation residuals do not follow a Laplacian distribution. Following SRC, He *et al.* [11] proposed a correntropy-based sparse representation (CESR) for robust face recognition, which introduced a Gaussian kernel-based fidelity term to regularize the coding residuals; and Gabor feature was also introduced in the framework of SRC to enhance its discrimination [12]. In order to deal with more general facial occlusion, Yang *et al.* [4] proposed a regularized robust coding (RRC) model by designing a robust representation term, which has shown the state-of-art performance in robust face recognition and attracted much attention in the field.



(a) (b) (c)

Fig. 1. The structured information of occluded image pixels. (a) a face image; (b) pixels' values; (c) pixels' occlusion patterns. It is easy to see occluded pixels' patterns but not the occluded pixels' values are spatial consistent.

Although RRC [4], CESR [11] and Gabor-SRC [12] have achieved leading performance in robust face recognition, all of them measure each pixel's representation residual independently with ignoring the structured information (i.e., spatial consistence) embedded in the 2D image space. In practical face recognition, most of occluded pixels are not independent but spatially consistent (e.g., illumination, expression, block occlusion, facial disguise). Here we should note that the spatial consistence is embedded in occluded pixels but not the occluded pixels' values. Fig. 1 gives an example to show the spatial consistence of image pixels and image pixels' values.

In this paper, we use a weight to indicate whether a pixel is occluded, then the structured information could be easily exploited in the pixel weight learning (PWL) without considering the difference among pixels' values. With the proposed PWL model, the structured information of image pixels could be effectively exploited and a novel framework of structured regularized robust coding (SRRC) was presented for robust face recognition. We also present efficient algorithms to solve PWL model. We evaluate the effectiveness of SRRC on several benchmark datasets, such as CMU Multi-PIE [21] and a joint face database of AR [13] and CAP-Peal [14]. The experiments on these datasets clearly show the advantage of SRRC in accuracy and effectiveness of robust face recognition.

The rest of this paper is organized as follows. Section 2 briefly reviews the related regularized robust coding model. Section 3 presents the proposed structured regularized robust coding framework. Section 4 conducts the experiments, and Section 5 concludes the paper.

2 Brief Review of Related Work

In order increase the robustness of SRC to various outliers, Yang *et al* [4] proposed a regularized robust coding (RRC) model, which was efficiently solved by using an iterative reweighted regularized coding algorithm. In each iteration RRC changes to

$$\min_{\boldsymbol{\alpha}} \left\| \mathrm{diag}(\boldsymbol{w})^{1/2} (\boldsymbol{y} - \boldsymbol{X\alpha}) \right\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_{l_p} \tag{3}$$

where l_p-norm on $\boldsymbol{\alpha}$ could be l_1-norn or l_2-norm ([2] indicated that the l_2-norm regularized coding could achieve similar accuracy to l_1-norm but with a faster speed), and diag($\boldsymbol{w}$) is a diagonal matrix with the weight vector $\boldsymbol{w}$ as its diagonal vector. Here the element of $\boldsymbol{w}$ is computed as

$$w_j = 1 \big/ 1 + \exp\left(\mu e_j^2 - \mu\delta\right) \tag{4}$$

where e_j=y_j-$\boldsymbol{r}_j\boldsymbol{\alpha}$, $\boldsymbol{r}_j$ is the j-th row vector of $\boldsymbol{X}$, and μ and δ are two automatically updated scalar parameters in the weight function [4]. Here w_j indicates the importance of the j-th element of $\boldsymbol{y}$ to the coding of $\boldsymbol{y}$. We can observe that the outlier pixels will have small weights to reduce their effects on the coding $\boldsymbol{y}$ on $\boldsymbol{X}$ since they have big residuals.

RRC could be solved by alternatively updating the weight vector $\boldsymbol{w}$ and the coding vector $\boldsymbol{\alpha}$. When the final coding vector $\boldsymbol{\alpha}$ is achieved, RRC conducts the classification via

$$\mathrm{identity}(\boldsymbol{y}) = \arg\min_i \left\| \mathrm{diag}(\boldsymbol{w})^{1/2} (\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\alpha}_i) \right\|_2^2 \tag{5}$$

where $\boldsymbol{X}_i$ the training samples of class i, $\boldsymbol{\alpha}$=[$\boldsymbol{\alpha}_1$; $\boldsymbol{\alpha}_2$;…; $\boldsymbol{\alpha}_c$], and $\boldsymbol{\alpha}_i$ is the coefficient vector associated with i-th disguise pattern.

3 Structured Regularized Robust Coding

Structured information (e.g., spatial consistence) is embedded in the pixels themselves but not the pixels' values. As shown in Fig.1, the nearby occluded pixels may have quite different values but their occluded patterns are same (e.g., their weight values, such as w in RRC, are same). Thus the proposed structured regularized robust coding (SRRC) could be represented as

$$\arg\min_{\boldsymbol{\alpha},\boldsymbol{w}} \left\{ \sum\nolimits_{i=1}^{n} \rho_{\theta} \left(y_i - \boldsymbol{r}_i \boldsymbol{\alpha} \right) + \lambda \left\| \boldsymbol{\alpha} \right\|_{l_p} \right\} \text{ s.t. } \boldsymbol{w} \text{ is structured regularized} \tag{6}$$

where $\boldsymbol{w}$ indicates weight values of image pixels, and the representation term is a robust fidelity term like RRC [4]. Here $\boldsymbol{w}$ could be reshaped to a weight map with the same size as face imges,

Similar to RRC, SRRC is solved by iteratively updating the coding vector $\boldsymbol{\alpha}$ and the weight value of each pixel. With known weight matrix $\boldsymbol{w}$, SRRC changes to the coding model of Eq. (3), which could be efficiently solved.

When the coding vector $\boldsymbol{\alpha}$ is known, the robust representation term, i.e., $\rho_{\theta} \left(y_i - \boldsymbol{r}_i \boldsymbol{\alpha} \right)$, could be represented by the approximation of its Taylor expansion (Please refer to the detailed Taylor expansion in [4]). So SRRC model changes to

$$\arg\min_{\boldsymbol{w}} \left\{ \left\| \boldsymbol{w} - \boldsymbol{w}^0 \right\|_{l_p} \right\} \text{ s.t. } \boldsymbol{w} \text{ is structured regularized} \tag{7}$$

where $\boldsymbol{w}^0 = \left[w_1^0 \ \cdots \ w_j^0 \ \cdots \right]$, $w_j^0 = 1 \big/ 1 + \exp\left(\mu e_j^2 - \mu\delta \right)$ is the estimated weight based on the Taylor expansion of $\rho_{\theta} \left(y_i - \boldsymbol{r}_i \boldsymbol{\alpha} \right)$, w_j^0 is the j-th element value of $\boldsymbol{w}^0$.

It can be easily seen that SRRC will degenerate to RRC if there is no structured regularization on $\boldsymbol{w}$. In order to make the robust representation model exploit the structured information, we present a pixel weight learning (PWL) model of Eq.(7) to introduce the structured information.

3.1 Pixel Weight Learning (PWL)

The structured information could be designed in many ways. In this paper, we only use the local consistence of image pixels as the structured information. Then the pixel weight learning (PWL) model could be rewritten as

$$\min_{\boldsymbol{w}} \left\| \boldsymbol{w} - \boldsymbol{w}^0 \right\|_{l_p} + \kappa \sum\nolimits_{i} \sum\nolimits_{j \in Ni} \left\| w_i - w_j \right\|_{l_p} \tag{8}$$

where k is a parameter to control the structured regularization, l_p-norm indicates the l_1-norm and l_2-norm when p=1 and p=2, respectively. For each pixel i, j is a neighboring pixels, and Ni is the set of neighboring pixels of pixel i. With the final term ensures the neighboring pixels have similar weight values. Here the neighboring size could be set by the users. A bigger neighboring region will introduce more global consistence. In this paper we use 4 neighborhoods for a pixel.

Different l_p-norm regularization will have different physical meanings. In order to make the learned $\boldsymbol{w}$ similar to $\boldsymbol{w}^0$,we use l_2-norm for the first term. When w_{pq}-$w_{pq'}$ is regularized by l_2-norm, Eq.(8) requires the weight map should be smooth, while l_1-norm regularized version could tolerate some sparse and sharp variance. Here we only consider the case that w_{pq}-$w_{pq'}$ is regularized by l_2-norm since we want the weight values be spatially consistent in general.

Thus the PWL model could be represented as

$$\min_{\boldsymbol{w}} \left\| \boldsymbol{w} - \boldsymbol{w}^0 \right\|_{l_p} + \kappa \sum\nolimits_i \sum\nolimits_{j \in Ni} \left\| w_i - w_j \right\|_2^2 \tag{9}$$

3.2 Solving Algorithm of PWL

In order to efficiently solve Eq.(9), we rewrite the final term of Eq.(9) as

$$\sum\nolimits_i \sum\nolimits_{j \in Ni} \left\| w_i - w_j \right\|_2^2 = \sum\nolimits_{j \in Ni} \boldsymbol{A}_j \boldsymbol{w} \tag{10}$$

where $\boldsymbol{A}_j$ is an indication matrix of the j-th neighboring pixel with all diagonal elements as 1s. For each row of $\boldsymbol{A}_j$ (i.e., each pixel in the image), the value of j-th neighboring pixel is set as -1, with all the elements as 0s. So $\boldsymbol{A}_j\boldsymbol{w}$ is a vector with each element as the difference of a pixel and its j-th neighboring pixel.

Denote $\boldsymbol{v}$=$\boldsymbol{w}$-$\boldsymbol{w}^0$, by replacing $\boldsymbol{w}$ as $\boldsymbol{v}$+$\boldsymbol{w}^0$ we rewrite the PWL model as

$$\min_{\boldsymbol{v}} \left\| \boldsymbol{v} \right\|_F^2 + \kappa \sum\nolimits_j \left\| \boldsymbol{A}_j \boldsymbol{v} + \boldsymbol{A}_j \boldsymbol{w}^0 \right\|_2^2 \tag{11}$$

In this case, we could derive an analytic solution, and the weight matrix solution could be presented as

$$\boldsymbol{v} = -\left(\kappa \sum\nolimits_j \boldsymbol{A}_j^T \boldsymbol{A}_j + \boldsymbol{I} \right)^{-1} \sum\nolimits_j \boldsymbol{A}_j^T \boldsymbol{A}_j \boldsymbol{w}^0 \tag{12}$$

Based on $\boldsymbol{v}$=$\boldsymbol{w}$-$\boldsymbol{w}^0$, we could further derive

$$\boldsymbol{w} = \boldsymbol{P}\boldsymbol{w}^0 \tag{13}$$

where $\boldsymbol{P}$=$\left(\kappa \sum\nolimits_j \boldsymbol{A}_j^T \boldsymbol{A}_j + \boldsymbol{I} \right)^{-1}$. Since the incidence matrix is predefined, the projection matrix $\boldsymbol{P}$ could be pre-computed and in testing time, only a projection operation with a low computation complexity is needed.

3.3 The Whole Algorithm of SRRC

Based on PWL, the whole algorithm of SRRC is summarized in Table 1.

Table 1. Algorithm of SRRC.

Solving algorithm of SRRC
1. Initialize $\boldsymbol{\alpha}$
2. Compute residual $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X\alpha}$.
3. Estimate weights $\boldsymbol{w}$ as via Eq.(4)
4. Weight updating $\boldsymbol{w}$ via PWL of Eq.(9)
5. Solve $\boldsymbol{\alpha}$ via the weighted regularized robust coding , i.e., Eq.(3)
6. Output $\boldsymbol{\alpha}$ until the condition of convergence is met, or the maximal number of iterations is reached.

After several iteration, we could get the final weight vector $\boldsymbol{w}$ and coding vector $\boldsymbol{\alpha}$, and then conduct face recognition via Eq.(5).

4 Experiments

We perform experiments on several benchmark datasets, such as CMU Multi-PIE [21], and a joint database [13][14] to demonstrate the performance of SRRC. In Section 4.1, we test SRRC on face recognition with illumination and expression variations; in Section 4.2 we compare the accuracies and running time on a joint face dataset. Here the joint database was constructed by using AR database (100 persons, 2599 images) [13] and a subset of CAS-Peal (101 persons and 843 images) [14]. For the experiments of face recognition without occlusion, we estimate the weight values of original face image and then use PCA to reduce the feature dimensionality like that in RRC [4].

In all experiments κ of PWL is set as 0.05 in face recognition without occlusion and 0.2 in face recognition with occlusion, respectively. The λ is set as the suggested value in RRC. The competing methods include the latest approaches, such as LLC [20], SRC [1], Gabor-SRC [12], CESR [11], RRC_L1 [4] and RRC_L2 [4]. Similar to RRC, SRRC_L1 and SRRC_L2 represent SRRC using l_1-norm and l_2-norm on $\boldsymbol{\alpha}$, respectively.

4.1 Face Recognition Without Occlusion

AR Database: As in [4], a subset (with only illumination and expression changes) that contains 50 male and 50 female subjects was chosen from the AR database [13] in this experiment. For each subject, the seven images from Session 1 were used for training, with other seven images from Session 2 for testing. The images were cropped to 60×43. The FR rates by the competing methods are listed in Table 2. We can see that SRRC could improve the performance of the second best method, RRC, in most cases. Especially, SRRC_L1 achieves the highest accuracy with visible improvement (e.g., 1.2% with 120-d feature).

Table 2. Face recognition rates on the AR database.

Dimension	54	120	300
NN	68.0%	70.1%	71.3%
SVM	69.4%	74.5%	75.4%
SRC [1]	83.3%	90.1%	93.3%
LLC [20]	80.7%	87.4%	89.0%
RRC_L$_2$	84.3%	94.3%	95.3%
SRRC_L2	84.4%	94.0%	95.9%
RRC_L$_1$	87.6%	94.7%	96.3%
SRRC_L1	**88.4%**	**95.9%**	**97.0%**

Multi PIE Database: The CMU Multi-PIE database [21] contains images of 337 subjects captured in four sessions with simultaneous variations in pose, expression, and illumination. Among these 337 subjects, all the 249 subjects in Session 1 were used for training. To make the FR more challenging, two subsets with both illumination and expression variations in Sessions 1 and 3, were used for testing. For the training set, as in [4] and [1], we used the 7 frontal images with extreme illuminations {0, 1, 7, 13, 14, 16, and 18} and neutral expression. For the testing set, 4 typical frontal images with illuminations {0, 2, 7, 13} and smile expressions (smile in Sessions 1 and 3) were used. Here we used the Eigenface with dimensionality 300 as the face feature for sparse coding. Table 3 lists the recognition rates in four testing sets by the competing methods.

From Table 3, we can see that SRRC_L1 achieves the best performance in all tests,. Compared to the baseline method, SRC, SRRC_L1 has 4.4% improvement in Smi-S1 and 16.3% in Smi-S3, respectively. Although the improvement of SRRC over RRC is not big, the introduction of structured information could still bring some benefits.

Table 3. Face recognition rates on Multi-PIE database. ('Smi-S1': set with smile in Session 1; 'Smi-S3': set with smile in Session 3).

	Smi-S1	Smi-S3
NN	88.7%	47.3%
SVM	88.9%	46.3%
SRC [1]	93.7%	60.3%
LLC [20]	95.6%	62.5%
RRC_L$_2$	95.9%	67.3%
SRRC_L2	96.2%	67.8%
RRC_L$_1$	97.8%	76.0%
SRRC_L1	**98.1%**	**76.6%**

4.2 Face Recognition on a Joint Face Database

In the test, we conduct FR with more complex disguises (e.g., sunglasses, scarf and hat) with variations of illumination and longer data acquisition interval. 340 images of the first 85 subjects (4 natural and non-occluded images with different illuminations in Session 1) in AR database and 263 images of the first 80 subjects (the non-occluded images) in CAS-Peal are used as the training sets. And 510 face images with sunglass and lighting variations, 510 face images with scarf and lighting variations, and 240 face images with hat and lighting variations are used as the testing dataset. Some samples are shown in Fig. 2.

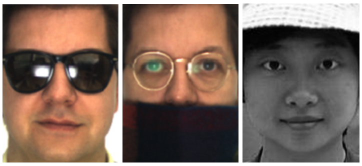


Fig. 2. The training and testing samples in the joint database.

Table 4. Recognition rates by competing methods on the joint database of AR and CAS-Peal with complex disguise occlusion.

Method	Sunglass	Scarf	Hat
SRC [1]	73.9%	24.9%	26.3%
GSRC [12]	52.4%	66.1%	34.2%
CESR[11]	80.2%	11.0%	26.7%
RRC_L_2	83.5%	75.3%	60.4%
SRRC_L2	87.4%	81.6%	71.3%
RRC_L_1	90.2%	77.3%	67.1%
SRRC_L1	**93.1%**	**83.3%**	**78.3%**

Table 4 lists the results of face recognition on the joint database by competing methods. Clearly, the SRRC methods achieve much better results than SRC, GSRC, CESR and RRC in most cases. RRC achieves the second best performance. SRRC_L2 outperforms RRC_L2 by 3.9%, 6.3% and 10.9% in face recognitions with sunglass, scarf and hat, respectively. SRRC_L1 outperforms RRC_L1 by 2.9%, 6.0%, and 11.2% in face recognitions with sunglass, scarf and hat, respectively;

Apart from recognition rate, computational expense is also an important issue for practical FR systems. In this section, the running time of the baseline method, SRC, and some competing methods which show not bad performance in all cases, including GSRC, RRC_L2, RRC_L1, and SRRC, is evaluated using the FR experiments on the joint face database. The programming environment is Matlab version R2013a. The desktop used is equipped with a 3.5 GHz CPU and 16G RAM. All the methods are implemented using the codes provided by the authors. For SRC, we use a fast l_1-minimization solver, ALM [15], to implement the sparse coding step.

Table 5 lists the average computational expense of different methods. We can observe that both SRRC_L2 and RRC_L2 have the least running time, followed by GSRC and SRC. Although the proposed SRRC has similar computation time to RRC, SRRC could achieve visibly better performance than RRC. Especially, SRRC_L1 has much better performance than RRC_L1 but with less running time.

Table 5. Average runnning time on the joint database with three facial disguises.

Method	Sunglass	Scarf	Hat
SRC (ALM)	0.610	0.579	0.574
GSRC	0.269	0.265	0.277
RRC_L_2	**0.177**	0.153	**0.171**
SRRC_L2	0.200	0.170	0.194
RRC_L_1	1.58	1.34	1.59
SRRC_L1	1.26	1.10	1.26

5 Conclusion

This paper presented a novel structured regularized robust coding (SRRC) framework and an associated pixel weight learning (PWL) model for robust face recognition. We also propose effective algorithms to solve the pixel weight learning model. One important advantage of SRRC is that the structured information (e.g., spatial consistence) could be exploited by the proposed SRRC with PWL. The proposed SRRC methods were extensively evaluated on FR with various variations, such as illumination, expression, random block occlusion, and real facial disgusie. The experimental results clearly demonstrated that SRRC outperforms previous state-of-the-art methods, such as SRC, CESR, GSRC and RRC.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China under Grants no. 61402289, 61272050 and 61403257, National Science Foundation of Guangdong Province (2014A030313558 and 2014A030313556), and Shenzhen Scientific Research and Development Funding Program under Grant JCYJ20140509172609171.

Reference

1. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Analysis and Machine Intelligence **31**(2), 210–227 (2009)
2. Zhang, L., Yang, M., Feng, X.C.: Sparse representation or collaborative representation: which helps face recognition? In: Proc. ICCV (2011)
3. Fidler, S., Skocaj, D., Leonardis, A.: Combining Reconstructive and Discriminative Subspace Methods for Robust Classification and Regression by Subsampling. IEEE Trans. Pattern Analysis and Machine Intelligence **28**(3), 337–350 (2006)
4. Yang, M., Zhang, L., Yang, J., Zhang, D.: Regularized robust coding for face recognition. IEEE Trans. Image Processing **22**(5), 1753–1766 (2013)
5. Zhao, W., Chellppa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Survey **35**(4), 399–458 (2003)
6. Belhumeur, P.N., Hespanha, J.P., Kriengman, D.J.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Analysis and Machine Intelligence **19**(7), 711–720 (1997)
7. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Trans. Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006)
8. Leonardis, A., Bischof, H.: Robust recognition using eigenimages. Computer Vision and Image Understanding **78**(1), 99–118 (2000)
9. Chen, S., Shan, T., Lovell, B.C.: Robust face recognition in rotated eigenspaces. In: Proc. Int'l Conf. Image and Vision Computing, New Zealand (2007)
10. Martinez, A.M.: Recognizing Imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. Pattern Analysis and Machine Intelligence **24**(6), 748–763 (2002)
11. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. IEEE Trans Pattern Analysis and Machine Intelligence **33**(8), 1561–1576 (2011)

12. 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, Part VI. LNCS, vol. 6316, pp. 448–461. Springer, Heidelberg (2010)
13. Martinez, A., Benavente, R.: The AR face database. CVC Tech. Report No. 24 (1998)
14. Gao, W., Cao, B., Shan, S.G., Chen, X.L., Zhou, D.L., Zhang, X.H., Zhao, D.B.: The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations. IEEE Trans. on System Man, and Cybernetics (Part A) **38**(1), 149–161 (2008)
15. Yang, A.Y., Ganesh, A., Zhou, Z.H., Sastry, S.S., Ma, Y.: A review of fast l1-minimization algorithms for robust face recognition (2010). arXiv:1007.3753v2
16. Jia, H., Martinez, A.: Support vector machines in face recognition with occlusions. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2009)
17. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Proc. Neural Information Processing Systems (2014)
18. Zhu, Z., Luo, P., Wang, X., Tang, X.: Multi-view perceptron: a deep model for learning face identity and view representations. In: Proc. Neural Information Processing Systems (2014)
19. Zhou, Z., Wagner, A., Mobahi, H., Wight, J., Ma, Y.: Face recognition with contiguous occlusion using markov random fields. In: Proc. IEEE Conf. Computer Vision (2009)
20. Wang, J.J., Yang, J.C., Yu, K., Lvx, F.J., Huangz, T., Gong, Y.H.: Locality-constrained linear coding for image classification. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2010)
21. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing **28**, 807–813 (2010)

Research and Application of Moving Target Tracking Based on Multi-innovation Kalman Filter Algorithm

Pinle Qin[1](✉), Guohong Lv[1], MaoMao Liu[1], Qiguang Miao[2], and Xiaoqing Chen[1]

[1] School of Computer and Control Engineering,
North University of China, Taiyuan 030051, China
qpl@nuc.edu.cn
[2] School of Computer Science and Technology, Xidian University, Xi'an 710075, China

Abstract. Moving objects tracking is an active problem in computer vision and has a wide variety of applications. The Kalman Filter algorithm has been commonly used for estimation and prediction of the target position in target tracking domain, of which the algorithm is adaptive to linear system, but the error of Kalman Filter will become large or even diverging when the target status changes suddenly. In this paper, multi-innovation theory is applied to target tracking, and the Multi-innovation Kalman Filter is proposed. Multi-innovation Kalman Filter has better precision and stability, because Multi-innovation Kalman Filter takes not only the moving targets' current state of motion into consideration, but also the time before. In addition, the authors theoretically analyzed the convergence of improved Multi-innovation Kalman Filter algorithm. Finally, simulation results show that the improved algorithm Multi- innovation Kalman Filter is superior to the traditional Kalman Filter.

Keywords: Target tracking · Kalman filter · Multi-innovation · Multi-innovation Kalman Filter · Simulation analyses

1 Introduction

With the development of computer technology, computer capacity has been greatly improved, and the target tracking has become one of the hottest topics at home and abroad [1]. It has been importantly and widely application in the fields of civilian, military, transportation and others [2, 3]. But how to accomplish correct and fast target tracking and how to reach good real-time performance and robustness are the key problems to be solved.

KF algorithm [4, 5] is widely used for target tracking, because KF can achieve the optimal estimation and better results of target tracking if the system equation, system noise and observation noise are all known [6, 7]. However, generally, the motion state of the observed object keeps a time-varying motion, instead of a uniform linear one. In this case, traditional KF predictive algorithm loses its superiority in target tracking, the tracking precision declines and the rate of convergence slows down. It may even result in losing track of the object.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 90–105, 2015.
DOI: 10.1007/978-3-662-48570-5_10

Considering the above problems of traditional KF, we generally use the adaptive filtering technique to solve them [8-12]. The authors of [8,9] come up with an improved adaptive KF algorithm, which introduces the forgetting factors based on fading memory index weighting, and gets the best forgetting factor by the method of forecasting residual error. At the same time, it also takes measures to ensure the semi-positive definiteness and positive definiteness of the noise estimation variance matrix and the measurement noise estimation variance matrix, thereby avoids the filter divergence. In [10], the method of KF target tracking, based on genetic algorithm, is proposed. The KF prediction is used to determine the candidate regions of target and then the genetic algorithm is used for searching and matching. As a result, the real-time performance and robustness of tracking is significantly improved. In order to reduce the evaluated error caused by changes in modeling, noise and so on, the authors of [11] and [12] put forward a KF tracking algorithm based on neural network. However, the amount of calculation of KF algorithm is relatively large. When it is combined with other algorithms, although the tracking precision is improved, the complexity and calculation of algorithm are increased. Therefore, it has difficulty in being applied to some real-time applications. Accordingly, it is very meaningful to design a simple and accurate time-varying moving target tracking algorithm.

Kalman filter algorithm is the classic algorithm in target tracking. Kalman filter algorithm is widely used in engineering because of its good real-time performance, so, raise the standard Kalman algorithm prediction accuracy in target tracking is necessary. Currently, many improved algorithms are put forward by combined other algorithm with Kalman algorithm to improve the accuracy of the algorithm, although the accuracy of the algorithm improves, complexity of the algorithm will be greatly increased. In this paper, we present an improved algorithm of tracking moving objects based on Multi-innovation theory, which is the Multi-innovation KF (MI-KF). MI-KF is improved on the basic algorithm, and there is no integration of the other new algorithm, the algorithm is simple, and it also had better real-time performance. At the end of the paper, it is used respectively in multiple types of movements (smooth movement and mutational movement), and the simulation results show that the improved algorithm MI-KF is superior to the KF, especially under the condition of the latter.

2 The Traditional Kalman Filter (KF)

KF is a state sequence of the dynamic system, and the algorithm is used for the linear minimum variance error estimation. The system is described by using dynamic state equation and observation equation [13]. The basic idea of KF is: Firstly, establishing a prior model for describing the stochastic and dynamic variables change over time; then under the condition of the real-time observation of random variables, using the Kalman equations in real time to get the best estimate of the target state that is based on global information. A recursive algorithm is used for the predicted value of time $k+1$ by that of time k, and ensures that the forecast error covariance is minimal. For linear systems, the KF algorithm can be used to predict target state easily and accurately, so it has been widely used in target tracking.

The state equation and measurement equation of discrete KF observer:

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1} \tag{1}$$

$$z_k = Hx_k + v_k \tag{2}$$

where x_k is the target state vector to be estimated, z_k is the measured value of the system, A is the state transition matrix, B represents the optional gain control input of u, a random variable w_k represents the process noise, and its covariance matrix is denoted by Q. In addition, $E\left[w_k\right]=0$ and $E\left[w_k w_l^T\right]=Q_k\delta_{kl}$. A random variable v_k represents the measurement noise, its covariance matrix is denoted by R, $E\left[v_k\right]=0$ $E\left[v_k v_l^T\right]=R_k\delta_{kl}$, δ is the function of $Kronec$ ker$-\delta$, and H is the observation matrix.

The basic equations of KF are as follows:

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1} \tag{3}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{4}$$

$$K_k = P_k^- H^T \left(HP_k^- H^T + R\right)^{-1} \tag{5}$$

$$\hat{x}_k = \hat{x}_k^- + K_k\left(z_k - H\hat{x}_k^-\right) \tag{6}$$

$$P_k = \left(I - K_k H\right)P_k^- \tag{7}$$

where (3) and (4) are the first step prediction and the first step prediction of error variance matrix separately, (5) is filter gain matrix, (6) is state correction, and (7) is the estimation of error variance matrix.

In Formula (3), the predicted value of time *k* is given by time *k*-1, and Formula (4) describes the effect of this prediction. Formula (5)-(7) are used to correct the estimated value of the previous correction. In Formula (6), $e\left(k\right)=z_k - H\hat{x}_k^-$ is known as innovation, and it is used to give feedback to the correction of the observed deviation. As long as the estimated value of initial state $\hat{x}\left(0\right)$ and the initial error variance matrix of filter state estimation $P\left(0\right)$ are given, KF will be started and proceeded recursively all the time.

3 The Improved KF Based on Multi-innovation Theory (MI-KF)

According to (6), namely the state correction equation, only one innovation exists in the traditional KF algorithm. The state prediction of time *k* is only estimated by the status time *k*-1, and the past data information are not made full use of, so that the useful information hidden in the past data is lost. Multi-innovation identification theory, proposed by Chinese scholar Ding Feng [14, 15], extends the scalar innovation to innovation vector, and the innovation vector to innovation matrix. As a result, he puts forward and sets up a kind of theory and method based on the idea of Multi-innovation, which is referred to as multi-innovation identification theory and method, and includes the multi-step information in the iterative process [16].

3.1 Multi-innovation Identification Theory

Some identification algorithms [17], such as least squares [18] and stochastic gradient algorithm [19], have a common characteristic: all of them use a single innovation correction technology that uses single innovation correction technology.

For the following scalar systems [20]:

$$y(k) = \varphi^T(k)\theta + v(k) \tag{8}$$

where $y(k) \in \mathrm{R}$ is the output of the system, $\varphi^T(k)$ is the information vector that is formed by input-output data of the system, θ is the vector parameter to be identified, and $v(k)$ is the system noise.

To estimate the parameter vector θ in Formula (8)

$$\hat{\theta}(k) = \hat{\theta}(\text{k-1}) + L(k)e(k) \tag{9}$$

where $L(k) \in R^n$ is the algorithm gain vector.

Innovation is an important quantity of recursion method, and it is used to describe output prediction error of time k. Innovation is defined as:

$$e(k) = y(k) - \varphi^T(k)\hat{\theta}(\text{k-1}) \in \mathrm{R} \tag{10}$$

In (9), $e(k) = y(k) - \varphi^T(k)\hat{\theta}(\text{k-1}) \in \mathrm{R}$ is the scalar, and it is called the single innovation. (9) shows that, we can reformulate parameter estimation vector $\hat{\theta}(k)$ of time *k* by the product of gain vector $L(k)$ and scalar innovation $e(k)$, which amends the estimated vector $\hat{\theta}(k-1)$ of time *k*-1. That means $\hat{\theta}(k)$ is calculated by adding the product of gain vector $L(k)$ and innovation $e(k)$ to $\hat{\theta}(k-1)$.

Innovation is different from the residuals, and the residuals are used to describe output deviation of time k. Residuals is defined as:

$$\varepsilon(k) = y(k) - \varphi^T(k)\hat{\theta}(k) \in \mathrm{R} \tag{11}$$

There is a relationship between the innovation and the residuals:

$$\varepsilon(k) = \frac{y(k)}{1 + \Lambda(k)\varphi^T(k)P(k-1)\varphi(k)} \tag{12}$$

or

$$\varepsilon(k)=\left[1-\Lambda(k)\varphi^{T}(k)P(k)\varphi(k)\right]y(k) \tag{13}$$

On the basis of single innovation, the scalar innovation $e(k)\in R$ is promoted as innovation vector $\boldsymbol{E}(p,k)=\left[e(k)\ ,\ e(k-1)\ ,\ \cdots\ ,\ e(k-p+1)\right]^{T}\in R^{p}$, which is the multi-innovation. To make the matrix multiplication dimension compatible, we extend gain vector $L(k)\in R^{n}$ to $\Gamma(p,k)\in R^{n\times p}$, then the multi-innovation identification algorithm turns out to be:

$$\hat{\theta}(k)=\hat{\theta}(k-1)+\Gamma(p,k)\boldsymbol{E}(p,t) \tag{14}$$

where $\Gamma(p,k)\in R^{n\times p}$ is the gain matrix, $E(p,k)\in R^{p}$ is the innovation vector, $p\geq 1$ is the Length of innovation. (14) shows that, in the multi-innovation identification algorithm, parameter estimation $\hat{\theta}(k)$ is corrected by the product of gain matrix $\Gamma(p,k)$ and innovation vector $\boldsymbol{E}(p,k)$, on the basis of parameter estimation $\hat{\theta}(k-1)$.

3.2 Multi-innovation KF (MI-KF)

Based on the above multi-innovation theory, we promote the standard KF algorithm, and extend the original single innovation to multi-innovation. In (6), the $e(k)=z_{k}-H\hat{x}_{k}^{-}$ is defined as the innovation, and we extend it to innovation matrix $E(p,k)$. In the same way, extend the measured values z_{k} to $Z(p,k)$, extend the measurement noise v_{k} to $V(p,k)$, extend the gain matrix K_{k} to $K(p,k)$, then:

$$V(p,t)=\begin{bmatrix} v_{k} \\ v_{k-1} \\ \vdots \\ v_{k-p+1} \end{bmatrix},Z(p,k)=\begin{bmatrix} z_{k} \\ z_{k-1} \\ \vdots \\ z_{k-p+1} \end{bmatrix},\ K(p,k)=\begin{bmatrix} K_{k} \\ K_{k-1} \\ \vdots \\ K_{k-p+1} \end{bmatrix},$$

$$E(p,k)=\begin{bmatrix} e(k) \\ e(k-1) \\ e(k-2) \\ \vdots \\ e(k-p+1) \end{bmatrix}=\begin{bmatrix} z_{k}-H\hat{x}_{k}^{-} \\ z_{k-1}-H\hat{x}_{k}^{-} \\ z_{k-2}-H\hat{x}_{k}^{-} \\ \vdots \\ z_{k-p+1}-H\hat{x}_{k}^{-} \end{bmatrix}\in R^{(mp)}$$

there, p is the multi-innovation length.

Then, measurement equation (2) will be changed into measurement matrix equation:

$$Z(p,k)=\begin{bmatrix} z_k \\ z_{k-1} \\ \vdots \\ z_{k-p+1} \end{bmatrix}=H(p,k)x_k+V(p,k) \tag{15}$$

By substituting (15) into (6), we can get the multi-innovation Kalman state prediction algorithm, and expand (6) to:

$$\begin{aligned} \hat{x}_k &= \hat{x}_k^- + \left[L_1(k),L_2(k),L_3(k)\cdots L_p(k)\right]E(p,k) \\ &= \hat{x}_k^- + \sum_{i=1}^{p} L_i(k)e(k-i+1) \end{aligned} \tag{16}$$

We obtain the equations of MI-KF as:

$$\begin{cases} \hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1} \\ P_k^- = AP_{k-1}A^T + Q \\ K_k = P_k^- H^T\left(HP_k^- H^T + R\right)^{-1} \\ \hat{x}_k = \hat{x}_k^- + \left[L_1(k),L_2(k),L_3(k)\cdots L_p(k)\right]E(p,k) \\ P_k = \left(I - K_k H\right)P_k^- \end{cases} \tag{17}$$

In (17), we get the traditional KF when $L_1(k)=L(k)$, $L_2(k)=L_3(k)=L_4(k)=\cdots=L_p(k)=0$.

The improved MI-KF proposed in this paper compared with the traditional KF, MI- KF has the following advantages:

(a) In each step of parameter estimation, standard KF only uses the innovation at k-1 moment, but the improved MI-KF not only uses the innovation at k-1 moment, but also uses the useful data and innovation in the past, so that the algorithm convergence is improved.

(b) When predicted the state of time k, MI-KF uses useful data of time k-1and time k-2. Similarly, when predicted the state of time k+1, MI-KF uses useful data of time k and time k-1. Therefore, when predicted the state of two adjacent time (time k and time k+1), it uses the useful data and innovation at k-1 moment repeatedly, and this is the main reasons for improving the MI-KF algorithm accuracy.

4 Convergence Analysis of the Improved MI-KF Algorithm

In order to analyze the convergence of proposed algorithm, it is necessary to introduce the following theorem:

Theorem 1. About system equations (1), (2) and the improved MI-KF algorithm, suppose (that) the system measurement noise v_k is white noise, and the mean and the variance of v_k are zero and σ^2.

(A1) $E(v_k)=0$, $E\left(\|v_k\|^2\right)=\sigma_k^2=\sigma^2$

If there exist constants $0<\alpha\leq\beta<\infty$ and the innovation length $p\geq n$ such that the following persistent excitation condition holds,

(A2) $\alpha I\leq\frac{1}{p}\sum_{i=1}^{p}\varphi(k-i+1)\varphi^T(k-i+1)\leq\beta I$, a.s.,

Then, the parameter estimation root-mean-square error of improved MI-KF algorithm is bounded.

Proof. Define the estimation error:

$$\tilde{x}_k=\hat{x}_k-x_k$$

there, x_k is real value.

From formulas (15) and (17), we can obtain:

$$\begin{aligned}
\tilde{x}_k&=\hat{x}_k-x_k\\
&=\hat{x}_k^-+K(p,k)\left[Z(p,k)-H(p,k)\hat{x}_k^-\right]-x_k\\
&=\hat{x}_k^-+K(p,k)\left[Z(p,k)-H(p,k)\hat{x}_k^--\frac{1}{K(p,k)}x_k\right]\\
&=\hat{x}_k^-+K(p,k)\left[H(p,k)x_k+V(p,k)-H(p,k)\hat{x}_k^--\frac{1}{K(p,k)}x_k\right]\\
&=\hat{x}_k^-+K(p,k)\left[H(p,k)\left(x_k-\hat{x}_k^-\right)+V(p,k)-\frac{1}{K(p,k)}x_k\right]\\
&=\hat{x}_k^-+K(p,k)\left[-H(p,k)\tilde{x}_k^-+V(p,k)-\frac{1}{K(p,k)}x_k\right]\\
&=\tilde{x}_k^-+K(p,k)\left[-H(p,k)\tilde{x}_k^-+V(p,k)\right]\\
&=\tilde{x}_k^-\left[I-K(p,k)H(p,k)\right]+K(p,k)V(p,k)
\end{aligned}\tag{18}$$

Then, we calculate the norm on both sides of the equation (18), and use the inequality: $(x+y)^2\leq(1+a)x^2+\left(1+\frac{1}{a}\right)y^2, a>0$. Formula (18) is transformed as follows:

$$\begin{aligned}
\|\tilde{x}_k\|^2 &\le (1+a)\left\|\left[I-K(p,k)H(p,k)\right]\tilde{x}_k^-\right\|^2 + \left(1+\frac{1}{a}\right)\|K(p,k)V(p,k)\|^2 \\
&\le (1+a)\|I-K(p,k)H(p,k)\|^2\left\|\tilde{x}_k^-\right\|^2 + \left(1+\frac{1}{a}\right)\|K(p,k)\|^2\|V(p,k)\|^2 \\
&\le (1+a)\left\|\tilde{x}_k^-\right\|^2\|I-K(p,k)H(p,k)\|^2 + \left(1+\frac{1}{a}\right)p\sigma^2\|K(p,k)\|^2 \\
&\le (1+a)\left\|\tilde{x}_k^-\right\|^2\left[I+\|K(p,k)\|^2\|H(p,k)\|^2\right] + \left(1+\frac{1}{a}\right)p\sigma^2\|K(p,k)\|^2 \\
&\le (1+a)\left\|\tilde{x}_k^-\right\|^2 + (1+a)\left\|\tilde{x}_k^-\right\|^2\|K(p,k)\|^2\|H(p,k)\|^2 + \left(1+\frac{1}{a}\right)p\sigma^2\|K(p,k)\|^2
\end{aligned}$$

Because $H(p,k)$ remains the same in the simulation experiments, and let's take it to be a constant *h*, then, $\|H(p,k)\|^2 = h$. So, the above equation becomes:

$$\begin{aligned}
\|\tilde{x}_k\|^2 &\le (1+a)\left\|\tilde{x}_k^-\right\|^2 + (1+a)h\left\|\tilde{x}_k^-\right\|^2\|K(p,k)\|^2 + \left(1+\frac{1}{a}\right)p\sigma^2\|K(p,k)\|^2 \\
&\le (1+a)\left\|\tilde{x}_k^-\right\|^2 + \left((1+a)h\left\|\tilde{x}_k^-\right\|^2 + \left(1+\frac{1}{a}\right)p\sigma^2\right)\|K(p,k)\|^2
\end{aligned} \tag{19}$$

Calculate the mathematics expectation of formula (19):

$$\begin{aligned}
E\left(\|\tilde{x}_k\|^2\right) &\le E\left((1+a)\left\|\tilde{x}_k^-\right\|^2\right) + E\left(\|K(p,k)\|^2\right)E\left[(1+a)h\left\|\tilde{x}_k^-\right\|^2 + \left(1+\frac{1}{a}\right)p\sigma^2\right] \\
&\le (1+a)E\left(\left\|\tilde{x}_k^-\right\|^2\right) + E\left(\|K(p,k)\|^2\right)\left[\left(1+\frac{1}{a}\right)p\sigma^2 + h(1+a)E\left(\left\|\tilde{x}_k^-\right\|^2\right)\right]
\end{aligned} \tag{20}$$

$\lambda_{\max}$ is the largest eigenvalue of $K(p,k)$. On the basis of (A2), the following formula was established:

$$E\left(\|K(p,k)\|^2\right) \le E\left[\lambda_{\max}\cdot K(p,k)\cdot K^T(p,k)\right] \le p\beta \tag{21}$$

We Substitutes formula (21) into formula (20):

$$E\left(\|\tilde{x}_k\|^2\right) \le (1+a)E\left(\left\|\tilde{x}_k^-\right\|^2\right) + p\beta\left[\left(1+\frac{1}{a}\right)p\sigma^2 + h(1+a)E\left(\left\|\tilde{x}_k^-\right\|^2\right)\right] \tag{22}$$

Let the assumption stand: $T_k = E\left(\|\tilde{x}_k\|^2\right)$, then $T_{k-1} = E\left(\left\|\tilde{x}_k^-\right\|^2\right)$, according to (22), it is easy to get:

$$\begin{aligned} T_k &\leq (1+a)T_{k-1} + p\beta\left[\left(1+\frac{1}{a}\right)p\sigma^2 + h(1+a)T_{k-1}\right] \\ &\leq \left[1+a+p\beta h(1+a)\right]T_{k-1} + p^2\beta\sigma^2\left(1+\frac{1}{a}\right) \end{aligned} \tag{23}$$

Assuming $1+a+p\beta h(1+a)=M$, $p^2\beta\sigma^2\left(1+\frac{1}{a}\right)=N$, then formula (24) turn into:

$$T_k \leq MT_{k-1} + N = M^{k-1}T_1 + \frac{1-M^{k-1}}{1-M}N \tag{24}$$

The proof is completed.

5 Experimental Results

The purpose of the simulation is to demonstrate that the improved target status prediction methods, proposed in this paper, is superior to the traditional method, especially when the speed of moving object changes suddenly. For different moving patterns of the target (smooth motion, abrupt motion), we conducted the Matlab simulations respectively. By comparing the experiment, we concluded that the accuracy of prediction of the target state in the multi-innovation KF Algorithm is comparable or better than the standard KF Algorithm. At the same time, we used the root mean square error (RMSE) to undertake the quantitative analysis. In multi-innovation KF algorithm, we added another innovation to the standard Kalman prediction algorithm in this paper, namely in the concrete simulation experiments, the improvement of multi-innovation Kalman algorithm contains two innovations.

5.1 Slow Moving Target Tracking

In the simulation experiment, the KF algorithm parameters should be initialized at first.

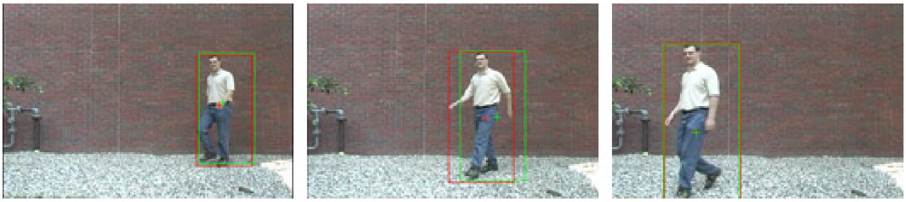
State transition matrix A is given as:

$$A=\left[[1,0,0,0];[0,1,0,0];[1,0,1,0],[0,1,0,1]\right]$$

and the transfer matrix of observation H as:

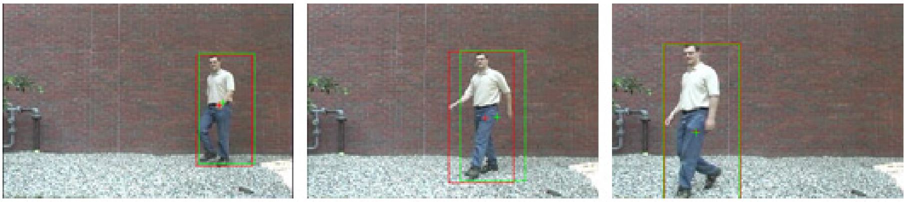
$$H=\left[[1,0];[0,1];[0,0];[0,0]\right]$$

Fig. 1 and 2 show the target tracking results, and they also show the tracking performance of some frames. Fig. 1 is the tracking performance of standard Kalman prediction algorithm. Fig. 2 is the tracking performance of improved multi-innovation Kalman prediction algorithm. The red rectangle box represents the position of the moving target, and the green rectangle shows the predicted position of the moving target. Comparing the tracking effect of two algorithms, we can see that the improved algorithm is better than standard Kalman algorithm. Fig. 3 is the trajectory of the moving target centroid in the video. In Fig.3, the red line represents the actual observation track of moving targets, and the green line represents the predicted trajectory of standard Kalman algorithm, and the black line represents the predicted trajectory of the improved multi-innovation Kalman forecasting algorithm. Similarly, in the enlarged part of the track curves, the paper improved algorithm MI-KF (black line) is closer to the real value of the predicted value (red line).



Frame 6 Frame 17 Frame 40

Fig. 1. The tracking effect of KF



Frame 6 Frame 17 Frame 40

Fig. 2. The tracking effect of MI-KF

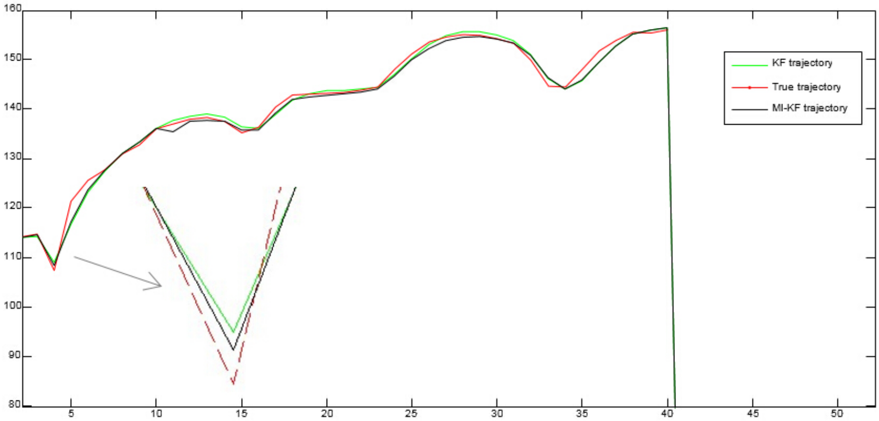


Fig. 3. Real trajectoriy and predicted trajectory of the moving target

In each frame of the video image, we define the real position of the centroid as $M(x_1, y_1)$, and the predicted position as $N(x_2, y_2)$, then the distance between the real position and the predicted position is $L_i = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, the root mean square error is $RMSE = \sqrt{\sum_{i=1}^{m} L_i^2 / m}$ (m is the total number of video frames). Fig. 4 shows the distance between the predicted position and the true location of moving target centroid in the video image. From the enlarged portion in Fig. 3 and Fig. 4, it can be seen that when the speed of the moving target is not stable and changes, the predicted performance of the multi-innovation Kalman algorithm is more accurate.

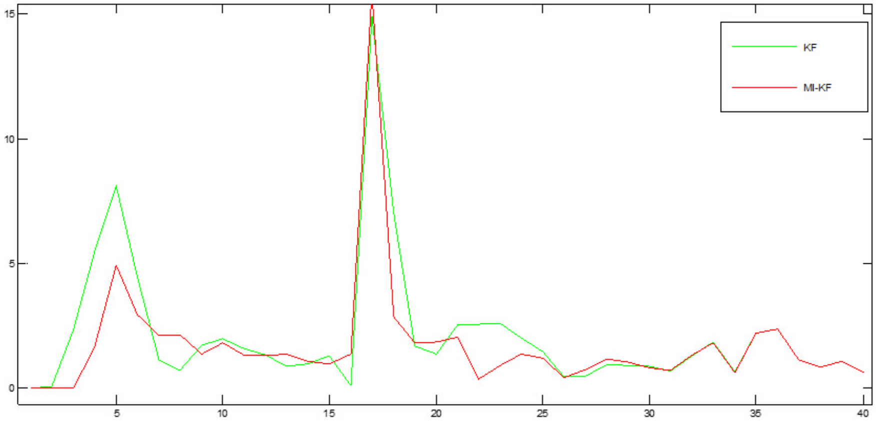


Fig. 4. Distance from the actual position and forecast position of the moving target

Table 1 shows the RMSE between the standard Kalman prediction tracking algorithm and the improved multi-innovation Kalman prediction algorithm. We can visually see from the table, the improved algorithm RMSE than the standard algorithm has decreased, and therefore a higher prediction accuracy of the algorithm.

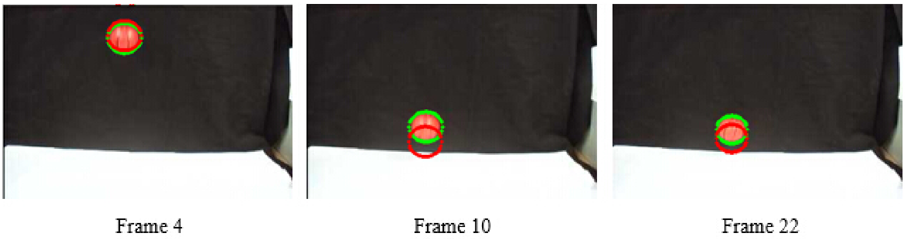
Table 1. RMSE of the two methods of tracking

Filtering algorithm	KF	MI-KF
RMSE	3.4051	3.2315

5.2 Abrupt Moving Target Tracking

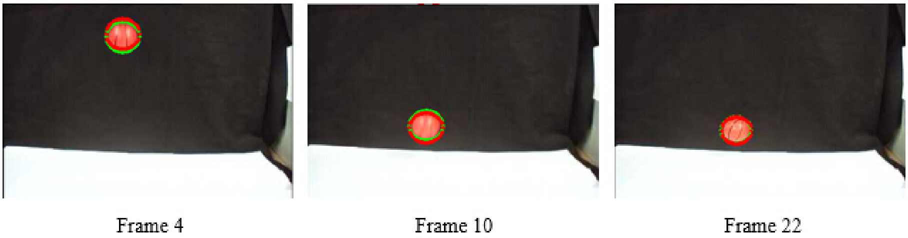
To further demonstrate that the improved multi-innovation Kalman prediction algorithm has better results when the target has mutational status. We conducted a second set of experiments, and selected the bouncing ball as the tracking target. The initialization of the MI-KF remains consistent with KF.

Fig. 5 and 6 show the ball tracking results of KF and MI-KF separately. Fig. 5 is the tracking performance of standard Kalman prediction algorithm. Fig. 6 is the tracking performance of improved multi-innovation Kalman prediction algorithm. The green circle represents the real position of the moving ball, and the red circle shows the predicted position of the ball. Comparing the tracking effect of two algorithms, we can see that the improved algorithm is better than standard Kalman algorithm, especially when the ball bounces (frame 10 and frame 22). Fig. 7 is the trajectory of the ball centroid in the video. In Fig.7, the red line represents the actual observation track of the ball, the green line represents the predicted trajectory of standard Kalman algorithm, and the blue line represents the predicted trajectory of the improved multi-innovation Kalman forecasting algorithm. The enlarged portion in Fig. 7 depicts that, when the ball bounces, MI-KF has a better result than KF. Fig. 8 shows the distance between the predicted position and the true location of ball centroid in the video image. We can see from Figure 8, when the whole tracking system is stable, the improved algorithm MI-KF predicted target position and the instance from the actual position than the standard predicted position KF smaller, higher forecast accuracy.



Frame 4 Frame 10 Frame 22

Fig. 5. The tracking effect of KF



Frame 4 Frame 10 Frame 22

Fig. 6. The tracking effect of MI-KF

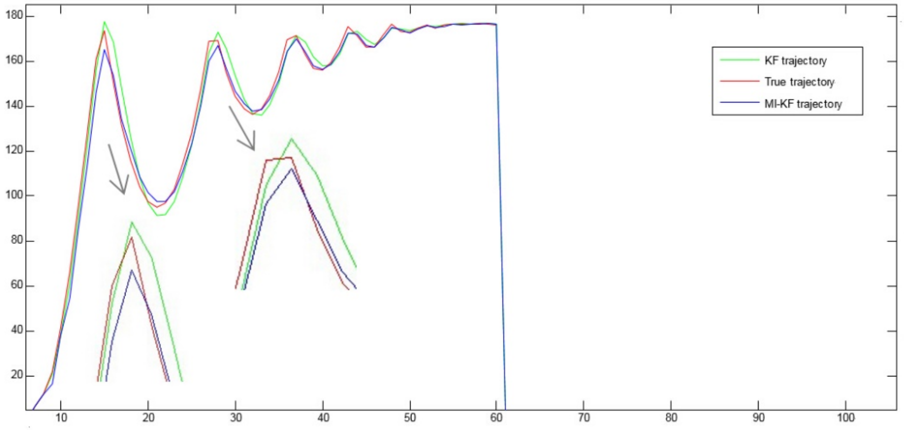


Fig. 7. The trajectory of the ball centroid

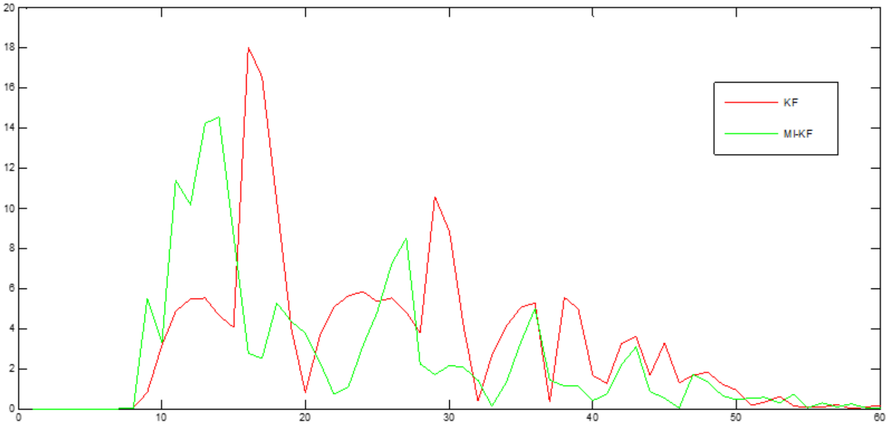


Fig. 8. Distance from the actual position and forecast position of the ball

Table 2 exhibits the RMSE between the standard Kalman prediction tracking algorithm and the improved multi-innovation Kalman prediction algorithm. We can visually see from the table, RMSE of the improved algorithm significantly lower than the standard algorithm, the prediction accuracy of the algorithm has been greatly improved.

Table 2. RMSE of the two methods of tracking

Filtering algorithm	KF	MI-KF
RMSE	4.9201	4.2624

5.3 Algorithm Performance and Complexity Analysis

Fig. 9 and Table 3 from two aspects of the step time and the total time to illustrate the difference between the standard KF algorithm and the improved MI-KF algorithm of abrupt moving target tracking. From Fig. 4 can be seen in every step of filtering, due to the improvement of the MI-KF algorithm requires the prior innovation iteration and participating in operation, so in single step will longer than the single innovation of the KF algorithm. As can be seen from table 3, the improved MI-KF algorithm than the standard KF algorithm for each step of the average use more than 0.01 milliseconds, more than the total time 5 milliseconds. However, MI-KF still used in the real-time requirements of the acceptable range, so a comprehensive comparison MI-KF algorithm is an overall advantage.

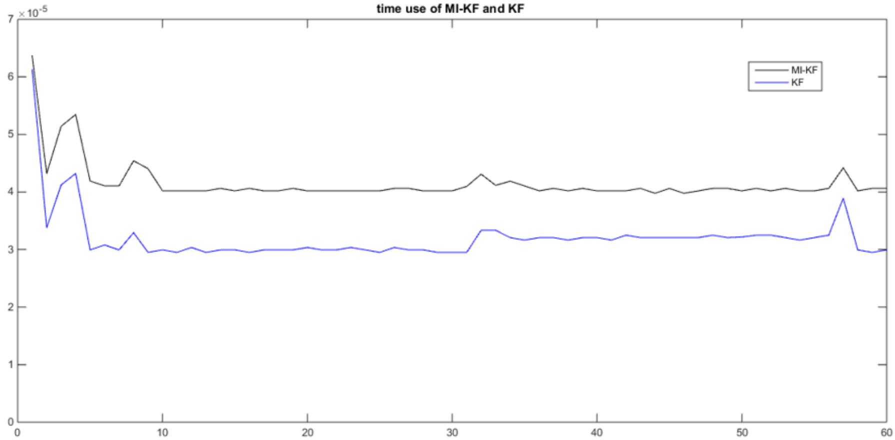


Fig. 9. Performance analysis of KF and MI-KF

Algorithm\Time	Average time(ms)	Total Time(ms)
KF	0.03241	1.94
MI-KF	0.04221	2.53

6 Conclusions

In this paper, we have presented an improved KF algorithm based on multi-innovation, which combines the multi-innovation theory, on the basis of the traditional KF algorithm. This algorithm gives full consideration to the useful information before the current time. These two algorithms are compared under different simulation conditions to demonstrate that the MI-KF can achieve better predictive effect than KF in predicting the target motion state, especially for the target whose state changes suddenly. Under such situation, our algorithm can avoid losing track of the target caused by the sudden change of state. Although the computation of MI-KF is slightly larger than the standard KF, but we have demonstrated by the above experiments that the increased computation has no effect on the real-time application of the MI-KF algorithm.

Acknowledgment. This work is partially supported by Shanxi Youth Foundation (No.2010021016-3). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Hou, Z.-Q., Han, C.-Z.: A Survey of Visual Tracking. Acta Automatica Sinica **32**(4), 603–617 (2006)
2. Xu, Y.-C., Wang, X.: Military test platform unmanned vehicle overall design. In: 2008 Annual Meeting of China Society of Agricultural Machinery, vol. 27, pp. 422–426 (2008)
3. Rad, R., Jamzad, M.: Real time classification and tracking of multiple vehicles in highways. Pattern Recognition Letters. **26**(10), 1597–1607 (2005)
4. Daum, F.E.: Nonlinear filters: Beyond the KF. IEEE Trans. Aerospace & Electronics Magazine. **20**(8), 57–69 (2005)
5. Evensen, G.: The ensemble KF for combined state and parameter estimation – Monte Carlo techniques for data assimilation in large systems. IEEE Control Systems Magazine **29**(3), 83–104 (2009)
6. Takeuchi, Y.: Optimization of liner observation for the stationary KF under a quadratic performance criterion. International Journal of Innovative Computing, Information and Control **7**(1), 85–99 (2011)
7. Sawada, Y., Kondo, J.: KF based LEQG control of parallel-Structured single-link flexible arm mounted on moving base. International Journal of Innovative Computing, Information and Control **6**(1), 29–42 (2010)
8. Karasalo, M., Hu, X.: An optimization approach to adaptive KF. In: 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, vol. 47, No. 3, pp. 1785–1793 (2011)
9. Sun, Y.-S., Li, Y.-M., Wan, L., Pang, Y.: An improved self-adaptive KF algorithm and its application in integrated navigation systems for AUV. Chinese High Technology Letters **23**(2), 174–180 (2013)
10. Hu, J.-H., Xu, J.-J.: Tracking of moving object based on genetic algorithm and KF. Journal of Computer Applications **27**(4), 916–918 (2007)
11. Qu, S.-R., Yang, H.-H.: Multi-target detection and tracking of video sequence based on Kalman_BP neural network. Infrared and Laser Engineering **42**(9), 2553–2560 (2013)
12. Vaidehi, V., Chitra, N., Chokkalingam, M.: Neural network aided KF for multitarget tracking applications. Computers and Electrical Engineering **27**, 217–228 (2001)
13. Hu, P.: Research on Video Object Tracking with KF. Chongqing University **28**(9), 11–15 (2010)
14. Ding, F., Xiao, D., Tao, D.: Multi-innovation stochastic gradient identification methods. Control Theory and Application **20**(6), 870–874 (2003)
15. Ding, F.: Several multi-innovation identification methods. Digital Signal Processing **20**(4), 1027–1039 (2010)

16. Ding, J., Xie, L., Ding, F.: Performance analysis of multi-innovation stochastic gradient identification for non-uniformly sampled systems. Control and Decision. **26**(9), 1338–1342 (2011)
17. Ding, F.: System identification – Part A: Introduction to the identification. Journal of Nanjing University of Information Science & Technology (Natural Science Edition) **3**(1), 1–22 (2011)
18. Liu, Y.J., Ding, F.: Hierarchical least squares identification method for periodically non-uniformly sampled systems. Control and Decision **26**(3), 453–456 (2011)
19. Ding, F.: System identification. Part F: Multi-innovation identification theory and methods. Journal of Nanjing University of Information Science & Technology. **4**(1), 1–28 (2012)

Visual Tracking via Structure Rearrangement and Multi-scale Block Appearance Model

Guang Han, Xingyue Wang(✉), and Jimuyang Zhang

Nanjing University of Posts and Telecommunications, Nanjing, China
wangxingyue2009@163.com

Abstract. In this paper, we propose a tracking method via structure rearrangement and multi-scale block based appearance model, plus a dynamic template mechanism within a two-stage search framework. Firstly, a wide range sampling is performed then confidence value of each candidate is calculated via a discriminative reverse sparse coefficient vote method, a part of candidates with large confidence value are selected to join in next stage. After a small range resampling in stage two, the candidates and target templates are divided into multi-scale patches, in addition, the background information is used to model the error term. Furthermore, a labeled template pool is maintained in the tracking process to dynamically generate an appropriate template set for next frame according to the occlusion map of current tracking result. Both qualitative and quantitative evaluations on challenging image sequences demonstrate that the proposed tracking algorithm performs favorably against several state-of-the-art methods.

Keywords: Visual tracking · Template structure rearrangement · Multi-scale patch · Dynamic template · Sparse representation

1 Introduction

Visual tracking [1] [2] is one of the classic computer vision problems, and has a wide application in numerous scenarios [3] [4] such as video surveillance, human-computer interaction, etc. Despite the success, to achieve a reliable object tracking still needs to overcome many difficulties, for example, the target partially occlusion, illumination variation, background clutter, appearance change, etc.

In general, a visual tracking algorithm [11][12][23]-[28] can be divided into three key components: a motion model, an appearance model and a tracking strategy. The motion model [5][14] is used to describe the state of the target motion and forecast the likely target position in next frame, the appearance model[10][19]-[22] is a description method of target information including intrinsic and extrinsic characteristics of the target object, and the tracking strategy[15]-[18] is used to build the main framework of the algorithm with the motion model and appearance model.

Since the first time sparse representation is introduced into object tracking by Mei and Ling [32], it has been used in various tracking algorithms [13][23][26][28][30]-[33]. Mei use a target template set and a trivial template set to linear represent each candidate

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 106–119, 2015.
DOI: 10.1007/978-3-662-48570-5_11

within particle filter framework. In spite of demonstrated success, this method consumes a large amount of computation and is sensitive to partial occlusion. Bai et al [34] apply a structured sparse representation model to visual tracking algorithm, and propose a block orthogonal matching pursuit (BOMP) algorithm based on orthogonal matching pursuit algorithm and mutual relations of the target on spatial structure. This algorithm fully combines the structural characteristics of the target to reduce the amount of calculation and has good robustness on the target occlusion. In addition, Jia et al. [28]proposed an align method to extract the target local sparse and spatial information, and also robust to partial occlusion, but it did not take the particle resampling process into consideration, once tracking result deviates from the actual location of the target, the tracker may completely lose the target. Zhong et al. [33] propose a sparse collaborative model that exploits both local and holistic templates, the algorithm is able to deal with partial occlusion. However, different sequences require a separate debug and different parameters, the performance is sensitive to parameters and needs high computational cost to solve L1 problem for so many local patches.

In view of above analysis, we propose the following method to deal with challenges during tracking: Firstly, to take full advantage of the holistic spatial structure information of target, we propose a template structure rearrangement method to rearrange the spatial structure of the template, this method can effectively deal with occlusion without any loss of target information. Secondly, we propose a multi-scale patch based method to resist partial occlusion with the horizontal and vertical multi-scale patches. Thirdly, we propose a dynamic target template set generated from the labeled template pool to deal with long-term occlusion or appearance change. Fourth, we use the background information to construct negative templates and model the error term, which can improve tracking robustness. Fifth, we use a two-stage search strategy to handle the situation that the target object moves fast and randomly. In addition, we use the APG [30] method to solve optimum problem to improve the processing efficiency.

2 Proposed Tracking Algorithm

2.1 Template Structure Rearrangement

We find that the traditional L1tracker is robust to Gaussian noise, however not to partial occlusion. Inspired by this phenomenon and analysis of existing methods, we propose a template structure rearrangement method to rearrange the spatial structure of template, this method has the ability to fragment the partial occlusion appeared anywhere and as close as possible to make partial occlusion obeys the Gaussian distribution. The template is divided into 9 patches shown in Figure 1(a) and numbered from 1 to 9, and then the positions of patch 1 and 2, 5 and 6, 7 and 9 are swapped respectively. Finally, all patches are re-assembled into a holistic template as shown in the right side in Figure 1(a), the Figure 1(b) is a demonstration with actual template image.

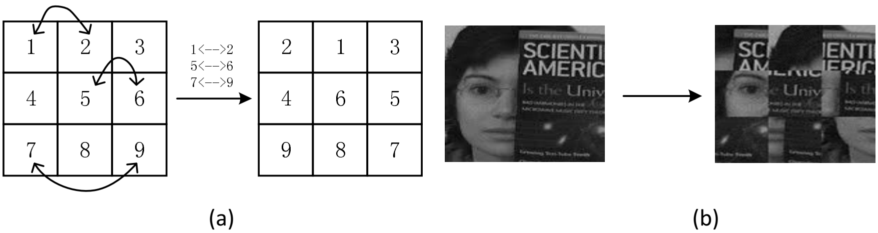


Fig. 1. A demonstration of structure rearrangement method

Usually the templates will be reshaped into a vector using the method demonstrates in Figure 2(a) when solving the L1 regularizations problem, if the target object is partial occluded (the colored patch represents occlusion), the occluded patch is still continuous distribution, which is seriously deteriorate the tracking performance. If the template structure is rearranged before reshaped into a vector shown in Figure 2(b), the occluded patch is divided into several smaller sub-patches so that the original continuous partial occlusion becomes close to trivial Gaussian occlusion.

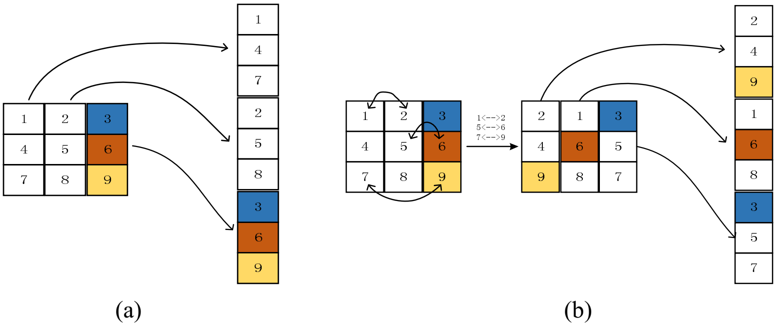


Fig. 2. A demonstration of handling partial occlusion

The proposed structure rearrangement method can effectively divide the partial occlusion into trivial occlusion and maintain spatial structure information of the target object itself without increasing the computational cost.

2.2 Discriminative Reverse Sparse Representation Based Vote Method

The traditional sparse representation based trackers use the target templates to linear represent the candidates and perform computationally expensive L1 regularizations problem at each frame for each candidate. To make sure the robust of L1 tracker, the candidates always amounts to hundreds or even thousands, which results in the tracker is unsuitable for an application with a real-time requirement. In this paper, motivated by a reverse thought of traditional sparse representation, we construct the dictionary with the candidate set Y to represent each target template as in Eq.1.

$$\mathrm{T} = \mathrm{Y}_{\mathrm{a}} + \mathrm{e} \tag{1}$$

Where T denotes the target template, Y donates the candidate set, a is the sparse coefficient vector, e is the error term, then the L1 regularizations problem can be re-write as Eq.2.

$$\arg\min_a \|T_i - Ya_i\|_2^2 + \lambda\|a_i\|_1,\ s.t.\ a_i \geq 0 \tag{2}$$

With sparsity constraint, only a few candidates which are similarly to the template would be involved in representing the template, and with non-negativity constraints the coefficient vector denotes the similarity between candidate and target template, therefore, a larger element of a means the corresponding candidate is more similar with the target template.

According to the characteristics of the sparse coefficient, each template contributes a vote for the candidate set, the vote value are a_i, combining all the voting values of the target template to arrive at a more accurate voting results a.

$$\mathrm{a} = \sum_i a_i \tag{3}$$

To further increase the accuracy of the voting, we introduce a weight W_{ij} for each voting value, W_{ij} represents the similarity between i-th template and j-th candidate:

$$W_{ij} \propto \cos < t_i, y_j > \tag{4}$$

Thus, the vote value of j-th candidate is W_j.

$$W_j \propto \sum_i \cos < t_i, y_j > \tag{5}$$

And the confidence value of each candidate is S_j^f.

$$\mathrm{S}_j^f = W_j \odot a_j \tag{6}$$

Where $\odot$ is the element-wise product, $W_j \propto \sum_i W_{ij}$,is the vote value of j-th candidate.

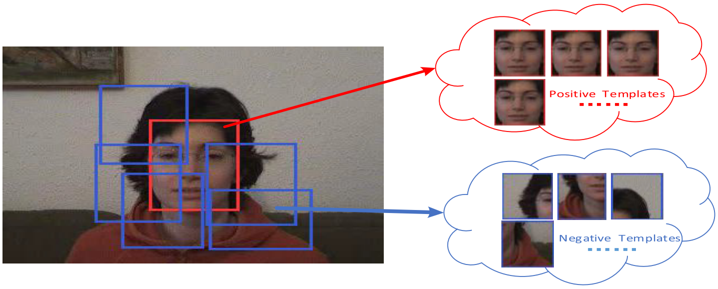


Fig. 3. Positive and negative templates

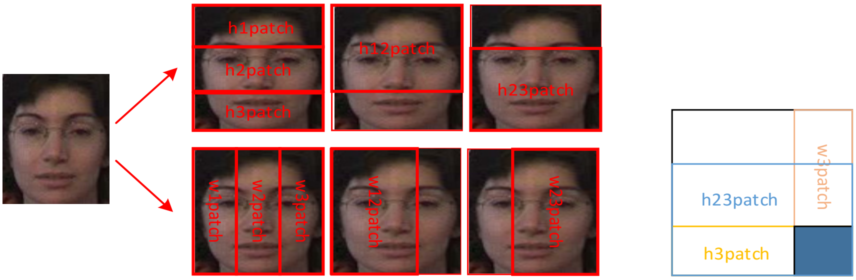
In order to further improve the tracking robustness, a negative template set (shown is Figure 3) is used to vote for each candidate in accordance with the above steps, then we can get the confidence value S_j^b generated from negative templates, thus, each candidate has a final confidence value S_j.

$$S_j = S_j^f - S_j^b \tag{7}$$

In this paper, the solution process is reduced to only dozen times by using the method of reverse sparse representation instead of the traditional sparse representation, and take full advantage of the inherent characteristics of the sparse coefficient to calculating a confidence value for each candidate.

2.3 Multi-scale Block Appearance Model

Multi-scale blocks are the patches divided from target template as shown in Figure 4(a). Each template has four scales of 10 sub-patches, and these sub-patches divided as following: trisect the template in the vertical direction and denoted by h1patch, h2patch, h3patch, respectively, then combine h1patch and h2patch as h12patch, combine h2patch and h3patch as h23patch; similarly, obtain w1patch, w2patch, w3patch, w12patch, w23patch in the horizontal direction. The benefits of this division method are: Firstly, the number of needed patches is less than the overlap patches based method, therefore the additional amount of calculation is small; Secondly, some patches contain other patches which maintains the structural information of the target object itself; Finally, the use of vertical and horizontal direction patch method having a similar function with coordinate axis which can be more quickly locate a smaller portion occlusion. As shown in Figure 4(b), when h3patch and w3patch are occluded, but h23patch or w23patch are not, then we can know that the colored patch in bottom right corner is occluded.



(a) (b)

Fig. 4. A demonstration of multi-scale target region division

2.4 Occlusion Map Creation

Firstly, we obtain the multi-scale patches of target template set and current tracking result, then calculate the cosine similarity between the patch $r_{i,\text{i}=\{1,2,\ldots,10\}}$ of tracking result and the corresponding position patch t_j of target templates, the confidence value con_i of each patch is proportional to the cosine similarity.

$$con_i \propto \sum_j \cos < r_i, t_j > \tag{8}$$

Thus, the occlusion map of current tracking result is written as $O = \{o_1, o_2, \ldots, o_{10}\}$,

$$o_i = \begin{cases} 1 & con_i < thr \\ 0 & con_i > thr \end{cases} \tag{9}$$

Where the thr is a predefined threshold. The occlusion map indicate whether 10 patches are occluded, depending on each element of O we can infer more accurate occlusion position and occlusion size.

2.5 Construction and Update of Dynamic Template Set

Due to partial occlusion occurs randomly, we consider not only the size but also the duration of occlusion. As shown in Figure 5, the dynamic template set of each frame is generated from the labeled template pool by the template factory, and the tracking result is used to update the labeled template pool and the template factory.

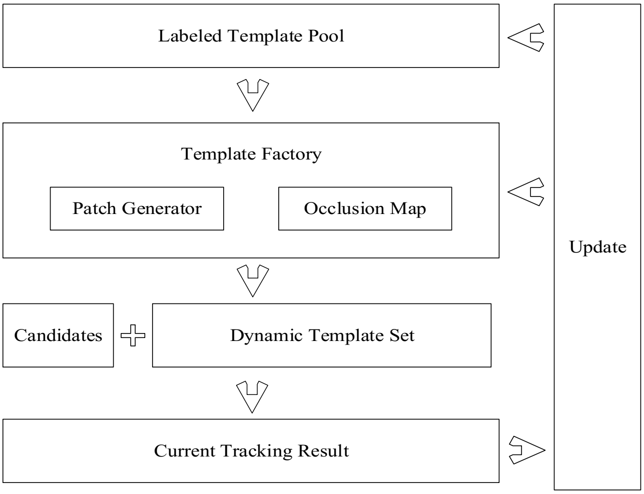


Fig. 5. Construction and update of dynamic template set

Labeled template pool, denoted by LTP. LTP contains 3 kinds of target template set with different labels: a template set without occlusion denoted by NT; a template set with small occlusion area (the occlusion area less than one-third of the area of the template), denoted by GT; a template set with large occlusion area (the occlusion area less than two-third of the area of the template), denoted by HT. Thus, LTP = {NT, GT, HT}.

Template factory consists of two parts: Patch generator and occlusion map. Patch generator is used to generate multi-scale template patches, and occlusion map used to indicate the occlusion position of the target.

Dynamic template set changes in a new frame according to the occlusion map of current tracking result, as shown in Eq.10.

$$T = \begin{cases} \{NT\} & sum(o) = 0 \\ \{NT, GT\} & 0 < sum(o) \leq 6 \\ \{NT, GT, HT\} & sum(o) > 6 \end{cases} \tag{10}$$

If current tracking result is not occluded, then it will be labeled as NT and only update the NT set, of course the template set is T={NT} in next frame. If current tracking result is partial occluded and satisfy $sum(o) > 6$, then it will be labeled as HT and only update the HT set, the template set is $T = \{NT, GT, HT\}$ in next frame, other cases handled according to the same principle. The proposed dynamic template set can effectively deal with a long term occlusion and reduce the negative impact of inappropriate update strategy for tracking.

In addition to the above holistic based dynamic template set, in this paper we also construct a multi-scale patch based dynamic template set, once the template T is obtained, T is divided into multi-scale patches, thus we can obtain a patch based dynamic template set $T_p = \{T_{h1}, T_{h2}, T_{h3}, T_{h12}, T_{h23}, T_{w1}, T_{w2}, T_{w3}, T_{w12}, T_{w23}\}$. The multi-scale patch based method can be more flexible to handle partial occlusion, pose change, etc.

2.6 Error Term Modeling

Motivated by that the occlusion usually comes from the background in tracking scenario, in this paper we use the background information to model the error term. As shown in Figure 6, we randomly select N patches with size 2×2 in our experiments near the target object in the red box, then assign to the trivial templates with the pixel values of these N patches, the other elements of trivial templates is set to zero. This background-patch templates take full advantage of the characteristics of occlusion itself, not only reduce the computational complexity by decreasing the number of dictionaries, but also improve the tracking accuracy.

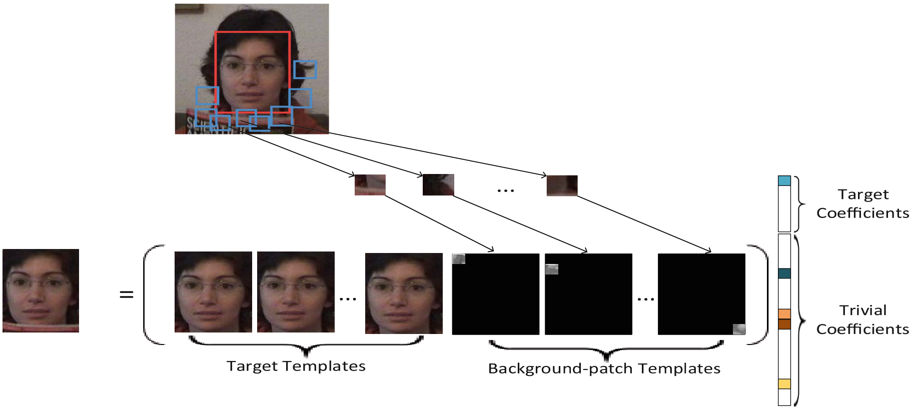


Fig. 6. Model error term with background-patch templates

2.7 Proposed Tracking Algorithm

After manually selecting target to be tracked, NT, the template set without occlusion, is generated by random tiny disturbance within the target region, and the negative templates are drawn further away from the marked location as shown in Figure 3, upon completion of the initialization process, the proposed tracking algorithm executed from the second frame. We uses a two-step search method to locate the tracking target, the specific process of this method is as follows: Firstly, we carry a large-scale search based on holistic template in the vicinity of the target location determined in last frame and sample a candidate set Y_1, then calculate the confidence values of each candidate using the discriminative reverse sparse representation method(see section 2.2) with the structure rearrangement template set Y_1 and dynamic template set T, the candidate with largest confidence value denoted by $\mathrm{r}_{\mathrm{s}1}$and the candidates whose confidence value are top k are denoted by $\mathrm{Y}_{\mathrm{s}1}$. In second stage, we draw a candidate set $\mathrm{Y}_{\mathrm{s}2}$with a small-scale particle sampling in the vicinity of $\mathrm{r}_{\mathrm{s}1}$, thus the candidate set of second stage is $\mathrm{Y}_2 = \{\mathrm{Y}_{\mathrm{s}1}, \mathrm{Y}_{\mathrm{s}2}\}$ (those candidates are raw images without structure rearrangement), then divide the candidate set Y_2 and dynamic template set T into multi-scale patches, now we can construct ten traditional sparse representation problems(different from the reverse sparse representation, in second stage we use the dynamic template set T to linear represent the candidates Y_2) as shown in Eq.11.

$$\begin{cases} Y_{h1} = T_{h1} * A_{h1} + e \\ Y_{h2} = T_{h2} * A_{h2} + e \\ \quad\quad \cdots \\ Y_{w23} = T_{w23} * A_{w23} + e \end{cases} \tag{11}$$

Where Y_{h1} and T_{h1} are composed of the h1patches of Y_2 and dynamic template set T, A_{h1} is the coefficient vector, e is the error term, other formulas are constructed in the same way. Then we can calculate the reconstruction error for each patch of each candidate,

$$\xi_i = \|y_i - T_i * A_i\|_2^2 \tag{12}$$

Where $\mathrm{i} = \{h1, h2, h3, h12, h23, w1, w2, w3, w12, w23\}$, then we can obtain the confidence value for each candidate in Eq.13,

$$conf_j \propto \sum_i e^{-\xi_i/w} \tag{13}$$

Where w is the weight parameter for reconstruction error, it is a predefined constant. We choose the candidate with largest confidence value as the tracking result in current frame.

2.8 Update Scheme

After locating the target position, the labeled template pool is updated firstly, we calculate the occlusion map of current tracking result, then determine its type by Eq.14.

$$\text{type}(\text{result}) = \begin{cases} NT & sum(o) = 0 \\ GT & 0 < sum(o) \leq 6 \\ HT & sum(o) > 6 \end{cases} \quad (14)$$

For example, if current tracking result belongs to NT set, we only update NT set by replacing the template with the smallest weight, the weight of template equals the corresponding confidence value. Similarly, GT set and HT set will be filled before updated. After updating the labeled template pool, we construct the dynamic template set T for next frame as shown in Eq.10, and the details of the update process will be found in section 2.5.

3 Experiments

The proposed algorithm is implemented in MATLAB and runs at 0.8 frames per second on an Intel 3.2 GHz i5-4570 Core PC with 4GB memory. In order to better evaluate the performance of our tracker, we conduct experiments on eleven challenging image sequences, these sequences focus cover the most challenging situations: heavy occlusion, illumination variation, fast motion, background clutter, in-plane and out-of-plane rotations and scale variation (See Figure 7). All the sequences are available online for public download. In this paper we compare with six state-of-the-art tracking methods including IVT[23]、FCT[29]、STC[27]、L1APG[30]、SCM[33] and ASLA[28]. For fair evaluation, all trackers run with the same initial positions of the targets. In our experiments, the target image patch is normalized to 32×32 pixels, the threshold thr in Eq.9 is fixed to be 9.5, the weight parameter w in Eq.12 is fixed to be 0.04, the labeled template pool is updated in every 3 frames, and all the parameters are fixed in all the experiments.

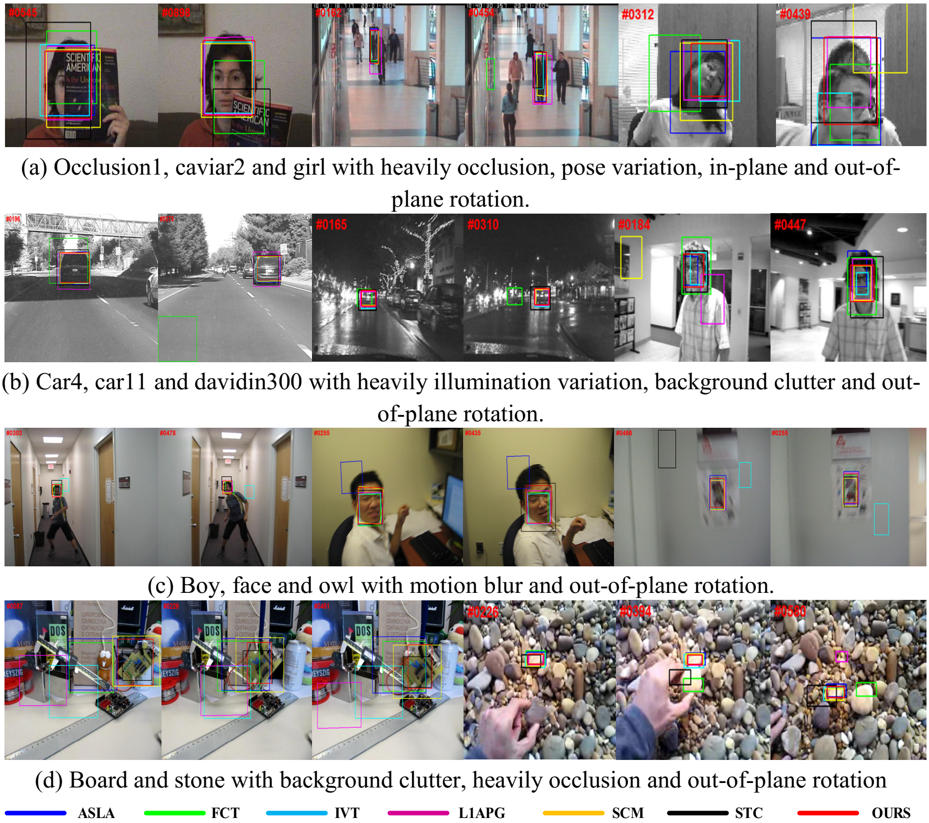
3.1 Qualitative Evaluation

Heavy Occlusion: We test three datasets occlusion1, caviar2 and girl, there are heavy occlusion and in-plane and out-of-plane rotations in these datasets shown in Figure 7(a). The proposed algorithm achieves better tracking performance judging from the experiments mainly because of the advantages of template structure rearrangement and the dynamic template mechanism, another importance reason is the two-stage search mechanism. The STC, IVT and FCT methods undergo some degree of drift, because their update mechanisms do not consider how to deal with partial occlusion. In contrast, L1APG、SCM and ASLA achieve better performance because of their update schemes deal with occluding patches.

Illumination Variation: We test three datasets car4, car11 and davidin300 with large illumination variation as shown in Figure 7(b), in addition, there are a certain scale variation and pose variation in these datasets. The FCT method does not track the target well when large illumination variation occur and has drift in all three sequences. In the davidin300 sequence, L1APG method tracks a wrong target which can be attributed to the fact that its trivial template mechanism mistakenly target pose and facial expression variation as occlusion. As a whole, the proposed tracker, SCM and ALSA perform well because they all use the local patches strategy based on sparse representation.

Fast Motion: We test three datasets boy, Owl and face with motion blur caused by target object fast motion as shown in Figure 7(c). The results show that, most tracking algorithms fail to follow the target right even lose target object. Some algorithm can continue to track the target object due to the fast moving target is in its place, but may be completely lose the target object if the fast moving target is not in the same place. Compared to other tracking algorithm, the proposed algorithm achieves the best tracking results, because the two-step search mechanism is applied.

Background Clutter: We test two datasets board and stone with complex background as shown in Figure 7(d), there are multiple objects in the background including some region which is similar to the target in terms of appearance. In addition, there are serious out-of-plane rotation and heavily occlusion. In the board sequence, the L1APG and IVT tracker almost both lose the target in all the frames, and the SCM, ALSA, FCT, STC tracker all drift to the background when the target object undergoes serious out-of-plane rotations. In contrast, our tracker performs well throughout this long sequence. In the stone sequence, the L1APG, FCT and STC all lose the target, but our tracker, SCM, ASLA, IVT perform well.



(a) Occlusion1, caviar2 and girl with heavily occlusion, pose variation, in-plane and out-of-plane rotation.

(b) Car4, car11 and davidin300 with heavily illumination variation, background clutter and out-of-plane rotation.

(c) Boy, face and owl with motion blur and out-of-plane rotation.

(d) Board and stone with background clutter, heavily occlusion and out-of-plane rotation

Fig. 7. Sample tracking results on eleven challenging sequences.

3.2 Quantitative Evaluation

We evaluate the above-mentioned algorithms using the center location error and overlap ratio, center location error is the center distance (in pixels) of the tracking results and the manually labeled location, the smaller center location error represents the better performance; and the overlap ratio is computed by intersection over union based on the tracking result R_T and the ground truth R_G, i.e., $\frac{R_T \cap R_G}{R_T \cup R_G}$, the larger overlap ratio represents the better performance.

Table 1. Comparison results in terms of average center errors (in pixels). The best three results are shown in red, Blue and Green fonts.

sequences	ASLA	FCT	IVT	L1APG	SCM	STC	Our
Board	84.1666	93.6998	164.9847	216.7564	31.7272	21.3653	11.7404
Boy	2.7294	8.4897	42.5409	6.0128	2.6866	10.7276	2.2062
car4	3.4738	178.9681	3.7852	13.2720	3.9399	14.7730	1.9223
car11	2.0578	26.9202	2.9854	2.4786	1.9946	3.5326	1.4043
caviar2	1.5334	61.1609	4.5668	12.3992	2.2586	6.5082	2.2100
davidin300	18.5372	10.1545	3.6339	24.0785	22.3952	8.5710	3.2389
Face	98.2796	30.6608	29.6935	28.9030	141.8946	28.3821	4.0860
Girl	16.3122	34.5508	30.4518	11.6667	29.6594	14.6307	10.2709
occlusion1	7.5252	38.3269	9.8410	7.7760	4.0931	34.2394	3.0815
Owl	27.7134	26.8186	99.2821	27.2591	29.8939	197.1934	1.7551
Stone	3.7389	27.6678	11.0253	7.6665	2.6097	19.8194	2.2837
average	24.1880	48.8562	36.6173	32.5699	24.8321	32.7039	4.0181

Table 2. Comparison results in terms of average overlap rates (in pixels). the best three results are shown in red, blue and green fonts.

sequences	ASLA	FCT	IVT	L1APG	SCM	STC	Our
Board	0.4007	0.3227	0.1462	0.0729	0.7048	0.3828	0.5601
Boy	0.7914	0.5764	0.3402	0.7215	0.7954	0.5389	0.8238
car4	0.9011	0.2607	0.9136	0.6871	0.8992	0.6873	0.9265
car11	0.8163	0.3762	0.7542	0.7930	0.7961	0.6341	0.8422
caviar2	0.7706	0.3109	0.6054	0.5991	0.6731	0.6681	0.7707
davidin300	0.4758	0.4733	0.6324	0.4488	0.5915	0.5605	0.7440
Face	0.2033	0.5048	0.5109	0.5374	0.3280	0.5235	0.9216
Girl	0.6443	0.5024	0.5888	0.7186	0.5763	0.5468	0.7016
occlusion1	0.8590	0.5172	0.8192	0.8461	0.9026	0.5012	0.9361
Owl	0.4959	0.4897	0.1976	0.4969	0.4693	0.0998	0.9417
Stone	0.5289	0.3338	0.5206	0.6309	0.6118	0.3446	0.6401
average	0.6261	0.4244	0.5481	0.5957	0.6680	0.4989	0.8008

Table 1 and Table 2 show the average center error and overlapping ratio where the red, blue and green fonts represent the top three tracking results. The sequences caviar2, girl, occlusion1 and stone have serious occlusion, even totally occluded. The average center errors of FCT, IVT and STC are larger in table 1, meanwhile, the average overlap rates in table 2 is smaller, which indicates these trackers lost their target, however L1APG, SCM and ASLA contain partial occlusion handling mechanism and the proposed tracker using a template structure rearrangement method and dynamic template mechanism to deal with occlusion. The sequences car4, car11 and davidin300 undergo a large illumination variation, the FCT method is not perform well, and the sparse based methods have more robustness to illumination. The average center errors is large and the average overlap rate is small for most trackers in sequences boy, Owl and face with motion blur, which indicates they almost lose target, however the proposed method introduce a two-step search mechanism to resist drifting. In the sequences board and stone with complex background the proposed method has the smallest average center errors and a larger average overlap rate, which indicates the proposed appearance model is robust to complex background. The last row in table 1 and table 2 represents the comprehensive performance of each algorithm in all sequences. Overall, the proposed algorithm is better than the other six kinds of popular algorithms.

4 Conclusion

In this paper, we propose a robust object tracking algorithm via structure rearrangement and multi-scale block appearance model. The structure rearrangement method in this algorithm rearranges the spatial structure of template without loss of target information and is ability to fragment the partial occlusion appeared anywhere, plus the multi-scale patches method the proposed tracker is robust to target appearance changes caused by the light, posture changes and heavily occlusion and so on. Furthermore, the use of labeled template pool and dynamic template set can not only effectively deal with the long-term occlusion and permanent deformation, but also reduce the negative impact of template update strategy for tracking performance. We also design a two-step search method to trim tracking results. Finally, both qualitative and quantitative evaluations on eleven challenging image sequences demonstrate that the proposed tracking algorithm performs favorably against the other six kinds of popular algorithms.

Acknowledgement. This work was supported by the China National Natural Science Funds (61302156) and the University Natural Science Research Project of Jiangsu province (13KJB510021).

References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Computing Surveys **38**(4) (2006)
2. Cannons, K.: A review of visual tracking, York University, Tech. Rep. (2008)
3. Trigueiros, P., Ribeiro, F., Reis, L.P.: Generic system for human-computer gesture interaction. In: Proceedings of the IEEE Conference on ICARSC (2014)

4. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer **29**(10), 983–1009 (2013)
5. Collins, R.T.: Mean-shift blob tracking through scale space. In: Proceedings of the IEEE Conference on CVPR (2009)
6. Collins, R.T., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(10), 1631–1643 (2005)
7. Yang, M., Yuan, J.: Spatial selection for attentional visual tracking. In: Proceedings of the IEEE Conference on CVPR (2007)
8. Kwon, J., Lee, K.M.: Visual tracking decomposition. In: Proceedings of the IEEE Conference on CVPR (2010)
9. Kwon, J., Lee, K.M.: Tracking by sampling trackers. In: Proceedings of the ICCV (2011)
10. Li, H., Shen, C.: Real-time visual tracking using compressive sensing. In: Proceedings of the IEEE Conference on CVPR (2011)
11. Sevilla-Lara, L., Learned-Miller, E.: Distribution fields for tracking. In: Proceedings of the IEEE Conference on CVPR (2012)
12. Oron, S., Bar-Hillel, A.: Locally orderless tracking. In: Proceedings of the IEEE Conference on CVPR (2012)
13. Zhang, T.: Robust visual tracking via multi-task sparse learning. In: Proceedings of the IEEE Conference on CVPR (2012)
14. Avidan, S.: Ensemble tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(2), 261 (2007)
15. Babenko, B.: Visual tracking with online multiple instance learning. In: Proceedings of the IEEE Conference on CVPR (2009)
16. Grabner, H., Bischof, H.: On-line boosting and vision. In: Proceedings of the IEEE Conference on CVPR (2006)
17. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: Proceedings of the ECCV (2008)
18. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N learning: Bootstrapping binary classifiers by structural constraints. In: Proceedings of the IEEE Conference on CVPR (2010)
19. Wang, S., Lu, H., Yang, F., Yang, M.-H.: Superpixel tracking. In: Proceedings of the on ICCV (2011)
20. Wen, L., Cai, Z.: Robust online learned spatio-temporal context model for visual tracking. IEEE Transactions on Image Processing **23**(2), 785–796 (2014)
21. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(5), 564–575 (2003)
22. Matthews, I.: The template update problem. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**, 810–815 (2004)
23. Ross, D., Lim, J., Lin, R., Yang, M.-H.: Incremental learning for robust visual tracking. International Journal of Computer Vision **77**(1), 125–141 (2008)
24. Bai, S., Liu, R., Su, Z.: Incremental robust local dictionary learning for visual tracking. In: Proceedings of the ICME (2014)
25. Zhang, J., Cai, W., Tian, Y., Yang, Y.: Visual tracking via sparse representation based linear subspace model. In: Proceedings of the IEEE Conference on Computer and Information Technology (2009)
26. Liu, B., Huang, J., Yang, L., Kulikowsk, C.: Robust tracking using local sparse appearance model and k-selection. In: Proceedings of the IEEE Conference on CVPR (2011)

27. Zhang, K., Zhang, L., Liu, Q., Zhang, D., Yang, M.-H.: Fast visual tracking via dense spatio-temporal context learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 127–141. Springer, Heidelberg (2014)
28. Jia, X., Lu, H., Yang, M.-H.: Visual tracking via adaptive structural local sparse appearance model. In: Proceedings of the IEEE Conference on CVPR (2012)
29. Zhang, K., Zhang, L., Yang, M.-H.: Real-time compressive tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 864–877. Springer, Heidelberg (2012)
30. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust L1 tracker using accelerated proximal gradient approach. In: Proceedings of the IEEE Conference on CVPR (2012)
31. Mei, X., Ling, H.: Robust visual tracking and vehicle classification via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(11), 2259–2272 (2011)
32. Mei, X., Ling, H.: Robust visual tracking using L1 minimization. In: Proceedings of the ICCV (2009)
33. Zhong, W., Lu, H., Yang, M.-H.: Robust object tracking via sparse collaborative appearance model. IEEE Transaction on Image Processing **23**(5), 2356–2368 (2014)
34. Bai, T., Li, Y., Tang, Y.: Structured sparse representation appearance model for robust visual tracking. In: Proceedings of the IEEE Conference on Robotics and Automation (2011)

Convolutional Neural Networks for Clothes Categories

Zhi Li, Yubao Sun, Feng Wang, and Qingshan Liu(✉)

B-DAT Lab, School of Information & Control, Nanjing University of Information Science and Technology, Nanjing 210044, China
qsliu@nuist.edu.cn
http://bdat.nuist.edu.cn/

Abstract. Clothes classification is a promising research topic. Due to the manually-designed features' limitation, the existing algorithms have a problem of low accuracy in attributes classification. In this paper, we propose a new method to utilize convolutional deep learning for clothes classification. We firstly set up a new database by downloading the images of each category from Internet via related software and manual work, which divides clothes into 16 categories according to the common clothing style in the market. Then, the paper designs convolutional neural networks(CNNs) architecture and adaptively learns the feature representation of clothes from our constructed dataset. The experiment adopts Bag of Words (BOW), Histogram of Oriented Gradient (HOG)+ Support Vector Machine(SVM)and HSV (Hue, Saturation, Value)+SVM to test the new database and compares these methods with our CNNs model. The results demonstrate the superiority of our CNNs to the other algorithms.

Keywords: Clothes categories · Deep learning · Convolutional neural networks

1 Introduction

The 2013 annual Chinese apparel e-commerce operation report showed that variety of goods in the online market were exponentially expanding to meet customer's increasing demands, especially clothes and footwear products. In 2013, clothes and footwear products took up the highest market share in the online market, purchasing rate reached up to 76.2%. Thus, clothes and footwear products have become the most promising goods in the online market. Nowadays, most commercial image retrieval systems mainly rely on the key words search, such as TaoBao, JingDong and SuNing e-commerce. However, these systems have two weaknesses: first, every original image needs to be marked with key word. With the widely spread of smartphones, numerous images are updated everyday. It costs a large amount of human resource and materials to mark images. Second, because of cognition subjectivity, people may have different understandings of the same image, which will result in subjectivity and inaccuracy when the images are marked by different key words.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 120–129, 2015.
DOI: 10.1007/978-3-662-48570-5_12

Many researchers have devoted to designing automatic classification of clothing. Pan et al. [1] proposed a BP neural network to recognize woven fabric. Ben et al. [2] recognized woven fabric based on the texture features and SVM classifier. Yamaguchi et al. [3] described clothes by labeling superpixels, which were obtained from image segmentation making use of a Conditional Random Field model. Liu et al. [4] had a proposal for describing clothes based on pose estimation and using the features like color, SIFT and HOG and classified clothes into 23 categories. Bourdev et. al. [5] proposed a system describe the appearance of people by using 9 binary attributes such as male/female with T-shirt and long hair. For clothes segmentation, Manfred et al. [6] presented an approach for segmenting garments in fashion stores databases. Hu et al. [7] proposed a new clothing segmentation method using foreground and background estimation based on the constrained delaunay triangulation (CDT), without any pre-defined clothing model. Weber et al. [8] introduced a novel approach to get the mask of the clothing starting from a set of trained pose detectors, in order to deal with occlusions and different poses inherent to humans. Also, the clothes can be classified by the attributes, such as color, pattern, neck type, sleeve and others. Chen et. al [9] proposed a system that is capable of generating a list of nameable attributes for clothes in unconstrained images. Lorenzo-Navarro et al. [10] presented an experimental study about the capability of the LBP, HOG descriptors and color for clothing attribute classification.

However, the previous clothing categorization algorithms have been trapped in two limitations. First, the traditional features can't achieve satisfactory results, especially for similar classes. Second, there has not a public clothing database yet to evaluate the algorithms in fair. Therefore, this paper contributes to proposing the clothes classification algorithm based on deep learning and setting up a new large clothing database. We design convolutional neural networks(CNNs) architecture which adaptively learns the feature of clothes representation. In additional, we set up a new clothing database by downloading the images of each category from Internet via related software and manual work, which divides clothes into 16 categories according to the common clothing style in the market. Comparing with some traditional manually-designed features methods, our algorithm obtains a better performance.

2 Construction of Clothing Database

So far, there is no a public clothing database, and also previous works often evaluated the method in a small database. In this paper, we build a new large database. We divide the clothes into 16 categories (8 categories of menswear and 8 categories of womenswear) according to the common clothing styles in the market, and we download the images from the Internet with human labeling. As the show of table 1, the new database contains 33965 samples, and we randomly select 27565 images as the training samples(14142 menswear samples, 13425 womenswear samples) and the rest 6400 images as the validation samples(3200 menswear samples, 3200 womenswear samples). The clothes are categorized into 16 clothing categories:

Jacket, Mens shirts, Men's windbreaker , Men's suits, Ski-wear, Men's knitwear, Men's down jacket, Men's T-shirts, Cheongsam, Women's shirt, Women's Windbreaker, Women's suits, Dress, Women's fleece, Women's down jacket, Women's T-shirt. The number of each categories samples is shown in the table 2 and table 3. The figure 1 shows us samples from our database.

Table 1. Train and validation total samples

samples	Men	Women	Total
Train	14142	13423	27565
Validation	3200	3200	6400
Total	17342	16623	33965

Table 2. Men's train and validation samples

	Train samples	Validation samples	Total samples
Jacket	1587	400	1987
Men's shirts	1582	400	1982
Men's windbreaker	2204	400	2604
Men's suits	1854	400	2254
Ski-wear	1652	400	2052
Men's knitwear	1873	400	2273
Men's down jacket	1636	400	2036
Men's T-shirts	1754	400	2154
Total	14142	3200	17342

Table 3. Men's train and validation samples

	Train samples	Validation samples	Total samples
Cheongsam	1610	400	2010
Women's shirt	1662	400	2062
Women's windbreaker	1603	400	2003
Women's suits	1662	400	2062
Dress	2017	400	2417
Women's fleece	1637	400	2037
Women's down jacket	1688	400	2288
Women's T-shirt	1544	400	1944
Total	13423	3200	16623

3 CNN Based Feature Learning

Deep learning model [11] is a class of machines that can learn a hierarchy of features by building high-level features from low-level ones. Such learning machines can be trained using either supervised or unsupervised approaches, and widely used in the field of computer vision such as object detection [12],

Fig. 1. Samples from the database

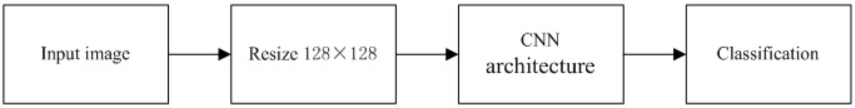


Fig. 2. Image processing for CNN

image classification [13] and image segmentation [14]. The convolutional neural network(CNN) [15] is a popular deep model in which trainable filters and local neighborhood pooling operations are applied alternatingly on the raw input images. CNN has been incorporated into a number of visual recognition systems in a wide variety of domains. CNN is previously proposed to contain many hidden layer of multilayer perceptron. By combining low-level features and discovering distributed characteristic presentation of data, deep learning forms more high-level characteristics stand by attribute categorisation and assembling. CNN attracted much attention in recent years, after obtaining much success in digit recognition [16], OCR [17] and object recognition tasks [18]. Due to the complex pattern of clothes, the common manually-designed features have the limitation of low accuracy in attributes classification. CNN can adaptively learn the high-level semantic features by the multiple layer architecture, which has the capacity to improve the performance of clothes classification. Thus, in this paper, we propose a new method to utilize convolutional deep learning for clothes classification.

From the figure 2, it can briefly show how to process image using CNN. We resize the image in the size of 128×128 for different image sizes from database. Then, the images are fed into CNN to learn network parameters. In order to improve the clothing recognition accuracy, the core is to design the effective network architecture which can learn appropriate features to represent the complex clothing appearance.

In Fig. 3, we design the architecture for our CNNs model. The architecture consists of 4 convolutions layers. We consider the image of size 128×128 as inputs to the CNN model. Then, we apply convolutions with a kernel of size 7×7, stride of 1, pad of 2 and C1 layer consists of 16 feature maps. We set pad as 2 in each convolutions in our architecture. In the subsequent subsampling layer S2, we apply 2×2 subsampling on each of the maps in the C1 layer. The next convolution layer C3 is obtained by applying convolution with a kernel

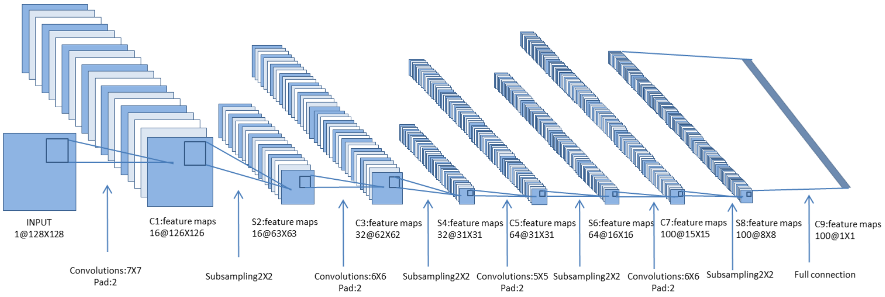


Fig. 3. The convolutional neural network architecture

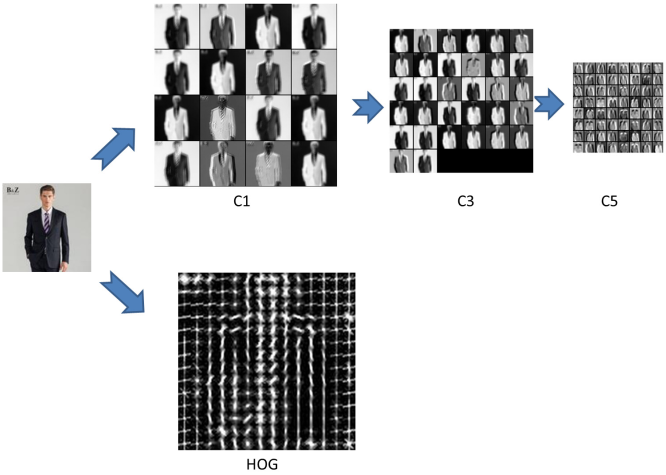


Fig. 4. Visual features learned by CNNs and HOG

of size 6×6 on each of feature maps separately. In the subsequent subsampling layer S4, we also apply 2×2 subsampling on each of the maps in the C3 layer. We set the convolution with a kernel of size 5×5 in the third layer and 6×6 in the fourth layer. The full connection layer consists of 100 feature maps of size 1×1. The size of image will be made 126×126, 62×62, 31×31, 15×15 after each convolution. Through the layers of convolution, the deep model can obtain the better features from shallow to deep.

Fig. 4 demonstrates the learned visual features by our designed CNNs. The output of C1, C3, C5 layer are displayed in the first column. The extracted HOG feature is also listed to compared with our learned features. It can be seen that HOG only represent the edge characteristic of clothes, lacking of the global pattern description and HOG features are sensitive to the noise result from HOG descriptors gradient operation. Different from the HOG features, our CNNs has the ability to abstract the features layer by layer. The features output form C5 can extract the global pattern of various clothes, not like the low-level edge information. Thus, the features leaned by our CNNs model can effectively represent the high-level semantic characteristic of clothes, which is more useful for clothes classification.

4 Experiments

4.1 Model

In order to evaluate the performance of our CNNs model, we adopt the classification accuracy as the measure criteria. Our model will also be compared with three baseline method, including BOW, HOG+SVM and HSV+SVM

BOW model is a common document representation method in image retrieval field. It consists of three steps. First, we extract visual vectors from different images by using the SIFT descriptor [19]. Second, we gather all feature points vectors together, and merge vectors with similar meaning through K-means algorithm [20]. Third, we compute the frequency that these words show up in images. Thus, these images are transformed into K-dimensional vectors. We put the feature vectors and labels into the SVM to train the classifier.

HOG initially proposes a descriptor which can implement human object detection. This method abstracts shape characteristic and movement information. In our work, we make use of a cell size of 8×8 pixels and the block is 32×32 cells.

HSV is a model that consists of three parameter: hue(H), saturation (S), value(V). The hue is measured in angle, and the range of hue is 0°∼360°. Hue counts from red counterclockwise. In this way, red represents 0°, green represents 120° and blue represents 240°. Saturation (S) varies from 0.0 to 1.0. The bigger the value is, the more saturated the color is. The range of value (V) is 0(black)∼255(white). In additional, HSV is a six pyramid model. We quantify hue into 64 intervals, and quantify saturation into 12 intervals, while value is not quantified. So we will establish a 768-order histogram.

4.2 Evaluation

Fig. 5 gives the classification results of four methods. We can see our CNN performs better than other methods and achieve the accuracy of 61.22%. HOG+SVM achieves the accuracy of 60.36% ranking in the second position. The third position is BOW with the accuracy of 56.27% and HSV+SVM performs worst with the accuracy of 20.58%.

Fig. 6 plots the accuracy curve of various method for each category. We can see the curve of our CNN compared with other curve which is overall at the top of the figure. In addition, we find an interesting phenomenon, if the accuracy of certain category is higher in CNN compared with other class, the same is happened in other three methods. On the other hand, if the accuracy of category compared with other class is lower, the same is true in other algorithms.

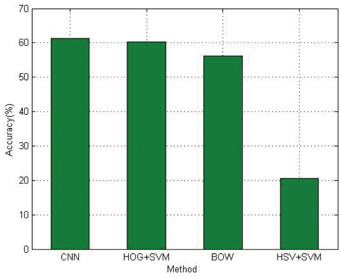


Fig. 5. The classification results of different methods

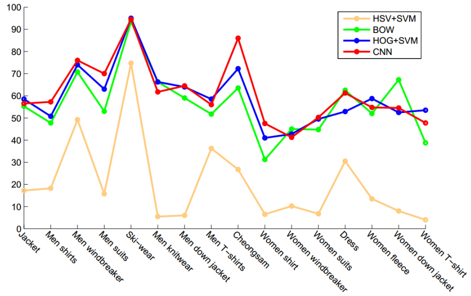


Fig. 6. Each category classification results for CNN, HOG+SVM, BOW and HSV+SVM in the database

Table 4. Detailed classification accuracy(%) for men's clothing

	CNN	HOG+SVM	BOW	HSV+SVM
Jacket	56.50	58.50	55.25	17.25
Men's shirts	57.25	50.75	47.75	18.25
Men's windbreaker	76.00	74.00	70.75	49.25
Men's suits	70.00	63.00	53.00	15.75
Ski-wear	94.50	95.00	93.50	74.75
Men's knitwear	61.75	66.25	66.25	5.50
Men's down jacket	64.50	64.00	59.00	6.00
Men's T-shirts	56.00	58.50	51.75	36.25

Detailed results of each category is shown in table 4 and table 5. We find that the accuracy of each category differs from each other. For examples, ski-wear and cheongsam accuracy are higher, and men's jackets, women's suits and women's shirts are relatively lower. We consider the main reason is that the ski-wears features of edge and color are relatively obvious and high degree of differentiation. However, the edge feature of jacket is confusing with windbreaker, suit and other kind of categories. In addition, their color features are not obvious which leads to the relatively lower accuracy.

Table 5. Detailed classification accuracy(%) for women's clothing

	CNN	HOG+SVM	BOW	HSV+SVM
Cheongsam	86.00	72.25	63.50	26.75
Women's shirt	47.50	41.00	31.25	6.50
Women's windbreaker	41.25	42.75	45.00	10.25
Women's suits	50.25	49.50	44.75	6.75
Dress	61.25	52.90	65.50	30.50
Women's fleece	54.75	58.75	52.00	13.50
Women's down jacket	54.50	52.50	67.25	8.00
Women's T-shirt	47.75	53.50	38.75	4.00

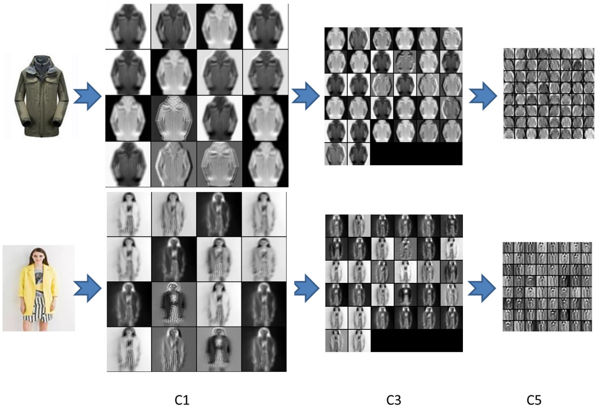


Fig. 7. Visual features of C1,C3,C5 for ski-swear and women's windbreaker

BOW is based on the regional block to extract feature, which can obtain more characteristics. However, compared with the HOG+SVM and CNN algorithm to extract the edge character, the accuracy of clothes recognition using BOW is slightly lower. But in some specific aspects such as Women's down jacket, dress and so on, it have certain advantages. The training of CNN model needs constantly iterative optimization. It can refer this iteration classification results to adjust the next iteration parameters. In addition, the convolution can capture good edge information of clothes and learn semantic feature. Therefore, the clothes with strong edge feature such as ski-wear, cheongsam and their accuracies are higher. However, for some similar clothes style, it's easily confused with each other on edge feature. Therefore, their accuracies are lower than other class. On the whole, our CNNs obtains a better performance.

4.3 Visual Analysis

Because of many clothes categories in the database, we wish to know what is the difference between these clothes categories by our CNNs. The figure 7 shows the visual features of ski-swear and women's windbreaker. We can see the original image and each features image after the C1, C3, C5. The size of original sample is 128×128. From the picture, we consider ski-wear is better than the women's windbreaker in features of edge. The profile of ski-wear's still clear even after C5 and these features can be learned easily by computer, which show better results. In contrast, the women's windbreaker doesn't show nice performance due to less strong edge features.

5 Conclusion

In this work, our convolutional neural networks obtains good results for clothes categories recognition. The experiments carry out with database which is set up by us. Our method learns the global information of image and semantic feature. The paper contributes to setting up a new clothing categories database and proposing the clothes classification algorithm based on CNN. We use the convolutional neural networks in deep learning, which can overcome the low accuracy in attributes classification. Comparing CNN with other traditional manually-designed features abstracted methods, our algorithm obtains a better performance. In future extensions of this work, we will optimize our deep networks architecture to improve the accuracy of database. In addition, database should be expanded with the increasing numbers of images furthermore, and we will publish our database in the right time.

Acknowledgments. This work was supported in part by the Natural Science Foundation of China under Grant 61272223, Grant 61300162, and in part by the Natural Science Foundation of Jiangsu Province, China, under Grant BK2012045 and Grant BK20131003.

References

1. Pan, R., Gao, W., Liu, J., Wang, H.: Automatic recognition of woven fabric pattern based on image processing and bp neural network. The Journal of the Textile Institute **102**(1), 19–30 (2011)
2. Salem, Y.B., Nasri, S.: Automatic recognition of woven fabrics based on texture and using svm. Signal, image and video processing **4**(4), 429–434 (2010)
3. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3570–3577. IEEE (2012)
4. Liu, S., Feng, J., Domokos, C., Xu, H., Huang, J., Hu, Z., Yan, S.: Fashion parsing with weak color-category labels. IEEE Transactions on Multimedia **16**(1), 253–265 (2014)
5. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1543–1550. IEEE (2011)
6. Manfredi, M., Grana, C., Calderara, S., Cucchiara, R.: A complete system for garment segmentation and color classification. Machine Vision and Applications **25**(4), 955–969 (2014)

7. Hu, Z., Yan, H., Lin, X.: Clothing segmentation using foreground and background estimation based on the constrained delaunay triangulation. Pattern Recognition **41**(5), 1581–1592 (2008)
8. Weber, M., Bauml, M., Stiefelhagen, R.: Part-based clothing segmentation for person retrieval. In: Advanced Video and Signal-Based Surveillance (AVSS), pp. 361–366. IEEE (2011)
9. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 609–623. Springer, Heidelberg (2012)
10. Lorenzo-Navarro, J., Castrillón, M., Ramón, E., Freire, D.: Evaluation of LBP and HOG descriptors for clothing attribute description. In: Distante, C., Battiato, S., Cavallaro, A. (eds.) VAAM 2014. LNCS, vol. 8811, pp. 53–65. Springer, Heidelberg (2014)
11. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. IEEE (2014)
13. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3642–3649. IEEE (2012)
14. Turaga, S.C., Murray, J.F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., Seung, H.S.: Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Computation **22**(2), 511–538 (2010)
15. 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)
16. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 41–551 (1989)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
18. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multistage architecture for object recognition? In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2146–2153. IEEE (2009)
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
20. Sangalli, L.M., Secchi, P., Vantini, S., Vitelli, V.: K-mean alignment for curve clustering. Computational Statistics & Data Analysis **54**(5), 1219–1233 (2010)

Multi-object Segmentation for Abdominal CT Image Based on Visual Patch Classification

Pan Li, Jun Feng(✉), QiRong Bu, Feihong Liu, and HongYu Wang

School of Information & Technology, Northwest University Xi'an, Xi'an, China
fengjun@nwu.edu.cn

Abstract. We introduce a robust multi-object segmentation algorithm based on visual patch classification for abdominal CT segmentation. Firstly, the proximity of the pixels is expressed by both intensity and spatial distance. And then clustering framework is employed to form various visual patches. In this way, the noise and embedded small tissues such as blood vessels and tracheas which often make other segmentation algorithms failed are filtered out during the cluster iteration. Afterwards, the visual patches are further grouped by the way of classification in the criteria of spatial relationship of visual pitches. Specially, the algorithm can be viewed as effectively tradeoff of bottom-up methods and top-down methods. The approach has been applied to the multi-object segmentation of abdominal CT images, such as the liver, kidney, spleen and gallbladder. We have test the method in American published TCIA database, whose efficiency and robustness is evaluated through quantification results on both sectional level and volumetric level, which exhibit the optimistic application and prospect in the field of medical image processing.

Keywords: Abdominal CT image · Segmentation · Visual pitch · Classification · Spatial relationship

1 Introduction

Medical image segmentation is the first phase of medical imaging data analysis and visualization, which are also the precondition and crucial to computer-aided diagnosis, image-guided surgery, virtual endoscopy and many medical image applications [1]. Compared with osseous organs, human abdominal soft-tissues are more complex and deformable, including the liver, the kidney, the gallbladder, the spleen as well as vascular such as the veins and arteries. However, due to the limitation of imaging device and the peristalsis of tissues, they often exhibit intensity inhomogeneity and overlapping in abdominal CT series. Besides, the blurred organs and the ambiguous of the edge of lesions also bring some considerable difficulties for segmentation. And for abdominal multiple organs segmentation, there will be much more influential factors, including high similarity of adjacent organs, partial volume effects and the relatively high variations of organ position and shape. Therefore, multi-object segmentation in abdominal CT image is still a challenge task [2].

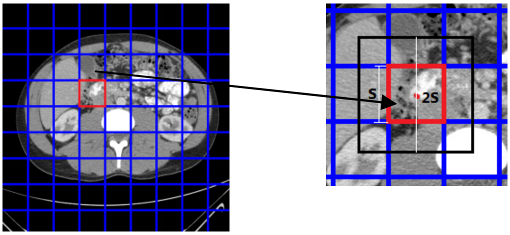
H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 130–138, 2015.
DOI: 10.1007/978-3-662-48570-5_13

Multi-object segmentation for abdominal CT image has become a research tendency, which has already reached some achievements. Daniel Freedman's algorithm, which is highly depend on learned shape and appearance models, compares the probability distributions instead of computing a pixelwise correspondence between the model and the image [3]. Robin Wolz et al. presented a multi-organ abdominal segmentation method based on a hierarchical atlas registration and weighting scheme that generates target specific priors from an atlas database, which finally obtain the segment result by applying an automatically learned intensity model [4]. Toshiyuki Okada et al proposed a method for finding and representing the interrelations based on canonical correlation analysis, which is developed for constructing and utilizing the statistical atlas [5]. Although the above two approaches are able to capture the organ location and appearance, it both relies on a subject-specific atlas model which impacted by inter-subject variability. Yinxiao Liu has developed an automatic threshold and gradient strength selection algorithm for unknown number of abdominal object regions by combining class uncertainty and spatial image gradient features, which is only in the view of mathematic and does not incorporate the spatial knowledge into segmentation[6]. To address the problem of spatial relationship among multiple organs of abdomen, Xiaofeng Liu et al extended the MAP framework by modeling the inter-organ spatial relations using a minimum volume overlap constraint, which focused on the posteriori probability and volume overlap without of characteristics of the organ themselves[7]. Besides, most of the traditional classification methods are put forward to single object with the ignorance of spatial information [8,9].

In this paper, we introduce a classification method for abdominal multi-object segmentation based on visual patch, which explicitly incorporate the spatial relationship of abdominal organs into classification. We introduce the definition of visual pitch, which is proved effective in color image pre-segmentation called superpixel [10], into the medical image segmentation. This pre-segmentation technology divides the image into several visual pitches by clustering the pixels with intensity similarity and spatial proximity, which are regarded as the unit of classification. Besides, through a large number of experiments, we discover an implicit spatial regularity among the visual patches of abdominal organs, which is implemented by establishing undirected adjacency graph of visual patches and finally integrated into classifier. In order to reducing the impact of similar intensity of different organs, we also consider the texture of visual patch into classification. Thus, the algorithm we proposed is the combination of low-level visual features and spatial physical characteristics, which also can be viewed as effectively tradeoff of bottom-up methods-clustering and top-down methods-classification. Experiments demonstrate that our method achieves the comparable performance with the state-of-art algorithms.

2 Visual Patches Generation Based on Superpixel

Here, we propose an unsupervised method for medical image, which generates the visual patch by clustering pixels based on both intensity similarity and spatial proximity. According to the complexity of image, there should be a K as the initial number of cluster center, and approximately equally divided the image into rectangle pitches shown in Figure 1(a). If the image is $N*M$, each initial visual patch is assigned about $N*M/K$ pixels. S is defined as the side-length of visual pitch, which is calculate by $S=\sqrt{N*M/K}$. The cluster region ranges $2S*2S$ around the center, which is shown in Figure 1(b).



(a) Initial visual patches (b) Cluster region

Fig. 1. Initial visual pitches

Intensity distance in grayscale space is perceptually effective for small distances, but it is no longer working when the space perception of the pixel exceeds the limit of intensity distance. Thus, we use the following measure instead of Euclidean distance:

$$d_{xy}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \tag{1}$$

$$d_g=\left|g_i-g_j\right| \tag{2}$$

$$D_s=\frac{\mu}{S}d_{xy}+d_{Gray} \tag{3}$$

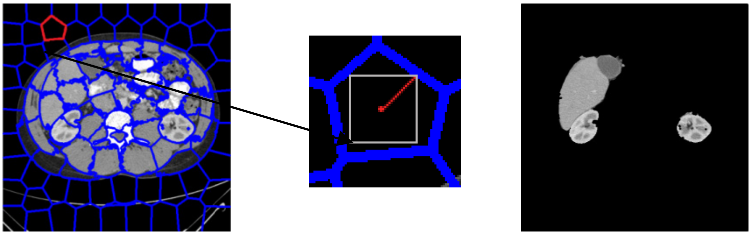
xy is the pixel spatial position, and g is the intensity value of pixel xy in grayscale image. Spatial proximity d_{xy} is calculated by Euclidean distance in two-dimension of image plane, and intensity distance d_g is regarded as feather similarity. D_s is the distance of the two pixels. μ is the parameter of the pixel compactness. The greater it is the spatial proximity is more important, and the clusters are more closely. Therefore, μ effectively balances the spatial proximity in grayscale space. It is known to us all that the standard deviation is an important method to measure the dispersion of numerical data, which also weighs the volatility of samples. Thus, we take the standard deviation of gray feature as the parameter μ which controls the compactness of pixels.

After above, select the minimum gradient pixel of each visual patch as the initial cluster center, which will avoids the impact of the boundary and noisy pixels. And then, use D_s incorporated the intensity similarity and spatial proximity as the clustering condition to cluster the pixels. Figure 2(a) shows the visual patches of the cluster result.

3 Multi-object Segmentation Based on Visual Patches Classification with Spatial Relationship

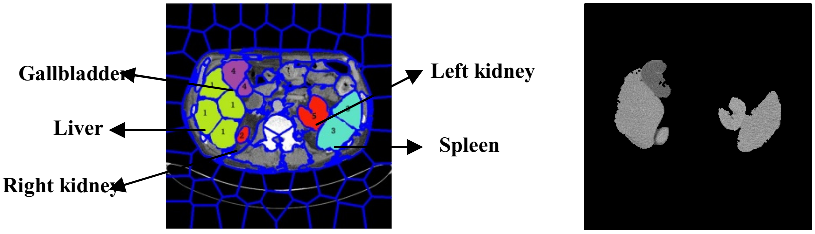
Classification is the essential methods of statistical analysis in the field of pattern recognition, which utilizes a known training sample to achieve segmentation purposes in the feature space of image. Without considering the spatial information, traditional classification segmentation algorithms are not sensitive to inhomogeneous intensity images, which will lead to large error in segmentation. There, we propose a method that exploits the supervised learning classification algorithm based on visual patches to segment the abdominal CT image to achieve the multi-object regions, in which we take the spatial relationship into consideration implemented by establishing an undirected adjacency graph of visual patches. In this way, even though the visual patches, which belong to the different organs, have the similar feathers, we will classify the patches exactly by the mutual positional relationship of organs.

Above all, the prior knowledge of the supervised classification algorithm in our method is the texture and the intensity of visual patches. Thus, it is crucial to extract the feather of the irregular patches. We put forward a measure that looking for the point in the edge of visual patch which is nearest to the cluster center of the patch, and take the distance as the half of the diagonal of rectangle, whose focus is the cluster center, and finally obtain the square shown in Figure 2(b). The gray level co-occurrence matrix method is exploited to extract the visual patches textural feathers [11]. Meanwhile, the intensity of visual patch is the average of all pixels' in patch.



(a) Visual patches (b)Texture region (c) Classification segmentation

Fig. 2. The visualization result of the steps



(a) Undirected adjacency graph of visual pitches (b) Classification segmentation

Fig. 3. The visualization result of the undirected adjacency graph

Through a large number of experiments, we discover an implicit spatial regularity among the visual patches of abdominal organs. First, we sign the visual pitches of the liver, right kidney, spleen, gallbladder, left kidney and background as a label (1,2,3,4,5,0). Then, we find that the visual pitches marked the same label are connected with each other. In addition to, the right kidney is close to the liver, and so as the gallbladder. And the spleen is near to the left kidney, both of which far away from the liver. Equation (5) shows the close and far distance calculated by equation (4), which are taken as the constraint condition of the classification. According to [5,7], liver and spleen is stable and easy to be located. So liver and spleen will be calculated more accurate. Besides, Right kidney and gallbladder are decided by liver. Left kidney is related to spleen. Then, we establish an directed adjacency graph of visual pitches shown in Figure 3(a), and define the distance calculated by equation (4) as the weight of two visual pitches to implement the spatial relationship, which also contains the location information.

$$d_{pitch} = \left\| P_1 - P_2 \right\|^2 \tag{4}$$

$$\begin{cases} d_{close} < 2S \\ d_{far} > 3S \end{cases} \tag{5}$$

d_{pitch} is the distance of two visual pitches. P is the position (x, y) of the center pixel of the visual pitch. $\|.\|$ is L_2 norm. The side-length of the cluster region is 2S.

When constructing the training sample set for classification, which contains the abdominal organs like liver, kidney, gallbladder, spleen as well as the backgrounds, we should abandon the visual pitches whose intensity is 0 because of the black background of the medical image, such as the red pitch in Figure 2(a). When all the preparations are ready, it is time to start the classification stage, which results in labels of the visual patches. Finally, we utilize the spatial constraints to revise the label of patches. The result of classification segmentation based on visual patch is shown in Figure 2(c) and Figure 3(b), which are filled and isolated.

4 Experimental Results

4.1 Experimental Results of Visual Patches Generation

In order to verify the good performance of visual patches generation, we compare the result with ground truth and the traditional clustering methods in American TCIA database. Experiment shows that the result of the visual patches generation segmentation is closest to the ground truth than the other state-of-the-art algorithms, which is shown in Figure 6. Figure 4 is the result of the visual patches generation and Figure 6(d) is the regions of right kidney which is filled and isolated. DP method took advantage of the Bayesian methodologies and Markov chain to enhance model flexibility, which ignores the intensity overlap and spatial relationship [8]. Although C-GMMs algorithm made use of the powerful probabilistic statistical theory, it also spent time with the convergence function, which was proposed for image segmentation using the feather function to estimate the parameters of GMMs in each iterative [9]. Moreover, when compared to the traditional clustering segmentation methods, our algorithm select the lowest gradient pixel as the initial cluster centers of each visual patch, which no longer rely on the random and manual selection, and largely avoid the effects of noise and intensity inhomogeneity. It also considers the spatial relationship of pixels during the clustering, which inhibits the phenomenon of under-segmentation and over-segmentation significantly, and at the same time retains all of the information of the target area in the original image. Therefore, visual patches generation provides a good basis for clinical medicine diagnosis.

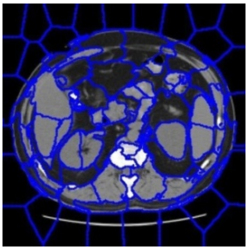


Fig. 4. Visual patches generation

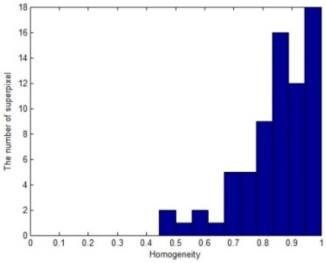
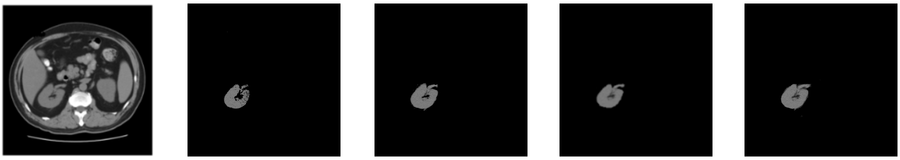


Fig. 5. Homogemeity



(a) Original image (b) DP[8] (c) C-GMM[9] (d) Visual patch (e) Ground turth

Fig. 6. Results of kidneys segmentation by visual patch generation

Besides, we put forward a new evaluate criteria for visual pitch. As known to us all, homogemeity is a measure of the change of texture, and the greater its value the texture is more uniform. Due to the homogemeity ranges from 0 to 1, we can calculate from Figure 5 that almost 80% of the homogemeity of visual pitches, regardless of the black background patches, are more than 0.8, and the rest are resulting from the uneven background clustering by spatial proximity. From this we can also prove the good performance of visual patches when used in small feather space of medical image.

4.2 Result of Multi-object Classification Segmentation

Traditional classification segmentations rely on the whole organ region feathers, whereas the method in this paper regards the organ patch as the classification unit. From the visual patch generation we can achieve several kinds of patches, each of which is likely to be a small part of the organs. Thus, it is easy to be high quality feathers without background, and the classification will be benefit from them. During the experiment, the method is validated on 200 CT sequence totally about 14000 slices which are opened in American TCIA database. Figure 7 shows the average classification accuracy of each organ. As the number of sample growing, the accuracy is gradually increasing and finally stabilized. Because the frequency of occurrence of gallbladder and spleen is lower than the liver and kidney in abdominal CT series, and their change is relatively large, the segmentation accuracy of them is lower. Table 1 shows the result of the segmentation of organs compared to the state-of-art method tested in American TCIA database, which is calculated by equation (6). T is the number of is the true positive pixels, and F is the number of the false negative pixels.

$$Accuracy = \frac{\mathrm{T}}{\mathrm{T+F}} \tag{6}$$

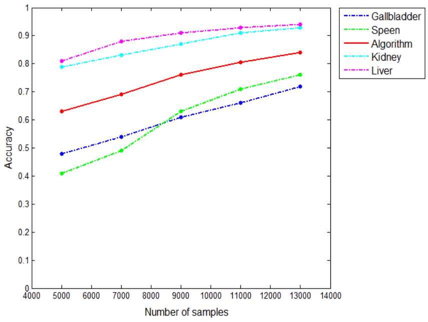


Fig. 7. The accuracy of our method.

Organs /Method	TG [6]	LVP [7]	Our method
Liver	**0.80**	**0.89**	**0.92**
Spleen	**0.55**	**0.72**	**0.78**
Kidney	**0.74**	**0.82**	**0.90**
Gallbladder	**0.38**	**0.56**	**0.73**

Table 1. The segment accuracy of each organ

From Table 1 we can see that classification based on visual pitch considering the spatial relationship of visual pitches is proved to be an excellent measure for segmentation. TG [6] using the threshold and gradient is inevitably affected by noise and intensity inhomogeneity. Besides, it causes over-segmentation with the intensity as the only feather when happened to grayscale overlapping. Although the LPV [7] takes the spatial relationship into consideration, it is highly depend on the inter-organ modeling by a minimum volume overlap constraint, which does not apply to individual diversity and also result in over- and under-segmentation. Our multi-object segmentation method, which combines the image low-level features and the spatial location information, is the fusion of visual features and physical characteristics. We test these methods in American published TCIA database, and the performance of our method is proved higher than the other state-of-art algorithms. Figure 8 shows the visualization result of the comparative algorithms.

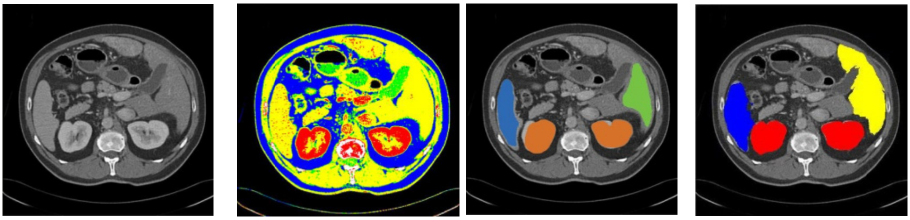

(a)Original image (b)TG[6] (c)LPV[7] (d)Our method

Fig. 8. Results of comparative algorithms segmentation

5 Conclusion

The method we presented is the combination of top-down method clustering and bottom-up method classification. In the visual patches generation phase, the pixels are clustered by intensity similarity and spatial proximity, which finally obtains the compact and uniform visual patches and inhibits the phenomenon of under-segmentation and over-segmentation. At next stage, with the difference of the traditional classification segmentation algorithms, our method regards the visual patch as classification unit instead of the pixels, which improves the speed of the segmentation, and takes the spatial relationship of visual patches into consideration to increase the accuracy of the classification. Therefore, multi-object segmentation based on visual patches classification is not only designed for multi-object segmentation, but also greatly increases the accuracy of segmentation. In conclusion, our image segmentation techniques have improved the segmentation accuracy segmentation speed, quality, multi-object and general applicability.

References

1. Withey, D.J., Koles, Z.J.: A Review of Medical Image Segmentation: Methods and Available Software. International Journal of Bioelectromagnetism **10**(3) (2008)
2. Selver, M.A., Kocaoglu, A., Akyar, H., Dicle, O., Guzelis, C.: Patient oriented neural networks to overcome challenges of abdominal organ segmentation in CT angiography studies. Electrical and Electronics Engineering, pp. 5–8, November 2009
3. Freedman, D., Radke, R.J., Zhang, T., Jeong, Y., Chen, G.T.Y.: Model-based multi-object segmentation via distribution matching. In: Computer Vision and Pattern Recognition Workshop (2004)
4. Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D.: Automated Abdominal Multi-Organ Segmentation With Subject-Specific Atlas Generation. Medicine and Biology Society, 1723–1730 (2013). doi:10.1109
5. Okada, T., Linguraru, M.G., Hori, M., Suzuki, Y., Summers, R.M., Tomiyama, N., Sato, Y.: Multi-organ segmentation in abdominal CT images. doi:10.1109/EMBC.2012
6. Liu, Y., Liang, G., Saha, P.K.: A new multi-object image thresholding method based on correlation between object class uncertainty and intensity gradient. Medical Physics, January 2012

7. Liu, X., Linguraru, M.G., Yao, J., Summers, R.M.: Abdominal multi-organ localization on contrast-enhanced CT based on maximum a posteriori probability and minimum volume overlap. Biomedical Imaging (2011). doi:10.1109
8. Adelino, R.F.D.S.: Bayesian mixture models of variable dimension for image segmentation. Compute. Meth. Prog. Biomed. **94**, 1–14 (2009)
9. Luo, S., Hu, Q., He, X., Li, J., Jin, J.S., Park, M.: Automatic liver parenchyma segmentation from abdominal CT images using support vector machines. Complex Medical Engineering, 1–5 (2010). doi:10.1109
10. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. Pattern Analysis and Machine Intelligence (2012)
11. Bo, H., Ma, F.L., Jiao, L.C.: GLCM of Image Texture. Acta Electronica Sinica 01, 155–158 (2006)

Improved 3D Local Feature Descriptor Based on Rotational Projection Statistics and Depth Information

Hui Zeng(✉), Rui Zhang, and Mingming Huang

School of Automation and Electrical Engineering,
University of Science and Technology Beijing, Beijing 100083, China
hzeng@163.com

Abstract. 3D local feature descriptor construction is a very challenging task in the field of 3D model analysis. In this paper, an improved Rotational Projection Statistics (IRoPS) descriptor is proposed. For each feature point, the local coordinate system is firstly built and its neighboring points are normalized. Then the normalized neighboring points are rotated and projected onto three coordinate planes. For each rotation, the distribution matrix is computed and the sub-descriptor can be obtained using the central moment, the Shannon entropy, the mean and variance of local depth values. Finally the IRoPS descriptor is constructed by concatenating all the sub-descriptors into a vector. Compared with the Rotational Projection Statistics (RoPS) descriptor, the IRoPS descriptor includes the local depth information and it has better discriminative power. Extensive experiments are performed to verify the superior performance of the proposed descriptor.

Keywords: 3D model matching · Rotational projection statistics · Local feature descriptor · Depth information

1 Introduction

As the increasing development of the computer vision theory, 3D model has gradually become the fourth type of multimedia data following the sound, image and video. Compared with image, 3D model contains much more information, which could be more conducive to the analysis and understanding of the scene. Therefore, 3D model has been applied in more and more fields, such as virtual reality, 3D games, building design, animation effect and medical diagnosis. Especially in recent years, with the emergence of 3D printer, the application of 3D model has begun to spread to households, which making the home users can print 3D model [1]. As a result, how to quickly search for the required model from so many 3D models has become an urgent research topic.

For most 3D model retrieval methods, 3D model matching is one of the basic steps [2]. Therefore, this paper mainly investigates the construction of the 3D model local descriptor and its application in 3D model matching. In general, model matching

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 139–147, 2015.
DOI: 10.1007/978-3-662-48570-5_14

includes the following process. Firstly, the feature descriptor is extracted from the 3D model. Then the similarity between the original model and the corresponding model is calculated to accomplish the 3D model matching. Up to now, according to the different extracted methods of feature descriptor, there are two kinds of methods to solve this problem: global matching methods and local matching methods. Global matching methods represent global shape of the entire 3D model as a feature descriptor, such as geometric distribution based methods [3], skeleton graph matching based methods [4] and projection map based methods [5], etc. However, when 3D models have similar local structure, pose variations and partial occlusions, global matching is difficult to achieve better performance. Over the past two decades, the most popular trend for 3D model matching is to use local features, due to their better robustness and descriptiveness. The local features contain abundant shape information, and they are not easily influenced by external environment, such as pose and lighting variations, shape deformation and occlusions. Local matching methods mainly have two kinds of methods. The first ones are feature points description based local matching methods. The keypoints are detected and their local descriptors are constructed for matching. This kind of methods is relatively robust to noise, but the global topology information is not been fully used in the process of matching. The second ones are topology description based local matching methods. This kind of methods obtains the corresponding local geometry features for sub-graph matching. But the normal vector and curvature calculation would be existed in descriptor construction, resulting that this kind of methods would be less robust to noise.

From the above analysis we can see that the construction of 3D local descriptor is one of key research topics of 3D local matching methods. So, in this paper, our research emphasis is 3D local descriptor construction. Up to now, much work about 3D local descriptor construction has been published. Chua et al. proposed point signature based feature descriptor [6]. They obtained a contour C by intersecting the 3D model surface with a sphere of radius r centered at the keypoint p. Then, they fitted a plane to these contour points. The distances between contour points and fitting plane would be used for descriptor construction. Johnson et al. proposed spin image based feature descriptor [7]. They used the normal vector of a keypoint p as the local reference axis to build a cylindrical coordinate system. A spin image based feature descriptor is generated by projecting a local surface onto the 2D cylindrical plane. Mian et al. proposed tensor based feature descriptor [8]. Firstly, local points were preprocessed by triangular mesh method. They then constructed a local 3D grid and summed the surface areas in each bin of the grid, to generate a "3D tensor" descriptor. Recently, Yulan Guo et al. proposed a novel local feature extraction algorithm, named Rotational Projection Statistics (RoPS) [9], and its performance is better than five state-of-art descriptors, including spin image [7], normal histogram (NormHist) [10], Local Surface Patch (LSP) [11], THRIFT [12] and signature of histograms of orientations (SHOT) [13]. The RoPS descriptor is generated by rotationally projecting the neighboring points onto three local coordinate planes and calculating several statistics (central moment and Shannon entropy) of the projected points, showing both high discriminative power and strong robustness to noise. However, this method only focused on the distribution statistics of the feature points, without considering the depth information of the projected points.

In this paper, in order to make full use of 3D models' local depth information, we propose an improved local feature descriptor, named improved Rotational Projection Statistics (IRoPS) descriptor. The IRoPS descriptor use the frame of RoPS descriptor, and the central moment and the Shannon entropy are also used for the sub-descriptor construction. The improvement of the IRoPS descriptor is that the mean and variance of local depth values are added in the process of sub-descriptor construction. Compared with the RoPS descriptor, the IRoPS descriptor can depict the 3D local structure more comprehensively. For two different local neighborhoods with different local depth information, they may have same RoPS descriptor and have different IRoPS descriptors. Comparative experiments have been performed and the results demonstrate the effectiveness and efficiency of our proposed IRoPS descriptor compared with RoPS descriptor.

The rest of this paper is organized as follows. Section 2 provides the description of the improved RoPS descriptor. Section 3 presents the results and analysis of 3D model matching experiments. Section 4 concludes this paper.

2 Improved Local Surface Descriptor construction

The construction of the improved RoPS descriptor includes the two following steps. Firstly, the 3D local coordinate system of each 3D feature point is built, which can provide invariance to 3D translation and rotation. Then the local feature descriptor is constructed, using the central moment, the Shannon entropy, the mean and variance of depth values of the projected points.

2.1 3D Local Coordinate System Definition

In this paper, we use the local reference frame proposed in [9] to build unique 3D local coordinate system for every 3D feature points, which can provide invariance to 3D rotation and transformation for the following local descriptor construction. For every feature point *p*, given a support radius *r*, the local neighborhood can be determined by cropping a sphere of radius *r* centered at *p*. For the i_{th} triangle with three vertices p_{i1}, p_{i2} and p_{i3}, the scatter matrix C_i can be defined as:

$$C_i = \tfrac{1}{12}\sum_{j=1}^{3}\sum_{k=1}^{3}(p_{ij}-p)(p_{ik}-p)^T + \tfrac{1}{12}\sum_{j=1}^{3}(p_{ij}-p)(p_{ij}-p)^T \tag{1}$$

Assume *N* is the number of triangles, the overall scatter matrix *C* is defined as:

$$C = \sum_{i=1}^{N} w_{i1} w_{i2} C_i \tag{2}$$

where $w_{i1} = \frac{|(p_{i2}-p_{i1})\times(p_{i3}-p_{i1})|}{\sum_{i=1}^{N}|(p_{i2}-p_{i1})\times(p_{i3}-p_{i1})|}$ $w_{i2} = (r - |\, p - \frac{p_{i1}+p_{i2}+p_{i3}}{3} |)^2$.

Then the descending eigenvectors $\{v_1, v_2, v_3\}$ of C are calculated, which offer a basis for local coordinate system definition. In order to eliminate the sign ambiguity, the unambiguous vector v_1 is defined as:

$$v_1 = v_1 \cdot sign(h) \tag{3}$$

where $h = \sum_{i=1}^{N} w_{i1} w_{i2} (\frac{1}{6} \sum_{j=1}^{3} (p_{ij} - p) v_1)$. Similarly, we can get the unambiguous vector v_2 and v_3. Consequently, 3D local coordinate system for given point p is finished, that p is the origin, v_1, v_2 and v_3 are the x, y and z axes respectively.

2.2 Improved RoPS Descriptor Construction

The RoPS (Rotational Projection Statistics) descriptor is a novel 3D local descriptor proposed by Yulan Guo et al. And it can be generated by rotationally projecting the neighboring points around a feature point onto three coordinate planes and calculating the statistics of the distribution of the projected points. Although extensive experiments have testified its superior performance, it does not make full use of the local structure information of the 3D point's neighborhood. Only the number of projecting points in each bin is used for the RoPS descriptor construction, without considering the depth information of the projection points. For two different local neighborhoods with same RoPS descriptor, they may have different local depth variance. The local depth information of the 3D local neighborhood is important for depicting 3D local structure. So we propose an improved RoPS descriptor to make the descriptor contains the depth information.

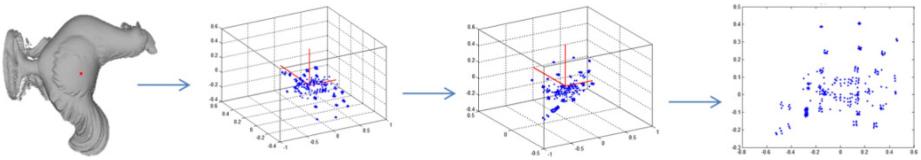
The construction process of a RoPS descriptor is shown in Fig. 1. Given a feature point p and its support radius r, local neighborhood of the feature point can be extracted and its local coordinate system can be determined using the method given in section 2.1. After rotation and translation normalization, the local neighborhood is rotated around every axis by an angle 2π/T each time, where T denotes the number of rotations. Then the local neighboring points are projected on xy plane xz plane and yz plane respectively. For each projection, the bounding rectangle of the projected points can be divided into L×L bins, and the number of points falling into each bin is counted to yield an L×L distribution matrix D. The distribution matrix D is then normalized to let the sum of its elements be 1, and it can make the descriptor have invariance to variations in mesh resolution. Then the central moment and the Shannon entropy are computed from the matrix D. The central moment μ_{mn} of the matrix D can be calculated using the following equation:

$$\mu_{mn} = \sum_{i=1}^{L} \sum_{i=1}^{L} (i - \bar{i})^m (j - \bar{j})^n D(i, j) \tag{4}$$

where, $\bar{i} = \sum_{i=1}^{L} \sum_{j=1}^{L} iD(i, j)$, $\bar{j} = \sum_{i=1}^{L} \sum_{j=1}^{L} jD(i, j)$. The Shannon entropy e of the matrix D can be calculated as:

$$e = -\sum_{i=1}^{L}\sum_{j=1}^{L} D(i,j)\log(D(i,j)) \tag{5}$$

For each rotation and projection, the sub-descriptor can be obtained using $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, e\}$. Finally the RoPS descriptor can be obtained by concatenating all the sub-decriptors into a vector.



(a) 3D model (b) local neighborhood (c) rotated points (d) projected plane

Fig. 1. An illustrative example of the RoPS method.

The RoPS method represents local points from a set of views, converting the matching between 3D models to that between the 2D projected planes. Under this frame, many 2D image matching methods can be used for 3D model matching. However, the difference between 3D model and 2D image is that 3D model has abundant depth information. The RoPS descriptor only uses the distributed information of the projected points from different views. The points with similar position in projected plane would have different depths in 3D local coordinate system. Unavoidably, the RoPS descriptor loses part of the 3D spatial structural information. In the history of machine vision, psychologist pointed that the human visual system uses a lot of depth information based on visual sense to understand and identify objects [14]. This give an important enlightenment: computer vision researchers can directly investigate the visual system based on depth information. As we all know, snapshot descriptor is obtained by using depth information. Malassiotis and Strintzis [19] first constructed an 3D local coordinate system by performing an eigenvalue decomposition on the covariance matrix of the neighboring points of a keypoint p. They then placed a virtual pin-hole camera at a distance d on the z axis and looking toward p. The x and y axes of the camera coordinate frame were also aligned with the x and y axis of the 3D local coordinate system at p. They projected the local surface points onto the image plane of the virtual camera and recorded the distance of these points from the image plane as a ”snapshot” descriptor. The snapshot descriptor is robust to self-occlusion and very efficient to compute. Snapshot achieved better pairwise range image alignment results compared to spin image. Mian et al. [20] also defined an 3D local coordinate system for a local surface and then fitted the local surface with a uniform lattice. They used depth values of the local surface to form a feature descriptor, which was further compressed using a PCA technique.

Inspired by this conclusion and some previous work, we improve the RoPS descriptor using the statistic of depth information. For each distribution matrix *D*, its mean and variance of depth information can be defined as:

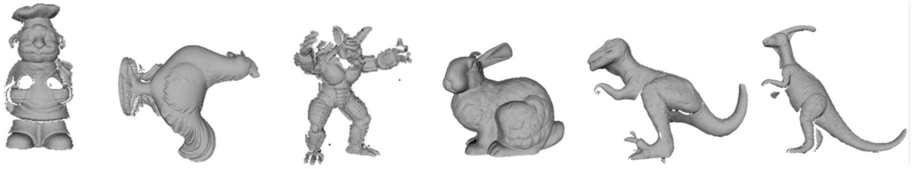
$$d_m = \frac{\sum_{i=1}^{num} | d_i |}{num} \tag{6}$$

$$d_v = \frac{\sum_{i=1}^{num} (| d_i | - d_m)^2}{num} \tag{7}$$

where d_i is the i_{th} point's depth value, *num* is the points' number. The mean value $\mathrm{d_m}$ can measure the distance between the local point set and its center point, and the variance value $\mathrm{d_v}$ can measure how far the local points are spread out. Therefore, they are useful for describing the 3D local structure. Finally, the sub-descriptor of each projection can be obtained using $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, e, d_m, d_v\}$ and the improved RoPS descriptor can be generated by concatenating all the sub-descriptors into a vector.

3 Experimental Results

In this section, comparative experiments are performed to verify the superiority of the proposed method. As shown in Fig. 2, the experimental data are six models ("Armadillo", "Bunny", "Chicken", "T-rex", "Parasaurolophus" and "Chef"), which were taken from the Stanford 3D Scanning Repository [15] and Mian's Dataset [16,17]. The corresponding model was synthetically generated by randomly rotating in order to create clutter and pose variances. Then Gaussian noise with a standard deviation of 0.1 mesh resolution was added to the model.

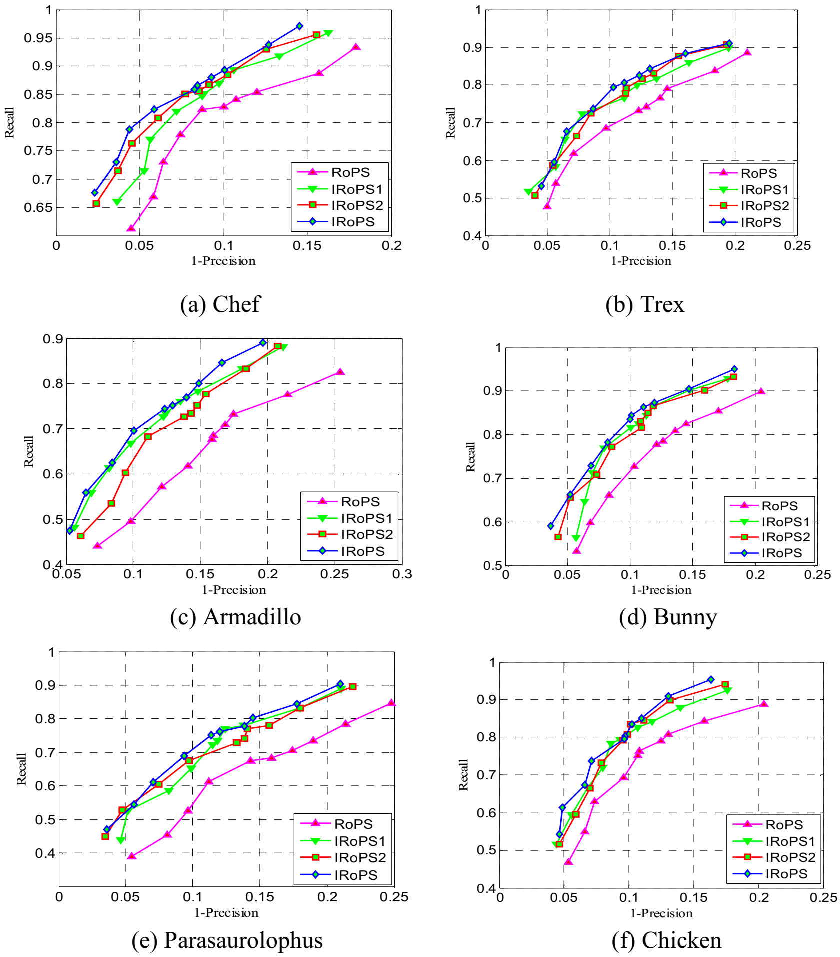


(a) Chef (b) Chicken (c) Armadillo (d) Bunny (e) Parasaurolophus (f) T-rex

Fig. 2. The experimental data.

In the experiments, we use the parameters according to Guo's suggestion in [8], which means the parameters are $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, e\}$, L=5, r=15mr (mesh resolution) and T=3. At first, 1000 feature points are randomly selected from the 3D model. Then the local descriptors are constructed for each feature point. Finally the feature points of two 3D models are matched by computing the distances of descriptors. To evaluate the effectiveness of the depth information, we compare four kinds of statistics to consruct the local descriptors.

The experimental results are evaluated using the Recall-Precision criteria based on the number of the correct matches and the number of the false matches. Fig. 3 is the matching results of the testing 3D models. Here “RoPS” denotes the descriptor proposed by Yulan Guo et al. and it is constructed by $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, e\}$,“IRoPS1” denotes the descriptor constructed by $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, e, d_m\}$, “IRoPS2” denotes the descriptor constructed by $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, e, d_v\}$, and “IRoPS” denotes the descriptor constructed by $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, e, d_m, d_v\}$. The goal in PR space is to be in the upper-left-hand corner, which means the better performance in 3D model matching [18]. From Fig. 3 we can see that the performance of the IRoPS1 descriptor is better than the performance of the IRoPS2 descriptor, and the IRoPS descriptor performs best. That is to say, the mean of the depth values has more discriminative information than the variance of the depth values for the local descriptor construction, and the IRoPS descriptor can character the local structure of the 3D model effectively.



(a) Chef

(b) Trex

(c) Armadillo

(d) Bunny

(e) Parasaurolophus

(f) Chicken

Fig. 3. The experimental results.

4 Conclusion

In this paper, a novel 3D local feature descriptor called IRoPS is proposed. It is the improvement edition of the RoPS descriptor. Similar to the RoPS descriptor, the IRoPS descriptor is invariant to rotation and translation by normalization using the local coordinate system. Compared to RoPS descriptor, the IRoPS descriptor not only contains the distribution information of the projected points but also contains the local depth information. It has better discriminative power, while maintaining the robustness of the RoPS descriptor. To verify the performance of the proposed descriptor, extensive 3D model matching experiments have been performed. The experimental results show that the IRoPS descriptor better performance than the RoPS descriptor. In our future work, we will search more discriminative information of the 3D local surface to improve the performance of the descriptor.

Acknowledgments. This article is supported by the National Natural Science Foundation of China (Grant No. 61375010, No.61005009), Beijing Higher Education Young Elite Teacher Project (Grant No.YETP0375) and the Fundamental Research Funds for the Central Universities under the Grant No. FRF-TP-14-120A2. The authors would like to acknowledge Yulan Guo for providing the code of RoPS descriptor.

References

1. Waran, V., Narayanan, V.: Utility of multimaterial 3D printers in creating models with pathological entities to enhance the training experience of neurosurgeons: Technical note. Journal of Neurosurgery **120**(2) (2014)
2. Wang, M., Gao, Y., Lu, K., Rui, Y.: View-based discriminative probabilistic modeling for 3d object retrieval and recognition. IEEE Transactions on Image Processing **22** (2013)
3. Ohbuchi, R., Otagni, T., Ibato, M., et al.: Shape similarity search of three-dimensional models using parameterized statistics. In: Proceedings of the Pacific Graphics, pp. 265–274. Beijing, China (2002)
4. Sundar, H., Silver, D., Gagvani, N., et al.: Skeleton based shape matching and retrieval. In: Shape Modeling International, pp. 130–139. IEEE (2003)
5. Chen, D.-Y., Tian, X.-P., Shen, Y.-T., et al.: On Visual Similarity Based 3D Model Retrieval. Eurographics V22(3) (2003)
6. Chua, C., Jarvis, R.: Point signatures: a new representation for 3Dobject recognition. International Journal of Computer Vision **25**(1), 63–85 (1997)
7. Johnson, A.E., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(5), 433–449 (1999)
8. Mian, A., Bennamoun, M., Owens, R.: Three-dimensional model-based object recognition and segmentation in cluttered scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(10), 1584–1601 (2006)
9. Guo, Y., Sohel, F., Bennamoun, M., Min, L., Wan, J.: Rotational Projection Statistics for 3D Local surface description and object Recongnition. Int. J. Comput. Vis. **105**, 63–86 (2013)

10. Hetzel, G., Leibe, B., Levi, P., Schiele, B.: 3D object recognition from range images using local feature histograms. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, issue II, p. 394 (2001)
11. Chen, H., Bhanu, B.: 3D free-form object recognition in range images using local surface patches. Pattern Recognition Letters **28**(10), 1252–1262 (2007)
12. Flint, A., Dick, A., Van den Hengel, A.: Local 3D structure recognition in range images. IET Computer Vision **2**(4), 208–217 (2008)
13. Tombari, F., Salti, S., Di Stefano, L.: Unique signatures of histograms for local surface description. In: European Conference on Computer Vision, pp. 356–369. Crete, Greece (2010)
14. Okoshi, M.T.: Depth Cues in the Human Visual System, Three-Dimensional Imaging Techniques. Academic Press, New York (1976)
15. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: 23rd Annual Conference on Computer Graphics and Interactive, Techniques, pp. 303–312. New Orleans, LA (1996)
16. Mian, A.S., Bennamoun, M., Owens, R.: A Novel Representation and Feature Matching Algorithm for Automatic Pairwise Registration of Range Images. International Journal of Computer Vision **66**(1), 19–40 (2006)
17. Mian, A., Bennamoun, M., Owens, R.: 3D Model-based Object Recognition and Segmentation in Cluttered Scenes. IEEE Transactions in Pattern Analysis and Machine Intelligence **28**(10), 1584–1601 (2006)
18. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of 23rd International Conference on Machine Learning, Pittsburgh, PA (2006)
19. Malassiotis, S., Strintzis, M.: Snapshots: A novel local surface descriptor and matching algorithm for robust 3D surface alignment. IEEE Trans. Pattern Anal. Mach. Intell. **29**(7), 1285–1290 (2007)
20. Mian, A., Bennamoun, M., Owens, R.: On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes. Int. J. Comput. Vis. **89**(2), 348–361 (2010)

Pedestrian Tracking in Infrared Image Sequence Based on Spatio-temporal Slice Trajectory Analysis

Dongmei Fu and Pan Shu(✉)

School of Automation and Electrical Engineering,
University of Science and Technology Beijing, Beijing 100083, China
panshupansy@gmail.com

Abstract. The research of pedestrian tracking algorithm in infrared image sequence is a curial part in video surveillance. But due to the special characteristics of infrared image, such as low contrast, fuzzy edge and unknown noises interference, the study of infrared pedestrian tracking algorithm becomes a great challenge. Spatio-temporal slice method is an effective analysis method considering both space and time. In view of the existing tracking methods based on spatio-temporal slice only considering horizontal slice analysis and usually with large amount of calculation, this paper proposes a novel tracking algorithm based on spatio-temporal slice trajectory analysis. The proposed algorithm uses multi-layer horizontal and vertical slices to obtain the complete trajectories in order to determine the information of target boundary and position, and then tracks the target in each frame of the infrared image sequence. The experimental results show that the algorithm has relatively high tracking accuracy at a fast computing speed. Moreover, it performs effectively in different infrared image sequences with various motion modes by single pedestrian from OTCBVS/05 Terravic Motion IR Database, namely, the algorithm has good robustness to some extent.

Keywords: Infrared image sequence · Spatio-temporal slice · Trajectory analysis · Pedestrian tracking

1 Introduction

Pedestrian tracking is a very important research hotspot in computer vision field and tracking pedestrian targets in infrared (IR) image sequences is especially a crucial part in many military or civil fields, such as video surveillance, position orientation and behavior analysis. However, the gray contrast between the targets and the background in IR image is quite low and the background in IR image is usually inevitably contaminated by unknown noises. Moreover, the distant surveillance determines that the tracked targets are usually small and blurred. Hence, developing an effective infrared pedestrian tracking algorithm is a great challenge.

Many researchers have focused on the tracking of infrared targets and numerous tracking algorithms have been proposed in this field, such as Meanshift [1], Camshift [2], histograms of oriented gradients [3] feature, SVM Classifier [4], local binary patterns [5], particle filter [6] and spatio-temporal slices analysis, etc. Among these

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 148–158, 2015.
DOI: 10.1007/978-3-662-48570-5_15

methods, spatio-temporal slices analysis has gained special attention for its unique advantages. Compared to traditional methods, it uses long time scale of information rather than a small amount of short time frames in image sequence for data processing, which enriches the visual information and track information and improves the tracking accuracy of the algorithm. Ngos, etc. [7] uses structure tensor histogram to do a statistics of the visual information in spatio-temporal slices, and do several video segmentations in space domain to realize the detection of foreground and background. Sato, etc. [8] proposes a TSV (Temporal Spatio - Velocity) transform method to calculate the rates of pixels, and to extract target in video image by binary TSV in order to implement the target tracking and interactive behavior recognition. Jingjing Yang, etc. [9] presents a slice-based approach for pedestrian detection in still images. Limited numbers of horizontal spatio-temporal sub-regions are employed to represent pedestrians. Meanwhile, a classifier is constructed to classify multiple sub-regions.

The existing tracking methods based on spatio-temporal slice almost only consider horizontal slice analysis, and usually have complex associated algorithms. Besides, they generally need a large amount of calculation. Considering the above mentioned, in this paper, a novel tracking algorithm based on spatio-temporal slice trajectory (both in horizontal and vertical direction) analysis is proposed by analyzing the spatio-temporal slice trajectory superposed by multiple layers to obtain target boundary and position in order to realize reliable and accurate infrared pedestrian tracking.

The remainder of the paper is arranged as follows. A brief review on spatio-temporal slice is provided in Section 2. Section 3 describes the detailed principles of the proposed tracking algorithm. Experimental results and conclusions are given in Sections 4 and 5.

2 Spatio-temporal Slices

The concept of spatio-temparal slice was initially proposed in 1985 by E. H. Adelson and J. Bergen [10] in the article ‘Spatiotemporal energy models for the perception of motion ’. Assuming a video as a 3-D image sequence in which the three dimensions are respectively the x, y, and time t, if you do segmentation along the axis of t, then the cross section is rightly the so-called spatio-temparal slice.

Spatio-temporal slice images record the history of long time scales of video information. There are usually three kinds of spatio-temparal slice depending on the segmentation direction: vertical, horizontal and diagonal slice. Among these slices, vertical and horizontal slices are more commonly used. One dimension of the cross section is time t, another dimension is x or y. If the cross section parallels to the x axis, it is called a vertical slice; if the cross section parallels to the y axis, then it is called a horizontal slice. The illustration of horizontal and vertical spatial temporal slices is shown in Fig.1. The image sequence includes 160 frames (T=160) and the size of each frame is 240 × 360 (M=240, N=360) pixels.

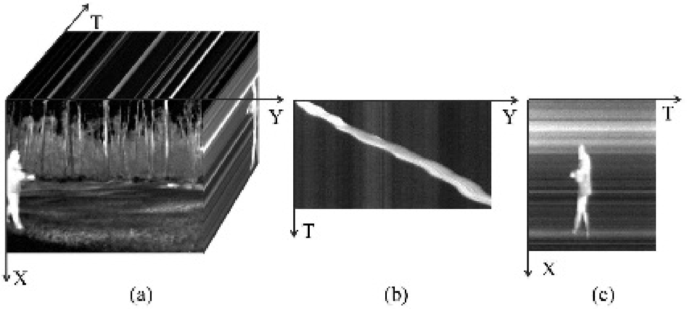


Fig. 1. Illustration of slicing: (a) image sequence (b) a horizontal slice (c) a vertical slice

3 Proposed Tracking Algorithm

In this paper, a novel tracking algorithm based on spatio-temporal slice trajectory analysis is proposed. First of all, multiple horizontal slices at different heights are obtained, then we respectively use background subtraction method to extract target trajectory in each slice. Next, do a superposition of multi-layer trajectories in order to get the complete horizontal trajectory. Meanwhile, the complete vertical target trajectory can be gained by using the similar process (instead, vertical slices are obtained from different widths).And then, vertical coordinate and height of moving target can be obtained via analyzing horizontal trajectory and abscissa along with width from vertical trajectory. Finally, tracking pedestrian with a rectangular box which is also called bounding box can be realized by using location information from the previous step. The process of this algorithm is depicted in Fig. 2.

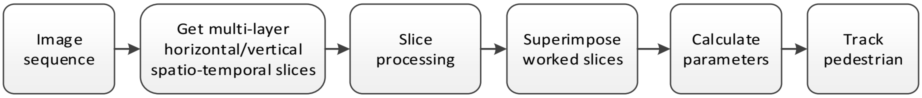


Image sequence
Get multi-layer horizontal/vertical spatio-temporal slices
Slice processing
Superimpose worked slices
Calculate parameters
Track pedestrian

Fig. 2. Flow chart of the proposed tracking algorithm

3.1 Multi-layer Spatio-temporal Slices Acquisition

First of all, assume a surveillance video as an XYT 3-D image sequence, where x and y are the image dimensions and t is the temporal dimension. Select one row of all pixels continuously from each frame in the 3-D image sequence at the same height-level, and put them together by order of dimension t which forms a horizontal slice. It can be seen as an image with dimension y and the temporal dimension t. In the same way, select one column of all pixels from each frame in the 3-D image sequence at the same width-level we can gain a vertical slice.

Take a video sequence with T frames for example, if the size of each frame is $M \times N$, then the size of one horizontal slice is $M \times T$, and the size of one vertical slice is $N \times T$. Assume that I is a video sequence and *Ii* represents each frame (*i* ranges from 1 to *T*), if we respectively select one row at the height of m and one column at the width of n, then the horizontal slice I_H and vertical slice I_V can be represented as the following formulas.

$$\begin{cases} I_H\left(i,:\right) = I_i(\mathrm{m},:) \\ I_V\left(:,i\right) = I_i(:,\mathrm{n}) \end{cases} \tag{1}$$

In order to gain complete trajectories, different rows and columns should be selected at an abundant and appropriate amount in the proper region. The amount of slices will have influence on the performance of this algorithm. It will be discussed in Section 4.

3.2 Slice Processing

In order to extract trajectory from slice image, background subtraction method is selected. So background modeling is a key issue to this method. The complete process is shown in Fig.3.

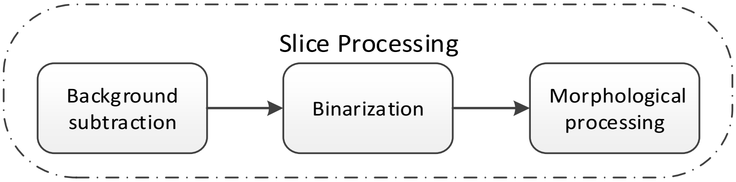


Fig. 3. Process of slice processing

There are many kinds of background modeling methods, such as Gaussian Mixture Model (GMM) [11], Codebook [12] Algorithm and so on. If one slice image is seen as a sequence of   lines in the time domain, then according to the reference definition of foreground and background in video sequence, one line is equivalent to one frame in a sequence and the trajectory is equivalent to the target section. In this section, single Gaussian Model was used for multi-layer slices background modeling. Each layer of the spatio-temporal slice corresponds to a row background model. In the same layer, the y value of all pixels are fixed, but in different layers y value are different. Therefore, on the whole, layers of time slice is still a two-dimensional Gaussian background model, which can be represented as the following formula.

$$I_B\left(x,y\right) \sim N\left(\mu,\sigma^2\right) \tag{2}$$

Assume B is a pixel from background in one slice, and x is its abscissa y is its ordinate, μ is average value, σ is variance.

Take the initial m rows in a single layer slice, μ_0 and σ_0 are initial average value and variance, and it can be obtained as the formula below.

$$\begin{cases} \mu_0 = \frac{1}{m}\sum_{t=1}^{m} x_t \\ \sigma_0^2 = \frac{1}{m}\sum_{t=1}^{m}(x_t - \mu_0)^2 \end{cases} \tag{3}$$

Then Judge by the formula (5) of pixel matching, if meet, then the pixels should update according to formula (6), otherwise, remain the same.

$$|x_t - u_t| < 2.5\sigma_t \tag{4}$$

$$\begin{cases} \mu_t = (1-\alpha)\mu_{t-1} + \alpha X_t \\ \sigma_t^2 = (1-\alpha)\sigma_{t-1}^2 + \alpha(X_t - \mu_t)^T (X_t - \mu_t) \end{cases} \tag{5}$$

Among them, α is the update rate of the background. From this process of Gaussian modeling, we can gain the background B_H and B_V. Then each line in one slice image make subtraction with this model B_H and B_V, the new slice is almost the trajectory T_H and T_V.

$$\begin{cases} T_H(\mathrm{i,j}) = |I_H(\mathrm{i,j}) - B_H(\mathrm{i,j})| \\ T_V(\mathrm{i,j}) = |I_V(\mathrm{i,j}) - B_V(\mathrm{i,j})| \end{cases} \tag{6}$$

Then the trajectory should be converted to binary image at an appropriate threshold.

$$T_{H/V}(\mathrm{i,j}) = \begin{cases} 1 & T_{H/V}(\mathrm{i,j}) \geq Threshold \\ 0 & otherwise \end{cases} \tag{7}$$

At last, some morphological processing were done on the trajectories, such as median filtering, open operation and close operation in order to obtain clear multi-layer trajectories.

3.3 Trajectory Extraction

The appearance of trajectory has its certain uncertainty. However, when the target motion in the video, the different parts of the same target at any of the same moment have similar levels of coordinates or abscissa as well as a consistent motion mode or movement trend. Thus, in the trajectory superposition phase, the adjacent layer trajectory images with similar levels of coordinate or abscissa in the path do a superposition then the complete trajectory can be obtained. This process can eliminate the target area of the shade or loss caused by the trajectory of fracture to some extent.

$$\begin{cases} TH = \sum_{h=1}^{rows}\sum_{i,j=1}^{M,T} T_H(i,j) \\ TV = \sum_{v=1}^{columns}\sum_{i,j=1}^{N,T} T_V(i,j) \end{cases} \tag{8}$$

In this formula, rows is the number of horizontal slices and columns is the number of vertical slices.TH is the complete horizontal trajectory and TV is the complete vertical trajectory.

3.4 Parameter Calculation and Pedestrian Tracking

The complete trajectory can reflect the scene changes and target motion states, such as the abscissa, ordinate, instantaneous velocity, direction, width and height information of the target as above mentioned. First state, the trajectory image is a binary image. It means that if one pixel is trace of trajectory, then its value is 1, otherwise it is background and its value is 0.

For horizontal trajectory, it is analyzed it by row. In each row, count the pixels with the value of 1, and record the coordinate of first appearance position as Min_y and last as Max_y. Then w, the width of the bounding box is determined: that Max_y minus Min_y. As for vertical trajectory, it is analyzed by column. In each column, count the pixels with the value of 1, and record the abscissa of first appearance position as Min_x and last as Max_x. Then h, the height of the bounding box is determined: that Max_x minus Min_x. Meanwhile, the Min_x and Min_y can be used as the position information of the starting point along with this bounding box.

$$\begin{cases} w = Max_y - Min_y \\ h = Max_x - Min_x \end{cases} \tag{9}$$

With the information about the starting point and bounding box, rectangles can be drawn to track pedestrian in the corresponding frames of one video sequence. Take one frame at time t as example, w is the width of bounding box and h is the height of bounding box. Fig. 4 illustrates this tracking process.

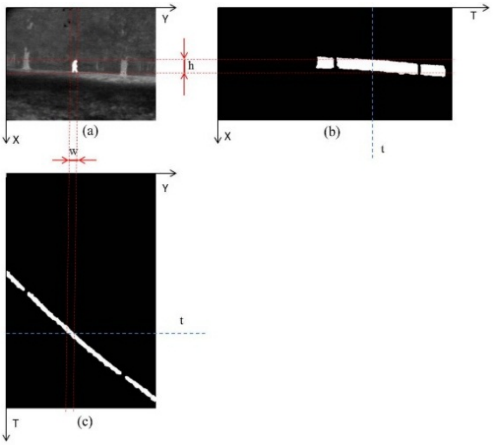


Fig. 4. (a): the t frame; (b): the horizontal trajectory; (c): the vertical trajectory

4 Experimental Results

In order to verify this proposed tracking algorithm based on the spatio-temporal slice trajectory analysis, this section we do simulation on multiple image sequences, including irw07, irw08, irw09, irw10 and irw11(these sequences are all from the OTCBVS/05 Terravic Motion IR Database [13]).

4.1 Complexity Performance

Traditional detection or tracking approaches usually consider all pixels in the image sequence, but in this approach target was detected by the low dimensional spatio-temporal slice images. Therefore, computational cost is relatively reduced in our approach. Their computational costs are shown respectively in formula (10) and formula (11).

$$O\left(M \times N \times T\right) \tag{10}$$

$$O\left[\left(k_1 \times N + M \times k_2\right) \times T\right] \tag{11}$$

where M, N, T represent the image height, width and total image frames, and k1,k2 represent the number of horizontal and vertical slice, meanwhile, $k_1 \ll M, k_2 \ll N$.

To validate time efficiency of our approach, experiments were carried on multiple surveillance videos of OTCBVS/05 Terravic Motion IR Database. The algorithm has been implemented by Matlab R2009a on a PC with Intel Core Quad CPU at 2.50 GHz running the Windows 7 Operating System. The size of each infrared image in this database is 320×240 pixels. Table 1 is the brief description about these image sequences. The average operating time of per frame is shown in Table 2.

Table 1. Sequence Information

Sequence Name	Description		
	Motion direction	Distance	Occlusion
irw07	away from the camera	close	no
irw08	across the scene	far	no
irw09	across the scene	far	no
irw10	across the scene	far	yes
irw11	across the scene	close	yes

4.2 Comparison of Tracking Performance

Just like the horizontal slices, the vertical slices also include motion information of the target. Especially the height and abscissa of the target can be obtained from the vertical trajectory. So it can improve tracking accuracy by analyzing both horizontal and vertical trajectories. The experimental results by only using horizontal slice and using both horizontal and vertical slices are shown in Fig. 5.

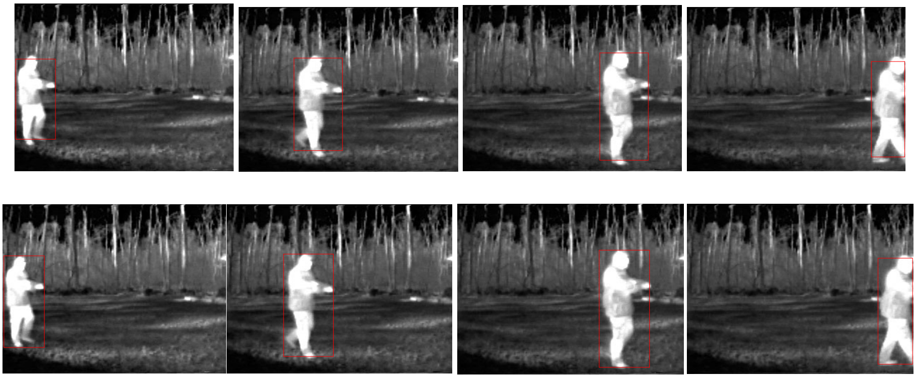


Fig. 5. Tracking results of irw08 (1) Row1: only use horizontal slices (2) Row2: the proposed method

From the comparison of these tracking results, we can see it can indeed track pedestrian with higher accuracy by using horizontal and vertical slices together.

4.3 Tracking Results

Sequences with different scenes and motion modes will have different trajectory patterns. The following images are different horizontal and vertical trajectories.

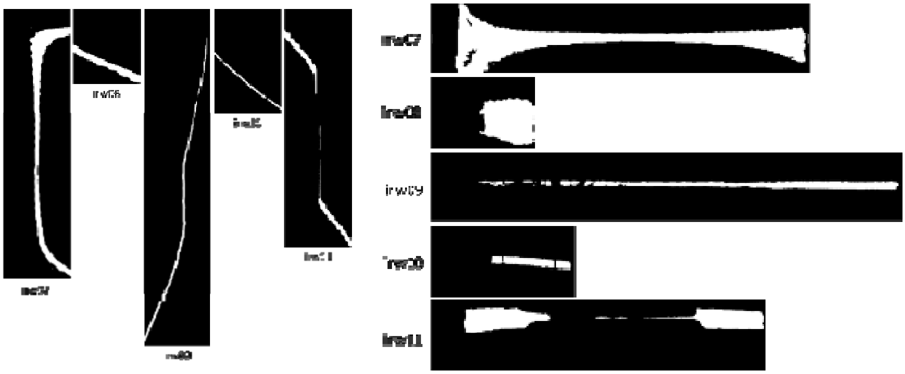


Fig. 6. Trajectories of all sequences (a) left: horizontal (b) right: vertical

The figures below are part of the tracking results of irw07, irw09, irw10 and irw11.

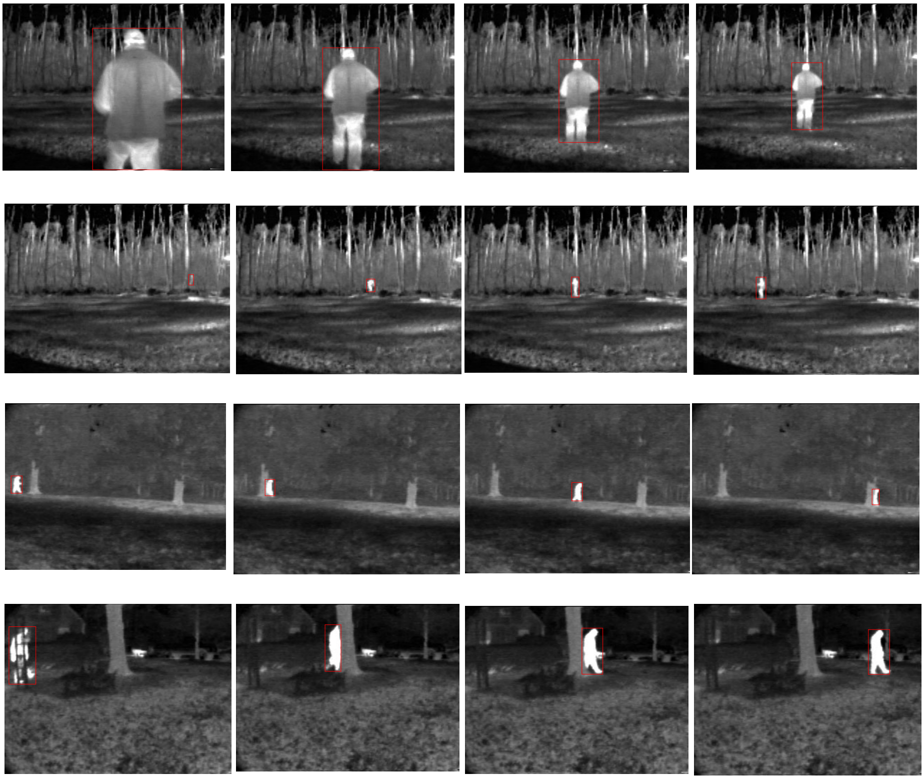


Fig. 7. Part of the tracking results from OTCBVS (1) Row1: irw07 (2) Row 2: irw09 (3)Row3:irw10 (4) Row4:irw11

If the foreground (pedestrian) is regarded as positive and background is negative, then we can use the following evaluation index to validate the efficiency of the proposed algorithm.

Table 2. Tracking results for the thermal video sequences

Name	Total Frames	TP	FP	FN	TN	TR	ACC	Comp.Time (s/frame)
irw07	1300	1206	0	0	94	1	1	0.05117
irw08	360	184	0	0	176	1	1	0.05071
irw09	1620	1354	0	48	218	0.9658	0.9704	0.05358
irw10	500	263	0	4	233	0.9850	0.9920	0.04972
irw11	1150	818	0	27	305	0.9680	0.9765	0.04898
Avg.	-	-	-	-	-	0.9838	0.9878	0.05083

(TP: True Positive; FP: False Positive; FN: False Negative; TN: True Negative; TR: Tracking Rate, TR = TP/ (TP + FN); ACC: ACC= (TP+TN)/ (P+N))

It can be observed from Table 2 that the proposed method has average 98.38% tracking rate at a relatively fast computation speed and zero false alarm rate. For sequence like irw09, irw10 and irw11, all the one pedestrian in motion are under the condition of target partially or completely occlusion for a period, so it was a little difficult to track the indetectable target and some false negative result (which means the target was regarded as background) occurred.

5 Conclusions

A novel tracking algorithm based on spatio-temporal slice trajectory analysis for single pedestrian tracking is proposed. Trajectories both in horizontal and vertical direction are used for determining the information of target boundary and position. Experiments show that the proposed tracking algorithm could perform well in different infrared image sequence, especially that it tolerates the situation that subregion of the pedestrian was with occlusion. In conclusion, it can provide relatively accurate pedestrian bounding boxes with low computational cost, and it has robustness to some extent.

References

1. Leichter, I., Lindenbaum, M., Rivlin, E.: Mean shift tracking with multiple reference color histograms. Computer Vision and Image Understanding **114**(3), 400–408 (2010)
2. Hsia, K.H., Lien, S.F., Su, J.P.: Moving target tracking based on CamShift approach and Kalman filter. Int. J. Appl. Math. Inf. Sci. **7**(1), 193–200 (2013)
3. Corvee, E., Bremond, F.: Body parts detection for people tracking using trees of histogram of oriented gradient descriptors. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 469–475. IEEE (2010)
4. Bai, Y., Tang, M.: Robust tracking via weakly supervised ranking svm. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1854–1861. IEEE (2012)
5. Ning, J., Zhang, L., Zhang, D., et al.: Robust object tracking using joint color-texture histogram. International Journal of Pattern Recognition and Artificial Intelligence **23**(07), 1245–1263 (2009)
6. Wang, X., Tang, Z.: Modified particle filter-based infrared pedestrian tracking. Infrared Physics & Technology **53**(4), 280–287 (2010)
7. Ngo, C.W., Pong, T.C., Zhang, H.J.: Motion analysis and segmentation through spatio-temporal slices processing. IEEE Transactions on Image Processing **12**(3), 341–355 (2003)
8. Sato, K., Aggarwal, J.K.: Temporal spatio-velocity transform and its application to tracking and interaction. Computer Vision and Image Understanding **96**(2), 100–128 (2004)
9. Yang, J., Su, X., Ma, P.: Fast pedestrian detection using slice-based motion analysis. In: 2010 First International Conference on Pervasive Computing Signal Processing and Applications (PCSPA), pp. 74–77. IEEE (2010)
10. Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. JOSA A **2**(2), 284–299 (1985)

11. Zhang, R., Ge, P., Zhou, X., et al.: An method for vehicle-flow detection and tracking in real-time based on Gaussian mixture distribution. Advances in Mechanical Engineering (2013)
12. Guo, J.M., Hsia, C.H., Liu, Y.F., et al.: Fast background subtraction based on a multilayer codebook model for moving object detection. IEEE Transactions on Circuits and Systems for Video Technology **23**(10), 1809–1821 (2013)
13. IEEE OTCBVS WS Series Bench; Roland Miezianko, Terravic Research Infrared Database. http://www.vcipl.okstate.edu/otcbvs/bench/Data/05/download.html

Exploring Deep Gradient Information for Face Recognition

Jianjun Qian(✉), Jian Yang, and Ying Tai

School of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China
{csjqian,csjyang}@njust.edu.cn, tyshiwo@gmail.com

Abstract. This paper presents a novel and simple image feature extraction method, called exploring deep gradient information (DGI), for face recognition. DGI first captures the local structure of an image by computing the histogram of gradient orientation of each macro-pixel (local patch around the central pixel). One image can be decomposed into L sub-images (also called orientation images) according to the structure information of each macro-pixel since there are L bins in the local histogram. For each orientation image, dense scale invariant feature transform (DSIFT) is used to further explore the gradient orientation information. All DSIFT feature are concatenated into one augmented super-vector. Finally, dimensionality reduction technology is applied to obtain the low-dimensional and discriminative feature vector. We evaluated the proposed method on the real-world face image datasets NUST-RWFR, Pubfig and LFW. In all experiments, DGI achieves competitive results compared with state-of-the-art algorithms.

Keywords: Feature extraction · Gradient information · Local structure feature · Face recognition

1 Introduction

Face recognition is a very popular visual recognition problem that attracted a lot of attentions due to its challenging nature and its potential values for practical applications. Especially, it's difficult to hand the appearance variations that commonly occur in face images owing to pose, illumination and facial expression. Therefore, a simple and robust image feature extraction method is one of the most urgent needs for an automatic face recognition system. In the past decade, there is a large amount of literatures on developing image feature extraction methods.

The previous work like SIFT [6] shows that gradient information plays an important role in image feature representation since it achieves good performance in the cases of image matching and scene recognition. To further explore sufficient gradient information behind the observed space, this paper borrows the idea of IDLS [11] and presents a novel image feature extraction method named exploring deep gradient information (DGI) for face recognition. DGI first computes the gradient orientation of each pixel. Thus, one image is decomposed into a series of orientation images

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 159–169, 2015.
DOI: 10.1007/978-3-662-48570-5_16

according to local gradient histogram of each macro-pixel. To achieve the more detailed gradient information, dense scale invariant feature transform (DSIFT) [14] is employed to represent each orientation images. Subsequently, we concatenated all DSIFT features into one augmented super-vector and used the dimensionality reduction technology to obtain the low-dimensional and discriminative feature. The pipeline of the proposed method is shown in Fig. 1.

It's necessary to highlight the differences between DGI and IDLS. 1) IDLS describe the local structure by computing the relationship between central macro-pixel and its neighbors in the local window. DGI utilize the histogram of gradient orientation in the macro-pixel to represent the local structure. 2) IDLS concatenated all the normalized structure images into a super-vector. However, DGI further explore the gradient information using DSIFT for each orientation images. Thus, DGI concatenated all DSIFT features into an augmented vector.

2 Related Work

It's known that image feature extraction method can be classified into two categories: subspace based global feature and local descriptor feature. In this section, we mainly focus on the related literature of local image feature extraction methods.

One of the most successful face image extraction methods is Gabor feature [1]. Gabor feature explores the desirable local characteristic structure of spatial frequency, spatial locality, and selective orientation, which are proven to be robust to illumination and facial expression changes. Local Binary Pattern (LBP) [2, 3] describes the difference between central pixel and its neighbors over the local patch. A lot of methods have been proposed to cover the shortcomings of LBP, like three-patch LBP, four-patch LBP [4] and local tensor pattern (LTP) [5]. SIFT, as one of the most popular local image descriptor, is widely used in the area of computer vision. For face recognition, SIFT performed not good since it is a sparse descriptor, which cannot provide rich detailed information. However, dense SIFT [14] overcome this drawback and gave better results than LBP [12].

Recently, Ngoc-Son et al. [7] applied the LBP-based structure on oriented edge magnitudes to construct a novel image descriptor, Patterns of oriented edge magnitudes (POEM), for face recognition. J. Qian et al. gave a novel scheme to compute dominant gradient orientation and presented a discriminative histogram of local dominant orientation (D-HLDO) for image feature extraction [8]. D-HLDO is robust to image noise and contrast changes. H. J. Seo et al. proposed LARK descriptor for face verification [9]. The key idea is that LARK describes the local structure information by using the geodesic distance to measure the similarities between the central pixel and its neighborhoods. Subsequently, ID-LARK is proposed for face recognition under difficult lighting conditions [10]. Based on the idea of image decomposition, J. Qian et al. presented a novel image feature extraction method coined IDLS and applied it for face recognition [11]. IDLS gave the competed results in various face image sets. Especially, IDLS is non-sensitive to illumination changes.

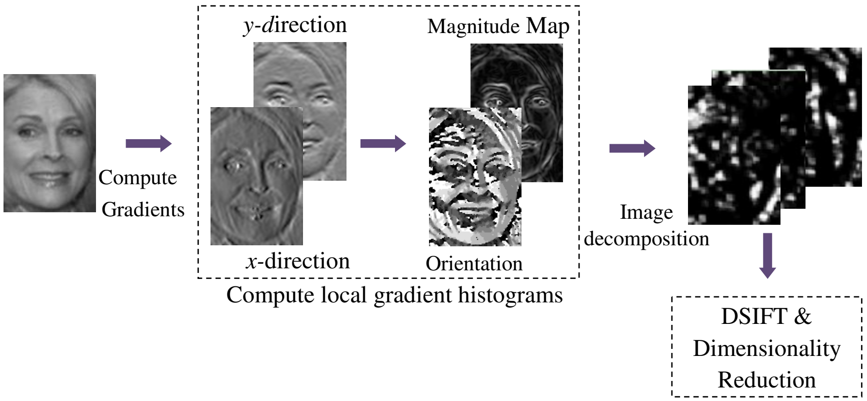


Fig. 1. An overview of the proposed method DGI

3 Exploring Deep Gradient Information

In this section, we will discuss how to explore deep gradient information in detail. There are mainly three steps: 1) orientation images are obtained by decomposing image via local histogram of gradient orientation; 2) re-exploring the gradient information on orientation images using DSIFT; 3) dimensional-reduction technology is employed to obtain the low-dimensional and discriminative feature vector.

3.1 Image Decomposition Based on Local Histogram of Gradient Orientation

To describe the local histogram of gradient orientation, let us characterize the macro-pixel firstly. Suppose there are N pixels in an image. We treat the *i*-th pixel in an image as a center and determine its corresponding macro-pixel. The macro-pixel is actually a $r \times r$ (e.g. *r*=5, 7, 9) local patch around the *i*-th pixel.

Further, we describe our method to obtain the local histogram of gradient orientation. Specially, we compute the gradient information in *x*-direction and *y*-direction, respectively. The orientation of the *i*-th pixel θ_i is defined as follows:

$$\theta_i = \begin{cases} \arctan(\frac{Gy(i)}{Gx(i)}), & Gx(i) > 0 \,\&\, Gy(i) > 0 \\ \arctan(\frac{Gy(i)}{Gx(i)}) + \pi, & Gx(i) < 0 \,\&\, Gy(i) > 0 \\ \arctan(\frac{Gy(i)}{Gx(i)}) + \pi, & Gx(i) < 0 \,\&\, Gy(i) < 0 \\ \arctan(\frac{Gy(i)}{Gx(i)}) + 2\pi, & Gx(i) > 0 \,\&\, Gy(i) < 0 \end{cases} \tag{1}$$

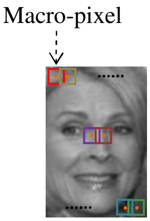


Fig. 2. The distribution of macro-pixels

where, $Gx(i)$ is the gradient value in *x*-direction of the *i*-th pixel, $Gy(i)$ is the gradient value in *y*-direction of the *i*-th pixel. The corresponding magnitude value w_i of θ_i is defined as:

$$w_i = \sqrt{Gx(i)^2 + Gy(i)^2} \tag{2}$$

As well known, the distribution of local gradients or edge directions can describe the local information effectively. To achieve the sufficient gradient information, we choose a series of macro-pixels with half-overlapping (as shown in Fig. 2) and compute the histogram of gradient orientation of them. For the *j*-th macro-pixel m_j, we construct a 1-D histogram of gradient orientation, $\mathbf{h}_j = [h_1, \cdots h_t \cdots h_L]$, which contains *L* bins (e.g. *L*=9), with each bin covering 360/*L* degrees for "signed" gradient. Each gradient orientation in macro-pixel m_j added to a histogram bin is weighted by its corresponding magnitude. In other words, the height of *t*-th bin of the histogram, h_t, accumulates all magnitudes that are associated with the orientation angles belonging to the *t*-th bin of the histogram. All the local histograms of each macro-pixel can form a gradient feature matrix as follows:

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_P] = \begin{pmatrix} h_{1,1} & h_{2,1} & \cdots & h_{P,1} \\ \vdots & \vdots & \iddots & \vdots \\ h_{1,L} & h_{2,L} & \cdots & h_{P,L} \end{pmatrix} = \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_L \end{bmatrix} \tag{3}$$

where, *P* is the number of macro-pixel.

Each row of **H**, a *P* dimensional vector can be reformulated into an image, which called orientation image. The gradient feature matrix thus generated a set of *P* orientation images. That is, one image can be decomposed into *P* orientation images according to local histogram of gradient orientation. Fig. 3 shows the orientation images are derived from a given image. Since each orientation image accumulate the different bin of local histogram in all macro-pixels, they reveal the structure information of different orientations.

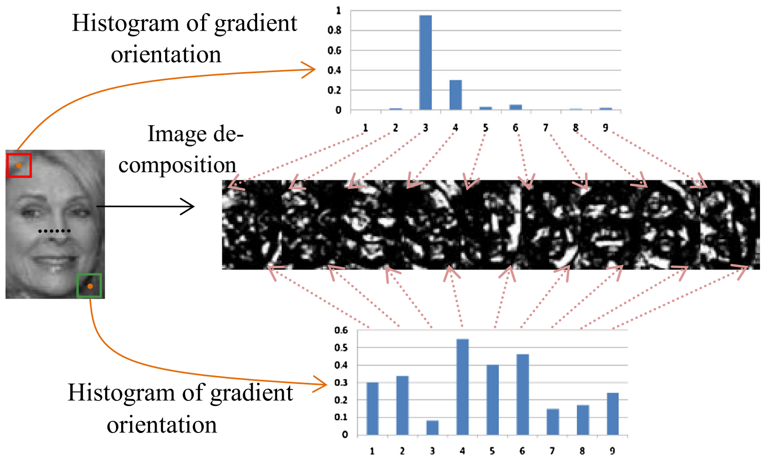


Fig. 3. Image decomposition according to local histogram of gradient orientation

3.2 Feature Representation and Dimensionality Reduction

From Section 3.1, we can obtain several orientation images of an image to describe the structure information in different orientations. To achieve the robust feature representation, we will further explore the detailed gradient information for each orientation image.

We know that SIFT is actually a 3-D histogram of gradient locations and orientations. The contribution of the location and orientation bins is determined by the gradient magnitude. We believe that SIFT reveal the gradient information in the local patch around the center-pixel. Here, we mainly focus on the descriptor representation of SIFT and select DSIFT to further extract gradient feature for all orientation images.

All the DSIFT features of each orientation images are concatenated into an augmented super-vector to include all deep gradient information from different orientations as shown in Fig. 4. Letting $\mathbf{v}_t$ $(t=1,\cdots,L)$ represents feature vector of the *t*-th orientation image. The augment super-vector $\mathbf{V}$ is defined as follows:

$$\mathbf{V}=(\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_L) \tag{4}$$

Now, the dimension of feature vector $\mathbf{V}$ is very high. To address this problem, FLDA is used to achieve a compact feature space to improve the discriminative power and the computational efficiency. The discriminative and low-dimensional deep gradient feature vector is:

$$\mathbf{DGI}=\boldsymbol{W}^T\mathbf{V} \tag{5}$$

where, $\boldsymbol{W}$ is the projection matrix of FLDA.

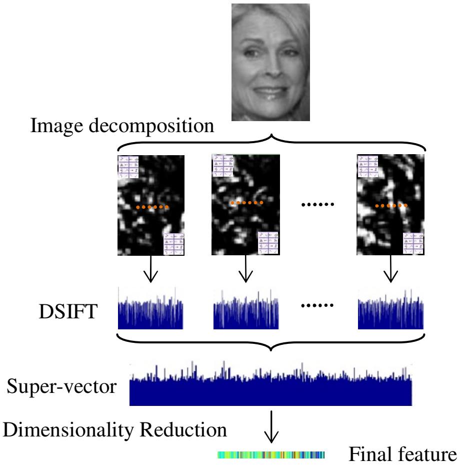


Fig. 4. An illustration of a DGI feature for a given image

4 Experiments

In this section, we will use the proposed method DGI for image feature extraction and compared the performance with those of state-of-the-art algorithms by experimenting on three real-world face databases (NUST_RWFR, Pubfig and LFW). For DGI, the number of bin is set to 9 through the paper.

4.1 NUST_RWFR

The NUST-RWFR database [11] contains 2400 color face images of 100 persons, including frontal views of faces with different facial expressions, lighting conditions and degrees of blurring. All the pictures are taken in a real world situation. The pictures of 100 persons were taken in two sessions and each session contains 12 color images. The quality of pictures in the first session is good, and that in the second session is poor. All face images of these 100 persons are used in our experiments. The face portion of each image is manually cropped and then normalized to 80×80 pixels. The sample images of one person as shown in Fig. 5.

In the first experiment, the images from the first session of each person are used for training, and images from the second session for testing. DSIFT plus FLDA, LBP plus Chi2, LTP plus Chi2, LARK, POEM, IDLS and proposed DGI are used for image feature extraction. The nearest neighbor (NN) classifier is employed for classification. Table 1 lists the recognition rates of each method. From Table 1 (Experiment 1), we can see clearly that the proposed DGI achieves remarkable results among all the methods. DGI significantly outperforms LBP, LARK and POEM. The probable reason is that these three methods are lack of discriminative power. Actually, DSIFT plus FLDA can be considered as a superficial gradient orientation feature. The performance of DGI is 8% higher than that of DSIFT plus FLDA. Compared to IDLS, DGI also decompose an image into several sub-images according to local structure information. However, DGI explore the deep gradient information using DSIFT to further represent the orientation images. This is reason why DGI give better performance than IDLS. Additionally, Fig. 6 illustrates that DSIFT plus FLDA, IDLS and DGI give the similar results when the dimension is lower. However, DGI gives better performance than the other two methods when the dimension is equal or greater than 50.

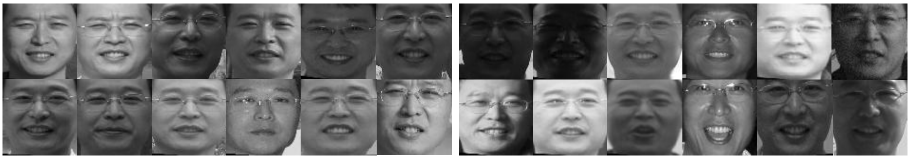

(a) (b)

Fig. 5. Sample images for one person of NUST-RWFR database (a) session 1, (b) session 2

Table 1. The recognition rates of each method on the NUST_RWFR database

Methods	**Recognition Rates**	
	Experiment 1	**Experiment 2**
DSIFT + FLDA	67.1	73.5
LBP [3]	49.2	60.7
LTP [5]	45.4	62.0
LARK [11]	47.2	62.7
POEM	56.4	65.3
IDLS [11]	73.8	78.0
DGI	75.1	79.3

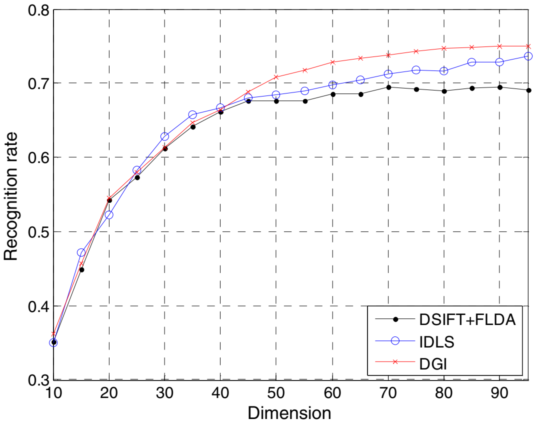


Fig. 6. The recognition rates of DSIFT, IDLS and DGI with the variations of dimensions on the NUST_RWFR database

The second experiment only selected images from the session 1 (with good quality) of each person. We just use the first six images of each person to construct the training set, and the remaining images are used for testing. Similar with the first experiment, seven competed methods are used for image feature extraction, respectively. The performance of each method is listed in Table 1. From Table 1 (Experiment 2), we see that the performance of all the methods are better than that in the column of Experiment 1 since the second experiment setting is easier than the first experiment. DGI always gives better recognition rates than other methods.

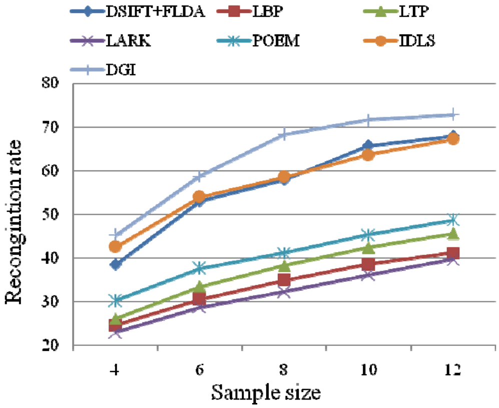


Fig. 7. The recognition rates of each method versus the variation of the training sample size on the Pubfig database

4.2 Pubfig

Pubfig is a large real-world face dataset consists of 58,797 images of 200 persons [15]. All images are taken in completely uncontrolled situations with non-cooperative subjects and there is large variation in pose, lighting, expression, scene, camera, imaging conditions and parameters, etc. Twenty face images of two hundred persons are selected and used in this experiment. We simply crop the aligned face images [17] to remove the background and resize them into 80×70 pixel.

In this experiment, we choose the first K (K varies from 4 to 12 with interval 2) images per subject for training, and the rest images for testing. The recognition rates of DSIFT plus FLDA, LBP plus Chi2, LTP plus Chi2, LARK, POEM, IDLS and proposed DGI with NN classifier are shown in Fig. 7. DGI always shows better results than the other methods, irrespective of the variations of training sample size. Fig. 7 also gives similar results with Table 1.

4.3 LFW

The LFW [16] database contains 13,233 target face images. There are 5,749 different individuals in the LFW. 1,680 people have two or more face images. The remainder 4,069 persons have just only one image. These images have a large degree of facial expression, occlusions, pose and illuminations since all of them are taken from the real-world. Here, we use the aligned version of images [18] and simply crop the face image to remove the background, leaving a 150×100 face image.

In this experiment, we mainly focus on the unsupervised setting, which is the most difficult case since there are no training examples available. Thus, DGI is applied to represent image feature for face verification and compared with the state-of-the-art algorithms include SIFT, LBP, LTP, LARK, POEM, LQP [13] and IDLS. Notice that DGI does not use the dimensionality reduction technology here. Additionally, we use the Euclidian distance to compute the similarity score. To reduce the effect of pose variation, LARK presents a 'mirror' operator to improve the performance. Similar with LARK, POEM gives 'flip' operator. Actually, both the idea of 'mirror' and 'flip' are the same. Here, we also give the mirror version of the proposed DGI to improve the performance. Table 2 lists the mean accuracy of each method follows the protocol as specified in [16]. It's clear that the mirror version of the proposed DGI outperforms all other competed image feature extraction methods. DGI also gives the better results than all considered descriptors like SIFT, LBP, LTP and also the very recent LARK, POEM and IDLS. Especially, the performance of DGI is 5.4% higher than SIFT. This result demonstrates that image decomposition based on local gradient orientation further improve the accuracy of face verification.

Table 2. The mean verification rates of the competed methods on the LFW database View 2 (the unsupervised setting)

Methods	Verification Rates
SIFT [12]	69.12
LBP [12]	68.24
LTP [11]	69.95
LARK [9]	70.98
IDLS [11]	73.70
POEM [7]	73.69
LQP [13]	75.30
DGI	74.52
LARK (Mirror) [9]	72.23
POEM (flip) [7]	75.22
DGI (Mirror)	75.65

5 Conclusions

This paper has presented a deep feature representation method coined DGI for face recognition. We have evaluated the new image feature extraction method through the extensive experiments on three real-world face image databases. Experimental results demonstrate that the proposed DGI achieves better or comparable results in comparison with state-of-the-art methods. Overall, our main findings are as follow. 1) Deep architecture provides valuable information behind the observed space. 2) Orientation images reveal the intrinsic structure of an image from different orientations. 3) Carefully adapted expressive gradient features for each orientation image are pivotal to the good performance of the practical system.

In our future work, we are planning to investigate diversified gradient features for orientation images and try to incorporate multi-feature image representation methods into one framework.

References

1. Liu, C.J., Wechsler, H.: Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing **11**(4), 467–476 (2002)
2. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 971–987 (2002)
3. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006)
4. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: Faces in Real-Life Images Workshop in European Conference on Computer Vision (ECCV) (2008)
5. Tan, X.Y., Triggs, B.: Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. IEEE Transactions on Image Processing **19**(6), 1635–1650 (2010)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**, 91–110 (2004)
7. Vu, N.-S., Caplier, A.: Enhanced patterns of oriented edge magni-tudes for face recognition and image matching. IEEE Transactions on Image Processing **21**(3), 1352–1365 (2012)
8. Qian, J., Yang, J., Gao, G.: Discriminative histograms of local dominant orientation (D-HLDO) for biometric image feature extraction. Pattern Recognition **46**(10), 2724–2739 (2013)
9. Seo, H.J., Milanfar, P.: Face Verification Using the LARK Representation. IEEE Transactions on Information Forensics and Security **6**, 1275–1286 (2011)
10. Qian, J., Yang, J.: A novel feature extraction method for face recognition under different lighting conditions. In: Sun, Z., Lai, J., Chen, X., Tan, T. (eds.) CCBR 2011. LNCS, vol. 7098, pp. 17–24. Springer, Heidelberg (2011)
11. Qian, J., Yang, J., Xu, Y.: Local Structure-based Image Decomposition for Feature Extraction with Applications to Face Recognition. IEEE Trans on Image Processing **22**(9) (September 2013)

12. Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: Asian Conference on Computer Vision (2009)
13. ul Hussain, S., Napoleon, T., Jurie, F.: Face recognition using local quantized patterns. In: British Machine Vision Conference (2012)
14. Vedaldi, A., Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008)
15. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: International Conference on Computer Vision (ICCV) (2009)
16. Huang, G., 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)
17. http://vfr.enriquegortiz.com
18. http://www.openu.ac.il/home/hassner/data/lfwa/

A Novel Dynamic Character Grouping Approach Based on the Consistency Constraints

Pengfei Xu[1(✉)], Qiguang Miao[3], Ruyi Liu[3], Feng Chen[2], Xiaojiang Chen[1], and Weike Nie[1]

[1] School of Information Science and Technology, Northwest University, Xi'an 710127, China
pfxu@nwu.edu.cn
[2] Office of Science and Technology, Northwest University, Xi'an 710127, China
[3] School of Computer, Xidian University, Xi'an 710071, Shaanxi, China

Abstract. In optical character recognition, text strings are extracted from images so that it can be edited, formatted, indexed, searched, or translated. Characters should be grouped into text strings before recognition, but the existing methods cannot group characters accurately. This paper proposes a new approach to group characters into text strings based on the consistency constraints. According to the features of the characters in the topographic maps, three kinds of consistency constraints are proposed, which are the color, size and direction consistency constraint respectively. In the proposed method, due to the introduction of the color consistency constraint, the characters with different colors can be grouped well; and this method can deal with the curved character strings more accurately by the improved direction consistency constraint. The final experimental results show that this method can group the characters more accurately, and lay a good foundation for text recognition.

Keywords: Grouping characters · Topographic maps · Color information · Character expandability · Consistency constraint

1 Introduction

Recognizing individual characters separately fails to take advantage of the whole word context, and the recognition results cannot represent the meaning of the word[1-2]. Character grouping is a difficult task, and much of the previous methods can only work on specific cases[3]. In order to solve these problems, researchers proposed character grouping methods to group characters into text strings, then these strings can be recognized to represent the meaning of the words more accurately.

In 1999, Goto proposed a method called Extended Linear Segment Linking, which was able to extract text strings in arbitrary orientations and curved lines[4]. This method works on touching characters effectively, and requires that the sizes of the characters are similar. A bottom-up approach was proposed by Pal[5], but it cannot work on the curved text strings. In 2008, Roy proposed a method based on the foreground and background information of the characters to extract individual text strings from multi-oriented and curved text document[6]. In 2009, another method was presented by him to

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 170–179, 2015.
DOI: 10.1007/978-3-662-48570-5_17

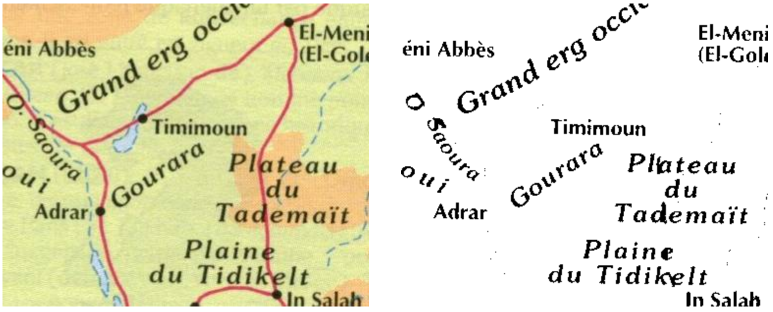
segment English multi-oriented touching strings into individual characters by using convex hull information[7]. These methods can deal with curved strings, but the directions of the strings were detected only in 4 directions. In 2004, a method for separating and recognizing the touching/overlapping characters was proposed by Velázquez[8]. In this method, OCR was applied to define the coordinate, size and orientation of the character strings, and four straight lines or curves were extrapolated to separate those symbols that were attached. In 2011, Aria Pezeshk grouped the individual characters into their respective strings using pyramid decomposition with Gaussian kernels[9,10], but this method cannot distinguish different text strings when they are nearby.

According to the analysis above, it is known that most researchers focused on the study of text separation and recognition, but the methods for grouping text strings are not researched deeply. Chiang has done lots of work on grouping characters[11] and text recognition[12], and a conditional dilation algorithm was presented for grouping characters into text strings[11]. Compared with other methods, Chiang's method can get better results. But there are still some problems, for example, the color information of characters is not considered, and the string curvature condition is not perfect. In order to solve these problems, this paper proposes a method to group characters into text strings based on the consistency constraints of the character color, size and directions.

The organization of the paper is as follows: In section 2, we make an analysis of the features of characters in maps. And the proposed method is described in section 3. In section 4, we compare the experimental results with Chiang's method. Finally, the concluding remarks are given in section 5.

2 The Analysis of Characters in Topographic Maps

In topographic maps, such as the map shown in Fig.1, the distribution of characters is very complex. The sizes of characters in different text strings are not the same, and the distance between some text strings is very small. In addition, there are broken and touching characters in the segmented maps. All these facts adversely affect the accuracy of the grouped text string.



(a) The topographic map (b) The text strings

Fig. 1. Text strings in topographic maps

At present, there are few methods for grouping characters into text strings, and these methods can only be used when the characters could be separated accurately. But some characters are mistakenly grouped into other text strings or leaved out when the characters in the map image have different sizes, directions, and colors, especially when some text strings are in a curved line.

3 Dynamic Character Grouping Based on the Consistency Constraints

According to the analysis in section 2, it is difficult to group the multi-oriented, multi-sized, and curved characters into text strings. This section gives a new method for grouping characters into text strings. The color information, which is not considered by the existing method[11], is used as an additional constraint due to that some characters are presented by different colors. The directional constraint is designed more perfect in the new method. So based on the consistency constraints of the character color, size and direction, the new character grouping method can be described by the following expression.

$$\mathrm{T} = \mathrm{G}(c,\ s,\ d) \tag{1}$$

Where, T is the grouped text string, $\mathrm{G}(\bullet)$ is the character grouping operation function. c , s and d are the color, size and direction consistency constraints, which means that the characters in one text strings have the similar color and size, and the centers of these characters are on a curved line，whose curvature is in a numerical range.

3.1 The Color Consistency Constraint

In topographic maps, some text strings are represented by different colors, so we can use this information to distinguish different text strings. And for a character α , its color feature is defined as:

$$C_\alpha = M\left(\alpha_{\mathrm{R}},\ \alpha_{\mathrm{G}},\ \alpha_{\mathrm{B}}\right) \tag{2}$$

Where C_α is the main color feature of the character α , which is obtained by color histogram. $M(\bullet)$ is the operation of color extraction, and the average value of RGB in the character area is used as the main color feature of this character. The color difference between the characters α_1 and α_2 can be obtained by Mahalanobis distance.

$$\mathrm{D}_{\alpha_1,\alpha_2} = \sqrt{\left(\mathrm{C}_{\alpha_1} - \mathrm{C}_{\alpha_2}\right)^{\mathrm{T}} S^{-1}(\mathrm{C}_{\alpha_1} - \mathrm{C}_{\alpha_2})} \tag{3}$$

Where S^{-1} is the inverse of the covariance matrix of the samples. The color consistency constraint of the characters means that $\mathrm{D}_{\alpha_1,\alpha_2}$ should be less than or equal to a minimum threshold T_c.

$$\mathrm{D}_{\alpha_1,\alpha_2} \le T_c \tag{4}$$

Where T_c is the average color difference in the area aound the current characters α_1 and α_2.

$$T_c = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D_{p_i,p_j}}{(\frac{N(N-1)}{2})} \tag{5}$$

Where N is the number of the pixels in the area, D_{p_i,p_j} is the color distance between the pixels p_i and p_j

If there are no color information, all the characters are viewed as that all of them have the same color. Further, they are grouped by the size consistency constraint and the direction consistency constraint.

3.2 The Size Consistency Constraint

For a character, its size is the max value of the length and the width of its bounding box, and the sizes of the characters in one text strings must be similar. So the size ratio of the characters must be smaller than a threshold T_s. According to the size features of the characters, such as the English letter '*f*' and 'a', we always choose T_s=3 [11].

3.3 The Direction Consistency Constraint

For the text strings in the binary image, each connected component is a single character, and each character needs to be connected to at least one character. Assuming that there is a character α_1, we can get another character α_2 which is the closest character to α_1 according to the size consistency constraint and the color consistency constraint. Furthermore, based on the distances between different connected components, we try to get the neighbor character $\alpha_{1,\mathrm{N}}$ of α_1, and $\alpha_{2,\mathrm{N}}$ of α_2. In this way, at least two and no more than four characters, which may belong to the same text string, can be obtained.

There are three cases when we check α_1 and α_2 belong to the same text string or not.

(1). If there are no neighbor characters of α_1 and α_2, these two characters are grouped into a text string directly.

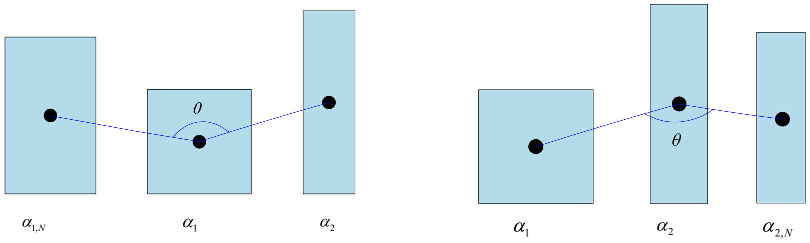
(2). If there is one neighbor character of α_1 or α_2, we need to check the direction consistency constraint of these three characters. Although the sizes of the characters of one text string meet the size consistency constraint, they are not the same, so the centers of these characters are on a curved line rather than on a straight line. The curved text grouping is similar with this case.

As shown in Fig.2, if the angle θ of the two lines is less than or equal to the threshold T_d , these three characters are satisfied with the direction consistency constraint.

$$|\theta - 180^\circ| \le T_d \tag{6}$$

$$T_d = 180^\circ \pm \beta \tag{7}$$

Here a curvature parameter β is used to control T_d, according to the distances and the size difference between the characters, or on the basis of the empirical data, this curvature parameter β is set to 50° generally.



(a) Neighbor character of α_1

(b) Neighbor character of α_2

Fig. 2. One neighbor character of α_1 or α_2

(3) If α_1 and α_2 have a neighbor character respectively, we need to check the direction consistency constraint of these four characters, as shown in Fig.3. If either the angle θ_1 or θ_2 is less than or equal to a minimum threshold T_d, these four characters are satisfied with the direction consistency constraint.

$$(|\theta_1 - 180^\circ| \le T_d) | (|\theta_2 - 180^\circ| \le T_d) \tag{8}$$

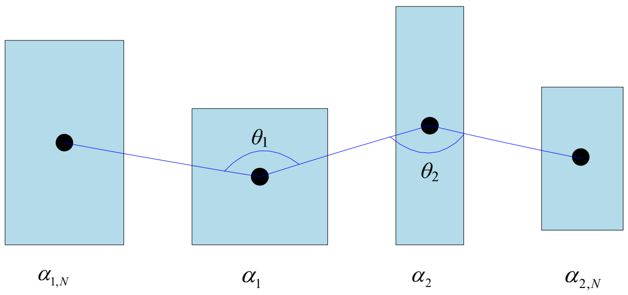


Fig. 3. Neighbor characters of α_1 and α_2

If the characters α_1, α_2 and neighbor characters are satisfied with the consistency constraint of the character color, size and direction, α_1 and α_2 are grouped into the same text string.

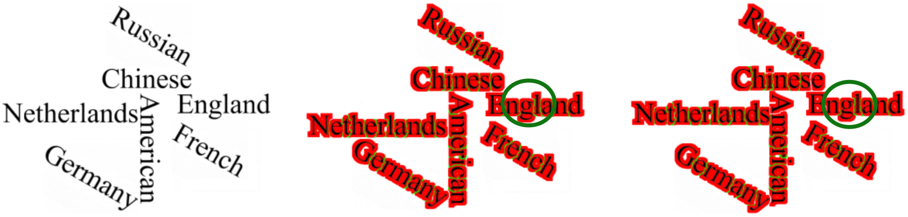
4 Experiments and Analysis

In this section, several experiments are made on artificial images and topographic maps to verify the accuracy of the proposed method. Because Chiang has made a comparison between the method proposed by him and several other methods in ref[3], and he came to the conclusion that his method has the best performances. therefore, in this paper, only Chiang's method[11] is chosen as a comparison method.

4.1 Experiments on Artificial Images

We made two artificial images which contain multi-color, multi-oriented, multi-sized, and curved strings to test our new method. In these images, the characters are seperated and unbroken. All the results are shown in Fig.4, the red pixels are the spreaded background pixles which connect to only one character, and each green pixel connects to two characters which would be grouped. By comparing the results, the proposed method can get more accurate results.

As shown in Fig.4 (c), Chiang's method can deal with the multi-oriented, multi-sized, and curved text strings well. But when there are two text strings in different directions, and the beginning or the end character of one string is close to one character of the other string. In this case, Chiang's method cannot deal with the beginning or the end characters of these two text strings, due to the disadvantages of Chiang's method in string curvature condition. In the proposed method, the direction consistency constraint is designed more perfect, so the results are better, as shown in Fig.4(b) .



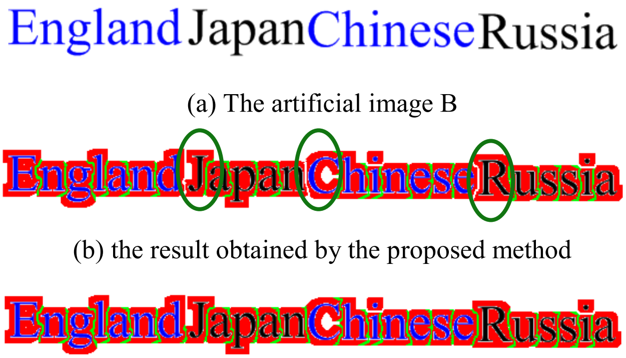
(a) The artificial image

(b) the result acquired by the proposed method

(c) the result acquired by the Chiang's method

Fig. 4. the results obtained from artificial images

In other case, Chiang's method cannot handle the close text strings with different colors, due to that color information is not used in his method, as shown in Fig.5(c). In contrast, the color consistency constraint is used in the proposed method, so the characters with different colors are not grouped into the same string, as shown in Fig.5(b).


EnglandJapanChineseRussia
EnglandJapanChineseRussia
EnglandJapanChineseRussia

(a) The artificial image B

(b) the result obtained by the proposed method

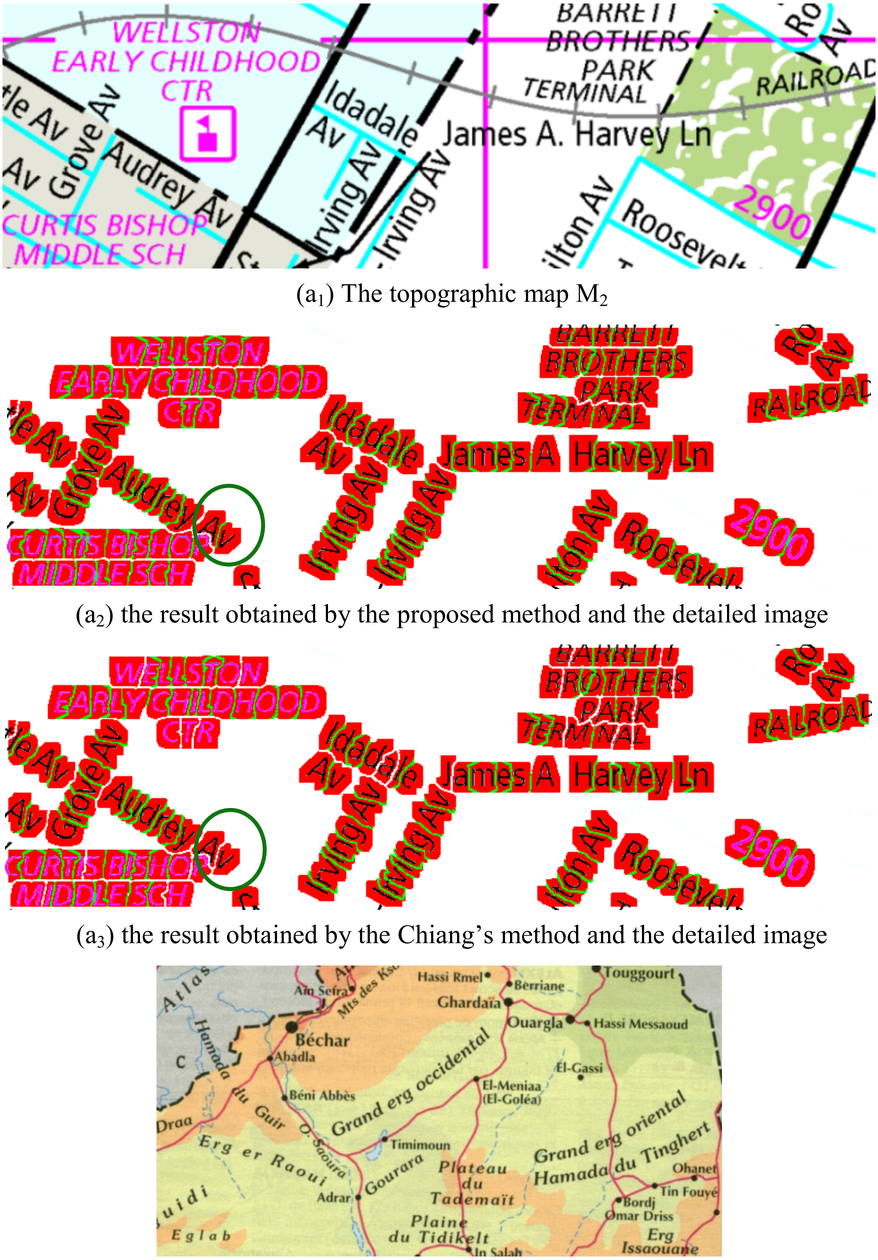
(c) the result obtained by the Chiang's method

Fig. 5. the results obtained from artificial images

4.2 Experiments on Topographic Map Images

In order to verify the accuracy of the proposed method in applications, two topographic maps are chosen as test images. Characters are separated based on color information and morphological features[2]. The grouped characters are shown in Fig.6.

From the results, the proposed method has advantages over Chiang's method. The former can deal with color and direction information better. From Fig.6(a_2) and （a_3), we can see that the proposed method can distinguish text strings with different colors, but Chiang's method cannot. Many characters, which is the beginning or end characters of the text strings, are not grouped into the corresponding text strings in the Chiang's results.



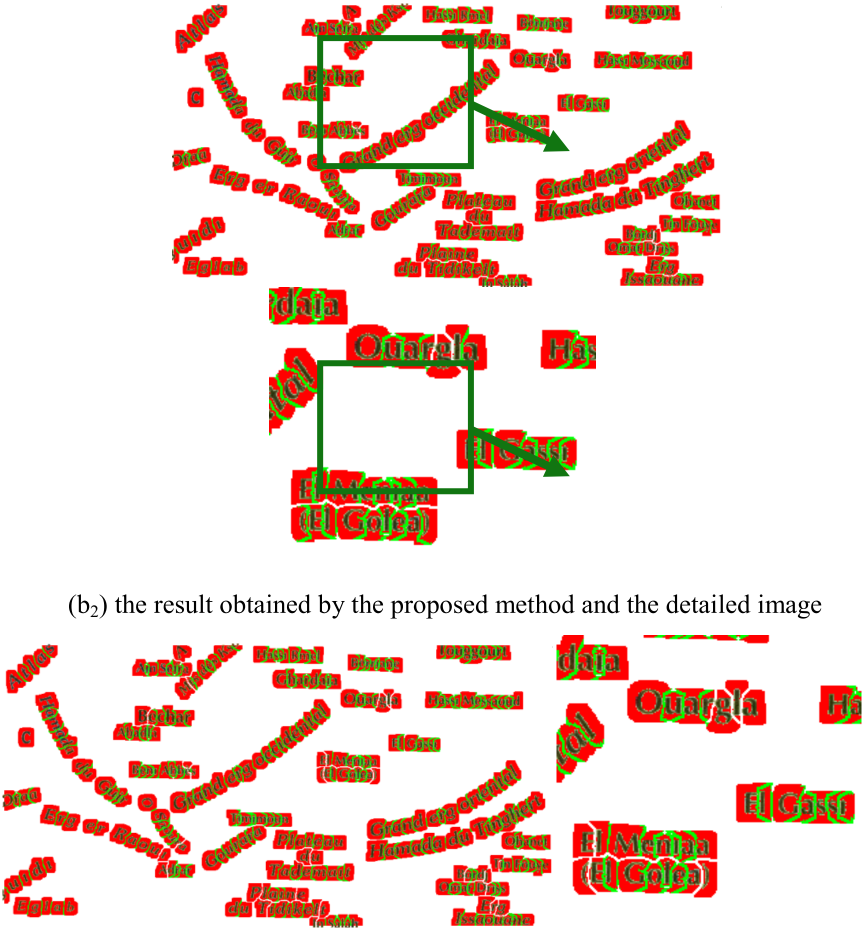
(a₁) The topographic map M₂

(a$_2$) the result obtained by the proposed method and the detailed image

(a$_3$) the result obtained by the Chiang's method and the detailed image

(b$_1$) The topographic map M$_3$

Fig. 6. The results obtained from topographic maps



(b$_2$) the result obtained by the proposed method and the detailed image

(b$_3$) the result obtained by the Chiang's method and the detailed image

Fig.6. (*Continued*)

5 Conclusion

It is well known that it is difficult to design a perfect method to group all the characters accurately in topographic maps. This paper proposes an algorithm to group characters into text strings based on the designed consistency constraint of the character color, size and direction. In this method, color features are introduced, and the direction consistency constraint is designed more perfect. Experimental results show that the proposed method can get better results. However, there are still many works to do. The proposed method cannot deal with the characters in the strings with big character spacing, so new methods should be studied.

Acknowledgments. The work was jointly supported by the National Natural Science Foundations of China under grant No. 61472302, 61272280, U1404620, 41271447, 61373177, and 61272195; The Program for New Century Excellent Talents in University under grant No. NCET-12-0919; The Fundamental Research Funds for the Central Universities under grant No. K5051203020, K50513100006, K5051303018, JB150313, and BDY081422; Natural Science Foundation of Shaanxi Province, under grant No.2014JM8310 ; The Creative Project of the Science and Technology State of xi'an under grant No. CXY1441(1); The State Key Laboratory of Geo-information Engineering under grant No.SKLGIE2014-M-4-4. The Science Foundations of Northwest University under grant No.14NW25, 14NW27, 14NW28.

References

1. Adam, S., Ogier, J.M., Cariou, C., et al.: Symbol and character recognition: application to engineering drawings. International Journal on Document Analysis and Recognition **3**(2), 89–101 (2000)
2. Pezeshk, A., Tutwiler, R.L.: Extended character defect model for recognition of text from maps. In: 2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI), pp. 85–88. IEEE (2010)
3. Chiang, Y.Y., Adviser-Knoblock, C.A.: Harvesting geographic features from heterogeneous raster maps. University of Southern California (2010)
4. Goto, H., Aso, H.: Extracting curved text lines using local linearity of the text line. International Journal on Document Analysis and Recognition **2**(2–3), 111–119 (1999)
5. Pal, U., Sinha, S., Chaudhuri, B.B.: Multi-oriented English text line identification. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 1146–1153. Springer, Heidelberg (2003)
6. Roy, P.P., Pal, U., Lladós, J., et al.: Multi-oriented English text line extraction using background and foreground information. In: The Eighth IAPR International Workshop on Document Analysis Systems. DAS 2008, pp. 315–322. IEEE (2008)
7. Roy, P.P., Pal, U., Llados, J., et al.: Multi-oriented and multi-sized touching character segmentation using dynamic programming. In: 10th International Conference on Document Analysis and Recognition. ICDAR 2009, pp. 11–15. IEEE (2009)
8. Velázquez, A., Levachkine, S.: Text/graphics separation and recognition in raster-scanned color cartographic maps. In: Lladós, J., Kwon, Y.-B. (eds.) GREC 2003. LNCS, vol. 3088, pp. 63–74. Springer, Heidelberg (2004)
9. Pezeshk, A., Tutwiler, R.L.: Automatic feature extraction and text recognition from scanned topographic maps. IEEE Transactions on Geoscience and Remote Sensing **49**(12), 5047–5063 (2011)
10. Burt, P., Adelson, E.: The Laplacian pyramid as a compact image code. IEEE Transactions on Communications **31**(4), 532–540 (1983)
11. Chiang, Y.Y., Knoblock, C.A.: Recognition of multi-oriented, multi-sized, and curved text. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 1399–1403. IEEE (2011)
12. Chiang, Y.Y., Knoblock, C.A.: An approach for recognizing text labels in raster maps. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 3199–3202. IEEE (2010)

A Digital Video Stabilization System Based on Reliable SIFT Feature Matching and Adaptive Low-Pass Filtering

Jun Yu[1], Chang-Wei Luo[1], Chen Jiang[1,2], Rui Li[1,2], Ling-Yan Li[1],
and Zeng-Fu Wang[1,2(✉)]

[1] Department of Automation, University of Science and Technology of China,
Hefei 230026, China
{harrtjun,zfwang}@ustc.edu.cn
[2] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China

Abstract. A real-time digital video stabilization system is proposed to remove unwanted camera shakes and jitters. Firstly, SIFT algorithm is improved to extract and match features between the reference frame and current frame reliably, and then global motion parameters are obtained based on the geometric constraint consistency between feature matches through random sample consensus algorithm. Secondly, multiple evaluation criteria are fused by an adaptive low-pass filter to smooth global motion for obtaining correction vector, which is used to compensate the current frame. Finally, stabilized video is obtained after each frame is completed by combining the texture synthesis method and the spatio-temporal information of video. The objective experiments demonstrate the system can increase the average peak signal-to-noise ratio of jittered videos around 6.12 dB, The subjective experiments demonstrate the system can increase the identification ability and perceptive comfort on video content.

Keywords: Global motion estimation · Motion filtering and compensation · Video completion

1 Introduction

Videos retrieved from hand-held video cameras are affected by unwanted camera shakes and jitters, resulting in video quality loss [1]. Digital video stabilization techniques have gained consensus, for they permit to obtain high quality and stable video footages by making use only of information drawn from footage images and do not need any additional knowledge about camera physical motion [2][3].

There are three stages for digital video stabilization: global motion estimation [4], motion filtering and compensation [5], video completion [6].

Global motion estimation can be performed by global intensity alignment approaches [7-10] or feature-based approaches [11-13]. Feature-based methods are generally faster than global intensity alignment approaches, while they are more prone to local effects. A good survey on global motion estimation can be found in [4].

After estimating the global motion, motion filtering is removing the annoying irregular jitter to recognize intentional movement. It can be performed by DFT filtered

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 180–189, 2015.
DOI: 10.1007/978-3-662-48570-5_18

frame position smoothing [10], Kalman filtering [14] and motion vector integration [15] according to real system constraints [16][17]. After motion filtering, motion compensation is applied to spatially displace image frames by correction vector from the filtering result.

The goal of video completion is filling in missing image areas in a video [18]. It can be performed by mosaicing [19], sampling spatio-temporal volume patches [20], multi-layers segmenting [21][22] and local motion estimation of missing image areas [23][24]. The texture synthesis method [25][26] searches the similar texture patch to replace the unknown part in the missing image area. Good result can be obtained if enough similar information are available.

In this paper, a digital video stabilization system is proposed (Fig. 1). Our work has following advantages: 1) To increase the reliability of invariant feature transform (SIFT) algorithm, non-maximum suppression is used to obtain evenly distributed feature points, and multi-objective optimization is used to improve the feature matching accuracy. 2) Feature matches are used to estimate global motion by random sample consensus (RANSAC) fitting. 3) An adaptive low-pass filter with adaptive length according to the variation of global motion is constructed, thus over stabilization and under stabilization are prevented effectively. 4) Multiple evaluation criteria are fused to increase the robustness of motion filtering and compensation. 5) The performance of image texture synthesis method [25] is promoted with the plentiful spatio-tempral information of video to conduct the video completion.

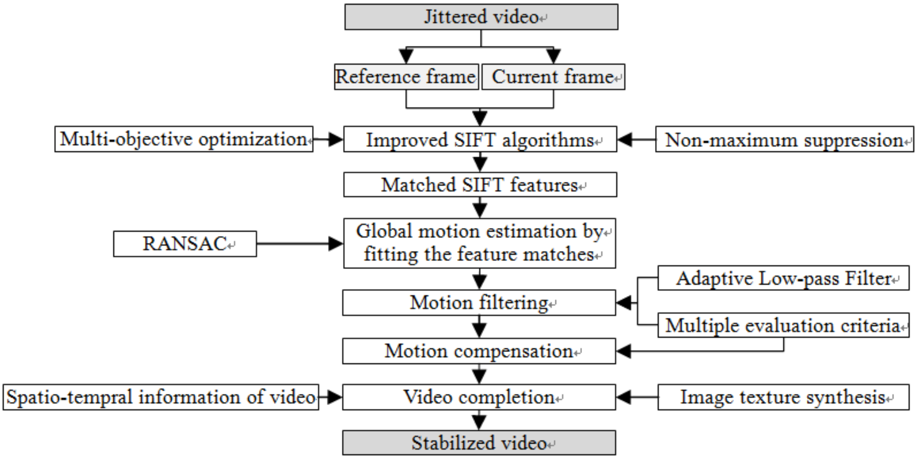


Fig. 1. Framework.

2 Image Matching

SIFT algorithm has been shown excellent performance in image matching [27]. It can be divided into three stages: feature detection, feature description and feature matching. However, there are two shortcomings of SIFT algorithm, namely non-evenly distribution of the feature points and non-adaptive feature matching strategy.

A large range, evenly distribution of feature points is a key factor ensuring the quality of image matching. In the feature detection stage of SIFT algorithm, the feature points are determined by the comparison of the extreme of the 26 surrounding pixels; thus the extreme value detected represents 27 pixels of the feature points, which are likely to fall into the local extremes. The spatial distribution of the detected feature points tend to be concentrated within a certain range, and the feature points may reflect only one or a few objects characteristics of the image. Whereas the required feature points should be able to reflect the overall characteristics of image, not just some local characteristics. In order to obtain a more uniform distribution of feature points, a large detection range should be considered. The reason is as follows. The greater the detection range, the greater the range of the local extremes represented by feature points. When the feature points are treated as local extremes in a wider range, the distances between them are more distant and a more uniform distribution of feature points will be obtained. Therefore, a detecting feature method with the non-maximal suppression [28] is used. There are (2×*r*+1)×2−1+18 pixels used to detect extremes with the radius of *r* at the current scale and 18 pixels at the adjacent scales, not only 26 pixels. In the original SIFT algorithm, feature detection is to compare feature points with 8 pixel at the current scale and 18 pixels at the neighbors and scales, while the used detecting feature method [28] is to compare feature points with 48 pixels at the current scale and 18 pixels at the neighbors and scales. Although the number of feature points detected by the used method is reduced, the features are distributed on a wider scope.

Because the SIFT feature points are disorder, and are not described regularly, such as corner, straight line, edge. So the normal image matching technology, such as relevant matching, is hard to achieve high accuracy. So the multi-objective optimization theory [29] is introduced into SIFT feature matching to reduce the error matching ratio, which consider Euclidean distance between correlation coefficient and feature point as the objective function and the confidence degree is taken as the constraint. The optimization purpose is to select the most satisfactory scheme from many alternative ones according to a more than one objective. The advantage of multi-objective optimization is that we can regulate the trade-off problem among multi-objectives to make them realize optimization at the same time under some certain constraint conditions. The details refer to [29].

As a result, the result of image matching is a list of keypoints pairs that can be easily used as the input of feature-based motion estimation stage.

3 Global Motion Estimation

Based on the perspective projection imaging model, the global motion, associating feature $\left(x_i, y_i\right)^T$ in frame I_n with feature $\left(x_j, y_j\right)^T$ in frame I_{n+1}, is described by:

$$\begin{aligned} x_j &= \left(a_1 x_i + a_2 y_i + a_3\right) / \left(a_7 x_i + a_8 y_i + 1\right) \\ y_j &= \left(a_4 x_i + a_5 y_i + a_6\right) / \left(a_7 x_i + a_8 y_i + 1\right) \end{aligned} \tag{1}$$

where $(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8)$ are the parameters to be solved.

The whole set of feature matches probably includes wrong matches or correct matches that indeed belong to self-moving objects in the filmed scene. Here RANSAC [30] is used to deal with this problem. Firstly, six couples of features are selected randomly from the feature set, and a solution is obtained from them. Then a subset of feature set is obtained by $\boldsymbol{S}_1^* = \left\{ \left((x_i, y_i)^T ; (x_j, y_j)^T \right) \middle| \left\| \begin{matrix} x_j - (a_1 x_i + a_2 y_i + a_3)/(a_7 x_i + a_8 y_i + 1) \\ y_j - (a_4 x_i + a_5 y_i + a_6)/(a_7 x_i + a_8 y_i + 1) \end{matrix} \right\| \le T \right\}$, T is a given threshold. Secondly, above process is repeated K times [13], and the subset with the most elements is selected. Finally, LM method is applied on the selected subset to obtain the final solution.

4 Motion Filtering and Compensation

An adaptive low-pass filter and multiple evaluation criteria are applied in the motion filtering. The following low-pass filter is used:

$$h(t) = \begin{cases} \sin(2\pi t/N - 1)/(2\pi t/N - 1) & -(N-1)/2 \le t \le (N-1)/2 \\ 0 & other \end{cases} \tag{2}$$

where N is the length of filter. To make N be adjusted according to the variation of global motion parameters, N is initialized as 5 manually after experiments, then following indices are computed:

Cumulative variation of gloabal motion parameters:

$$S = \sum_{i=1}^{N} \left| \boldsymbol{M}_i - \boldsymbol{M}_{ave} \right|, \boldsymbol{M}_{ave} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{M}_i \tag{3}$$

Max variation of gloabal motion parameters:

$$\rho = \max \left\{ \left| \boldsymbol{M}_i - \boldsymbol{M}_{ave} \right|, i = 1, 2, \cdots, N \right\} / S \tag{4}$$

Smoothness of gloabal motion parameters:

$$\lambda = S / \boldsymbol{M}_{ave} \tag{5}$$

Then a max threshold of λ (*threshold 1*) and a min threshold of ρ (*threshold 2*) are determined manually after experiments. Finally, N is adjusted online during stabilization process: if λ is smaller than *threshold 1*, and ρ is smaller than *threshold 2*, N is increased. Otherwise, N is decreased.

Firstly, the estimated global motion parameters $\boldsymbol{M}$ is chosen as one criterion to evaluate the video jitter. The adaptive low-pass filter is applied on $\boldsymbol{M}$, and the smoothing components are set as the motion filtering result. However, we found the motion filtering result of $\boldsymbol{M}$ is not satisfying when the jitter has very frequent tiny rotation component. The reason is: when the rotation component is very tiny, the filtering result is almost same to the original value, thus the compensation effect is very limited, and the human visual system still feel jittery when watching the compensated result. To alleviate this problem, the Euclidean distance of matched keypoints (EDMK) between adjacent frames is used as the second criterion, and the adaptive low-pass filter is also used to smooth the x component and y component of EDMK. Finally, the average of the filtering results by both criteria is set as the final result.

After motion filtering, the motion compensation is conducted as follow:

Firstly, the correction vector for the first criterion is obtained by computing the difference between original parameters $\boldsymbol{M}$ and filtering parameters $\hat{\boldsymbol{M}}$. Then the motion compensation is applied according to the equation (1). The only difference is (x, y) is the pixel position.

Secondly, the correction vector for the second criterion is obtained by computing the difference between original EDMK and filtering EDMK. Then the motion compensation is applied by displacing the pixel according to the correction vector.

Finally, the coordinates of pixels are set as the average coordinates of the pixels compensated by the first evaluation criterion and the pixels compensated by the second evaluation criterion.

5 Video Completion

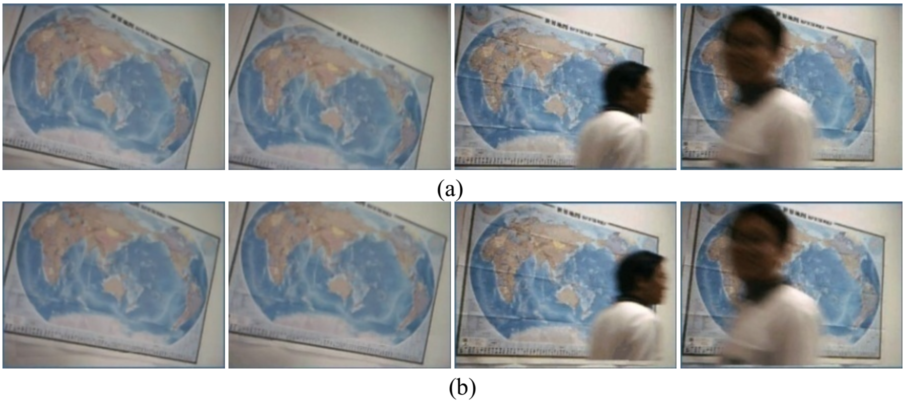
Good result can be obtained by the texture synthesis method [25][26] if enough similar information are available. However, it is hard to obtain satisfying result only by this method because the similar information in the single image is usually not enough. The texture synthesis method can be improved if the plentiful interframe information of video is introduced. The way of combining them is: the most similar texture patch of an original texture patch $\boldsymbol{A}$ in the current frame is searched in the adjacent frames by the texture synthesis method. If it is found, and the found texture patch is $\boldsymbol{B}$ in the adjacent frames, the neighbor texture patch of $\boldsymbol{B}$ will have the high priority to be the most similar texture patch of the neighbor texture patch of $\boldsymbol{A}$ during searching. If it is not found, it is searched in the current frame by the texture synthesis method.

6 Experiments

Experiments are conducted using a workstation with AMD Athlon (tm) II X4 640 3.01G, memory 2G, NVIDIA GT200 and CUDA 1.3.

Two jittered videos are captured [31]. The first is the video without moving object, and has 2476 frames, while the second is the video with moving object, and has 3124 frames. The GPU+CPU framework [32] is used to achieve the real-time ability. Because the global motion estimation, motion filtering and compensation need large computation, they are implemented in GPU, while other parts are implemented in CPU. In addition, the GPU implement of SIFT [33] is used to accelerate the process of feature extraction.

Fig. 2 shows the video stabilization results on the captured videos.



(a)

(b)

Fig. 2. (a) The frames before video stabilization. (b) The frames after video stabilization.

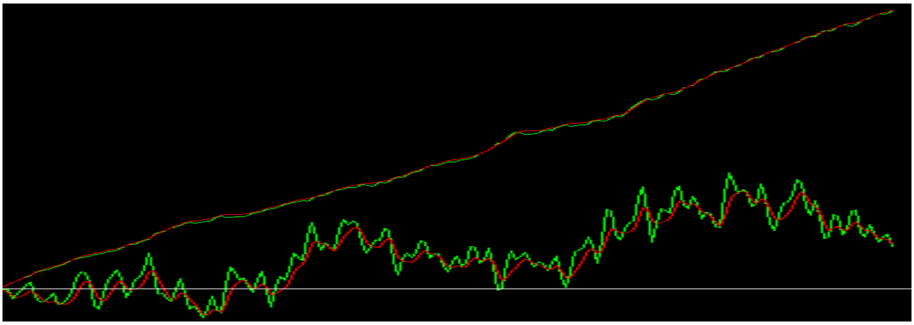


Fig. 3. Green curves are a_3, a_6 of the original video, red curves are a_3, a_6 of the stabilized video.

Fig. 3 is the motion filtering results of $\boldsymbol{M}$ by the adaptive low-pass filter. It shows the adaptive smoothing effect of the filter.

An index, peak signal-to-noise ratio (*PSNR*) between the reference frame $\boldsymbol{S}_0$ and current frame $\boldsymbol{S}_1$, is defined to evaluate the stabilization quality:

$$PSNR\left(\boldsymbol{S}_1,\boldsymbol{S}_0\right)=10\cdot\log_{10}^{255^2/MSE(\boldsymbol{S}_1,\boldsymbol{S}_0)} \quad (6)$$

where *MSE* is the mean square error of pixel value between two images. This index reflects the coherence between two images. The large the index, the better the video stabilization result.

Table 1 show the average *PSNR* on captured videos. As can be seen from it, the average *PSNR* is increased by the proposed video stabilization method around 6.12 dB, and the real-time ability is also achieved. Therefore, jittered video is stabilized by the proposed method nicely in real-time.

Table 1. Qualitative evaluation result of video stabilization.

	Average *PSNR* of original videos	Average *PSNR* of stabilized videos	Average time each frame takes
Captured video	25.35	31.47	0.045s

6.1 Improved SIFT Vs. Original SIFT

To evaluate the performance of the improved SIFT algorithm, experiment is conducted on different pairs of images from a standard LEAR image database [32]. Table 2 shows that the matching correct rate of improved SIFT is outperform that of original SIFT. From it, we can see the effectiveness of the improvements on the distribution of feature points and the feature matching strategy.

Table 2. Comparison of the matching correct rate between improved SIFT and original SIFT.

	Original SIFT	Improved SIFT
LEAR image database	84.56%	89.67%

6.2 Using Single Evaluation Criterion Vs. Fusing Multiple Evaluation Criteria

The effect of fusing multiple evaluation criteria is verified on a video clip, in which some very frequent tiny rotation component is added. From Table 3, we can see the superiority of fusing multiple evaluation criteria.

Table 3. Evaluation between single evaluation criterion and multiple evaluation criteria.

	Average *PSNR* of original videos	Average *PSNR* of stabilized videos	Average time each frame takes
Single evaluation criterion	18.56	23.67	0.037s
Multiple evaluation criteria	18.56	24.56	0.045s

6.3 Objective Comparison with Other Algorithm

The method in [24] is one of the state-of-the-art video stabilization methods. We have implemented it, then it and the proposed method are tested on the above video clip. We can see the proposed method is superior to the method in [24] from Table 4. This is because the proposed method fuses multiple evaluation criteria to conduct motion filtering by an adaptive low-pass filter, and the SIFT algorithm is improved to extract and match features between the reference frame and current frame reliably.

Table 4. Evaluation of several video stabilization algorithms.

	Average *PSNR* of original videos	Average *PSNR* of stabilized videos	Average time each frame takes
The proposed method	18.56	24.63	0.045s
The method in [24]	18.56	24.21	0.053s

6.4 Subjective Comparison with Other Algorithm

The problem with an objective evaluation is that the absolute truth of camera motion is not known. However, it is less problematic for the subjective evaluation since the human visual system is very sensitive to the video jitter. Therefore, user's reactions interacting with this system are evaluated.

34 users participate in the evaluation. The goal of the evaluation is to decide if the system can remove the discomfort on human visual system, and if the objects in the stabilized video can be identified easily.

In the first stage, the questionnaire is chosen for participants. Table 5 shows the constructs and questions of the survey related to the system performance. The answers to these questions are given from 'disagree' to 'agree' on a ten point scale. A Cronbach's alpha test [34] is carried out to determine if these constructs refer to the same topic. Typically, an alpha of 0.7 or greater is considered acceptable in psychological experiments. As Table 5 shows, all the alpha values obtained are greater than 0.7, indicating that the questionnaire is suitable for the evaluation in this paper.

Table 5. Cronbach's alpha results of questionnaire and mean scores after evaluation.

Construct	Question	Cronbach's alpha	Mean score of the proposed method	Mean score of the method in [24]
Smoothness	If the stabilized video is smooth and coherent.	0.743	7.79	6.73
Identification	If objects in the stabilized video can be identified easily.	0.811	7.75	6.48

In the second stage, the developed system and the method in [24] perform stabilization on captured videos, then participants compare stabilized videos with original videos. Finally, the questionnaire is filled. Table 5 shows the result of mean scores after evaluation. The maximum is 10, while the minimum is 0. For the developed system, all the scores obtained are greater than 7.5, and are higher than those of the method in [24], indicating that it has the ability to remove the discomfort on human visual system, and the objects in the stabilized video can be identified easily.

7 Conclusion

A real-time, reliable and adaptive digital video stabilization system is proposed. The SIFT algorithm is improved to match the adjacent frames robustly. Global motion parameters are obtained by RANSAC effectively. Multiple evaluation criteria are fused to conduct motion filtering by an adaptive low-pass filter. The spatio-temporal information are combined with the texture synthesis method to obtain a complete video. In future, the accuracy of motion estimation will be further improved.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61303150), the Open Project Program of the State Key Lab of CAD&CG (No. A1501), Zhejiang University.

References

1. Ejaz, N., Wonil, K., Soon II, K., et al.: Video stabilization by detecting intentional and unintentional camera motions. In: ICISMS, pp. 312–316. IEEE Press, New York (2012)
2. Chen, C.H., Chen, C.Y., Chen, C.H., et al.: Real-Time Video Stabilization Based on Vibration Compensation By Using Feature Block. IJICIC **7**, 5285–5298 (2011)
3. Seok-Jae, K., Tae-Shick, W., Dae-Hwan, K., et al.: Video stabilization based on motion segmentation. In: ICCE, pp. 416–417. IEEE Press, New York (2012)
4. Dung, T.V., Lertrattanapanich, S., et al.: Real time video stabilization with reduced temporal mismatch and low frame buffer. In: ICCE, pp. 61–62. IEEE Press, New York (2012)
5. Puglisi, G., Battiato, S.: A Robust Image Alignment Algorithm for Video Stabilization Purposes. TCSVT **21**, 1390–1400 (2011)
6. Puglisi, G., Battiato, S.: Robust video stabilization approach based on a voting strategy. In: ICIP, pp. 629–632. IEEE Press, New York (2011)
7. Abraham, S.C., Thomas, M.R., Basheer, R., et al.: A novel approach for video stabilization. IEEE Recent Advances in Intelligent Computational Systems **1**, 134–137 (2011)
8. Ko, S.J., Lee, S.H., Lee, K.H.: Digital image stabilizing algorithms based on bit-plane matching. TCE **44**, 617–622 (1998)
9. Ko, S.J., Lee, S.H., Jeon, S.W., Kang, E.S.: Fast digital image stabilizer based on gray-coded bit-plane matching. TCE **45**, 598–603 (1999)
10. Erturk, S., Dennis, T.J.: Image sequence stabilization based on DFT filtering. IEE Proceedings on Image Vision and Signal Processing **127**, 95–102 (2000)
11. Bosco, A., Bruna, A., Battiato, S., Bella, G.D.: Video stabilization through dynamic analysis of frames signatures. In: ICCE, pp. 312–316. IEEE Press, New York (2006)
12. Veon, K.L., Mahoor, M.H., Voyles, R.M.: Video stabilization using SIFT-ME features and fuzzy clustering. In: IEEE/RSJ ICIRS, pp. 2377–2382. IEEE Press, New York (2011)
13. Windau, J., Itti, L.: Multilayer real-time video image stabilization. In: IEEE/RSJ ICIRS, pp. 2397–2402. IEEE Press, New York (2011)
14. Erturk, S.: Image sequence stabilization based on kalman filtering of frame positions. Electronics Letters **37**, 95–102 (2001)
15. Paik, P.: An adaptative motion decision system for digital image stabilizer based on edge pattern matching. Consumer Electronics, Digest of Technical Papers (1992)
16. Auberger, S., Miro, C.: Digital video stabilization architecture for low cost devices. In: ISISPA, pp. 474–483. IEEE Press, New York (2005)

17. Tico, M., Vehvilainen, M.: Constraint translational and rotational motion filtering for video stabilization. In: ESPC, pp. 1474–1483. IEEE Press, New York (2005)
18. Zhiyong, H., Fazhi, H., Xiantao, C., et al.: A 2D-3D hybrid approach to video stabilization. In: ICCADCG, pp. 146–150. IEEE Press, New York (2011)
19. Litvin, A., Konrad, J., Karl, W.: Probabilistic video stabilization using kalman filtering and mosaicking. In: IS&T/SPIE SEIIVC, pp. 663–674. IEEE Press, New York (2003)
20. Wexler, Y., Shechtman, E., Irani, M.: Space-time video completion. In: CVPR, pp. 120–127. IEEE Press, New York (2004)
21. Jia, J., Wu, T., Tai, Y., Tang, C.: Video repairing: inference of foreground and background under severe occlusion. In: Proc. CVPR, pp. 364–371. IEEE Press, New York (2004)
22. Cheung, S.C.S., Zhao, J., Venkatesh M.V.: Efficient object-based video inpainting. In: ICIP, pp. 705–708. IEEE Press, New York (2006)
23. Cheung, V., et al.: Video epitomes. In: CVPR, pp. 42–49. IEEE Press, New York (2005)
24. Matsushita, Y., Ofek, E., Ge, W.N., et al.: Full-frame video stabilization with motion inpainting. TPAMI **28**, 1150–1163 (2006)
25. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. TIP **13**, 1200–1212 (2004)
26. Tang, F., Ying, Y.T., Wang, J., et al.: A novel texture synthesis based algorithm for object removal in photographs. In: ACSC, pp. 248–258. IEEE Press, New York (2005)
27. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60**, 91–110 (2004)
28. Gao, T., et al.: Multi-Scale Image Registration Algorithm based on Improved SIFT. Journal of Multimedia **8**, 755–761 (2013)
29. Zheng, Y., et al.: Video Image Tracing Based on Improved SIFT Feature Matching Algorithm. Journal of Multimedia **9**, 130–137 (2014)
30. Hoper, P.J.: Robust statistical procedures. SIAM (1996)
31. Yu, J., Luo, C.-w., Jiang, C., Li, R., Li, L.-y., Wang, Z.-f.: Real-time robust video stabilization based on empirical mode decomposition and multiple evaluation criteria. In: Zhang, Y.-J. (ed.) ICIG 2015. LNCS, vol. 9219, pp. 125–136. Springer, Heidelberg (2015)
32. Juang, C., et al.: Speedup of implementing fuzzy neural networks with high-dimensional inputs through parallel processing on graphic processing units. TFS **19**, 717–728 (2011)
33. http://cs.unc.edu/~ccwu/siftgpu/
34. Marcosa, S., Gómez-García-Bermejob, J., Zalama, E.: A realistic, virtual head for human-computer interaction. Interacting with Computers **22**, 176–192 (2010)

A Tree-Structured Feature Matching Algorithm

Xiongwei Sun[1], Xiubo Ma[2], Lei Chen[1], Li Wan[1], and Xinhua Zeng[1](✉)

[1] Hefei Institutes of Physical Science, Chinese Academy of Science, Hefei, China
xiongweisun@163.com, xhzeng@iim.ac.cn

[2] Department of Computer Engineering, AnHui SanLian University, Hefei, China
maxiubo0310@126.com

Abstract. Feature matching is essential in computer vision. In this paper, we propose a robust and reliable image feature matching algorithm. It constructs several matching trees in which nodes correspond to traditional sparsely or densely sampled feature points, and feature lines are constructed between the nodes to build a cross-references based on a Difference-of-Gaussians down-sampling pyramid. This can make patch-based descriptors combine efficiently with spatial distributions. By comparing with SIFT, SURF and ORB, our method can get much more correct correspondences on both synthetic and real data under the influence of complex environments or transformations especially in irregular deformation and repeated patterns.

Keywords: DoG · Image feature matching · Tree structured matching · Feature line

1 Introduction

Feature correspondence is a fundamental task in many applications of computer vision such as feature tracking[13], image classfication[11], object detection[4], 2D and 3D registration[10,17]. A large number of applications promote various kinds of feature matching algorithms. At the same time, the continuous development of the applications put forward some new and higher requirements such as precision, speed and robust ability. In order to meet the needs of all these practical applications, much attention has been paid to improve the matching performance. A widely used method is computing variety of feature descriptors and select a threshold carefully to filter out large outliers. Therefore, variety of feature descriptors have been proposed, such as SIFT[8], SURF[9], BRISK[15], ORB[14] and LDB[19]. Further more, some people turned to combine more flexible geometric features and spatial characteristics. For instance, Chui et al.[5] introduced a feature based method named TPS-PRM (thin-plate spline-robust point matching) for non-rigid registration. C.Schmid[16] and Y.Zheng[20] use the thought of proximity. They assumed that two adjacent points in the original image should be matched to the couples which are also neighbours in the target image. X.Xu[18] use RANSAC and strong space constraints to obtain relatively stable feature point set first and then use a selection model[10] to decide which transformations are the most appropriate one. Finally, it constructs a global geometric

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 190–200, 2015.
DOI: 10.1007/978-3-662-48570-5_19

transformation model as the matching constraint. O.Duchenne[6] accommodate both (mostly local) geometric invariants and image descriptors and search for correspondences by casting it as a hyper graph matching problem using higher order constraints.

Although many existing algorithms are general and could cover both rigid and non-rigid matching problems according to the problem definition, most of the them are either too computationally expensive to achieve real-time performance, or not sufficiently distinctive to identify correct matches from a large database with various transformations.

In this paper, we propose a tree structured hybrid feature matching algorithm, called DoG-based Random Grow (DoG-RG). In order to solve the problems mentioned above, we summarize our contributions as follows:

1. Flexible tree structure can effectively improve the patch-based discrimination of feature matching by combine feature lines with spatial distributions.
2. In order to increase the distinguish, we build a iterator method on the feature lines correspondences based on down sampling DoG pyramid.
3. Cell-space partitioning algorithm is used to reduce the selection number of candidate points in a limited area and which drastically speed up the matching process.

The remainder of the paper is organized as follows: Section 2 presents details of the proposed algorithm. In section 3, we compare performance of DoG-RG with some existing outstanding algorithms on public benchmarks. Section 4 gives the concluding remarks.

2 Algorithm Details

Suppose we have extracted two sets of feature points P^S and P^T from source image I^S and target image I^T. The overview of our proposed framework is shown in Figure 1. Firstly,we present the feature line extraction method; Secondly, we show the tree structure's start points(we call it anchors); Thirdly, we show the details of the exploring random tree(DoG-RG).

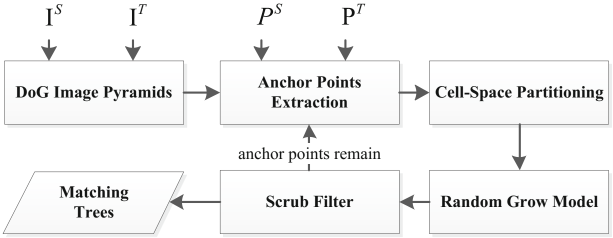


Fig. 1. The framework of the algorithm

2.1 Image Pyramids and Feature Lines

Considering the computational complexity, antinoise ability and characteristics of resolution, we use the DoG pyramid as the reference matching substrate. The lines between feature points are casted on the surface of the pyramid called feature lines.

Image Pyramids. We define L(x,y,δ) as a level in the multi-scale images and it is formed by a gaussian function G(x,y,δ) and an image I(x,y) convolution[12],described as follows:

$$L(x,y,\delta) = G(x,y,\delta) \otimes \; I(x,y) \tag{1}$$

We set $\otimes$ as the operator of convolution and G(x,y,δ) is:

$$G(x,y,\delta) = \frac{1}{2\pi\delta^2}\; e^{\frac{-(x^2+y^2)}{2\delta^2}} \tag{2}$$

We set σ as 1.5 and make a substraction between two adjacent scale-space images. The difference of gaussian image is denoted as D(x,y,δ). The function is given as follows:

$$D(x,y,\alpha) = S(k) * ((G(x,y,\sigma\delta) - G(x,y,\delta)) \otimes I(x,y) \tag{3}$$

where k is the down sampling factor, S is the down sampling function which is introduced to reduce the computational burden of feature lines's extraction and increase the robustness of feature lines.

Feature Lines. We cast the feature points onto the reference matching substrate of DoG pyramid through coordinate conversion. As is shown in Figure 2, the feature lines projected on the surface of substrate present different fluctuations. The higher the level of the substrate is, the more stable of its fluctuations will become. On the contrary, The lower level substrate is, the stronger the resolution of the feature line will be.

In order to construct feature lines, we sample discrete pixels normally along the corresponding spaced feature points from DoG images. The similarity evaluation can be expressed as the follow mathematical formula:

Let FL^S and FL^T be the sequence of points extract from different levels of DoG images and for any $P_i^S(x,y,z) \in FL^S$,$P_i^T(x,y,z) \in FL^T$. Then normalize the length of FL^S and FL^T as $N = max(length(FL^S),length(FL^T))$. Here similarity η is defined as

$$\eta = \frac{\sum\limits_{i=1}^{N-1} sign(P_i^S) \bigodot sign(P_i^T)}{N-1} \tag{4}$$

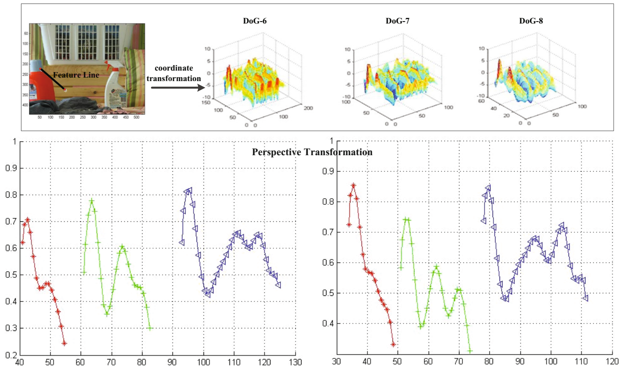


Fig. 2. An illustration of feature line extract from source and target image DoG-6:blue;green:DoG-7;red:DoG-8

where, $P_i.z$ means the gray value of DoG image and $sign(P_i)$ is defined as

$$sign(P_i) = \begin{cases} 1 \ , P_i.z - P_{i-1}.z > \varepsilon \\ -1 \ , P_i.z - P_{i-1}.z < -\varepsilon \\ 0 \ , abs(P_i.z - P_{i-1}.z) \leq \varepsilon \end{cases} \tag{5}$$

Where, the ε is a precision control factor given as follows, M is a feature line's resolution parameter.

$$\varepsilon = \frac{max(P_{set}.z - min(P_{set}.z))}{4M} \tag{6}$$

Generally, we set $N \geq 20$,otherwise, this feature line will be deemed invalid. More over, we set $\varepsilon = 0.7$, M = 13 and these settings are used in the subsequent experiments in this paper.

Anchor Points Extraction. In order to find more stable feature points (Anchor points), we establish a triangular structure, the start position of matching trees, which is constructed by three feature points and three feature lines. Based on the patch-based descriptors and more restricted similarities, we get small pieces of matched feature points. After that, several three-point combinations are randomly selected and checked by the spatial similarity and feature lines. If the triangular correspondence is wrong, the matching trees will always become low and will be filtered in the procedure Scrub Filter latter.

2.2 Tree Structured Random Grow Method

Space Partitioning. In order to speed up the matching process, we divide the feature points into grid cells according to space distribution. Assuming that feature points are under relatively uniform distributions, the division can contribute to avoiding a large number of outliers' operations.

For the subdivision, two basic principles are proposed as follows:

1. Minimize the number of points in each subdivision cell to improve the matching speed.
2. Retain enough cell size can increases the length of a feature line, strengthen the resolution and improve the matching accuracy.

As these two principles are contradictory, we propose an empirical formula. Suppose P^S's distribution area is $posW_S \times posH_S$, P^S size is $FeNum_S$,N is the minimum acceptable length of the feature lines. The division of cell number X_S can be calculated as:

$$X_S = min(\frac{\sqrt{posW_S * posH_S}}{2N}, \frac{\sqrt{FeNum_S}}{2}) \tag{7}$$

If the scale transformation happened, denoted by s, we can get the target point set P^T's division cell number X_T as:

$$X_T = min(\frac{\sqrt{posW_T * posH_T}}{2N}, \frac{\sqrt{FeNum_T}}{2}, \frac{X_S}{s}) \tag{8}$$

In the division, we put the feature points to the split cell grids. This strategy quite good to the quick index of feature points and exclude the outliers. Thus, It drastically increase the matching speed and improves the precision.

Random Grow Method. Suppose A^S is anchor point and let it as the root of a whole matching tree. Matching process starts from A^S and then search new points in the range of r^{branch} around. Assuming point D^S is belong to the range of r^{branch}, we use feature line $\overrightarrow{A^S D^S}$ and position relations to check the corresponding point D^T in the target point set P^T.

If the feature point D matching success, we set D^S as a new growing point and shrink the searching area to eight-neighborhood region. As shown in the Figure 3. In order to maintain the distinction of feature lines, our eight-neighborhood area ignore the center cell which is marked blue in the Figure 3. With the reference of the position D^S, we find out the corresponding position E^{vir} in I^T. Then we draw a circle (E^{vir}, r^{leaf}) and extract the contact cells, obtain all the candidates in the cells such as E_1^T and E_2^T. Finally, we use the descriptors and feature lines to sift out the best match. Traditionally, patch-based descriptors may hard to distinguish local repeat mode. Here,with the back-trace strategy of tree structure, we can easily find out local repetitive patterns and determine which one is the best. For example, if the feature lines from D^T failed to distinguish E_1^T and E_2^T, we just backtrack to A^T and build new feature lines to avoid passing through duplicate regions.

In the entire search process, we continually use the matched points to deduce the next points nearby till the end of matching process.

We continue the performs of algorithm till it can not find new anchor points to generate more matching trees. Since we can not ensure all the anchor points are correct matches, we have to filter the scrub to ensure a better accuracy.

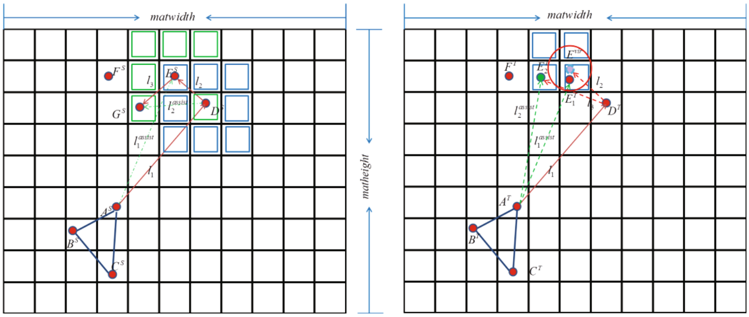


Fig. 3. An illustration of the matching process;The image at left reveals origin strategy and the one on the right shows the target

Due to the wrong matching trees don't match with the growing image region which always grow shorter, we can easily purify the matching trees by remove the scrubs.

3 Experimental Results

In the experiments presented here, we dived them into two parts: first we try to assess our method's actual capability on several image transform conditions(illumination, blur and compression). In order to guarantee the justice of experiment, we use OpenCV 2.4.6 to extract different size of feature points. To rule out the influence of patch-based descriptors, we repeat the experiments with the frequently-used descriptors: SIFT, SURF, ORB and BRISK.

Second, we give a set of images including local area transformation, irregular deformation and high repetitive pattern to show the good robustness of the algorithm.

3.1 Experiment Based on Different Local Feature Descriptors

Using the image groups of Tree, UBC and Leuven from data set [3], we exam the performance of blur, compress and light respectively. For each image group, the task is to match the first image to the remaining five, yielding five image pairs per sequence which are denoted as pair 1/2 to pair 1/6. Under the reference of the descriptors performance research [1], we carefully selected SIFT, SURF, ORB and BRISK as feature descriptors for the contrast experiments. To be fair, we compare DoG-RG with several the most frequently used feature descriptor combination algorithms integrated in OpenCV2.4.6, they are RADIUS, NNDR and BRUTE-FORCE. In order to verify the validity of corresponding

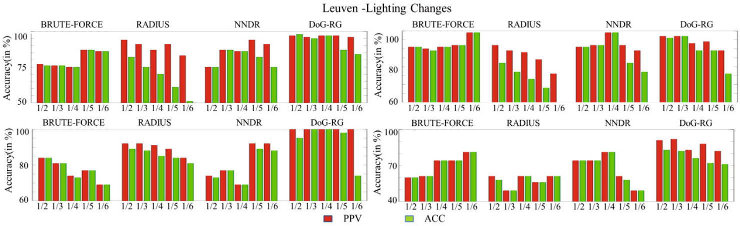


Fig. 4. PPV and ACC obtained by SIFT(Top left),SURF(Top right), ORB(Left bottom), BRISK(Right bottom) for the six image sequences of leuven

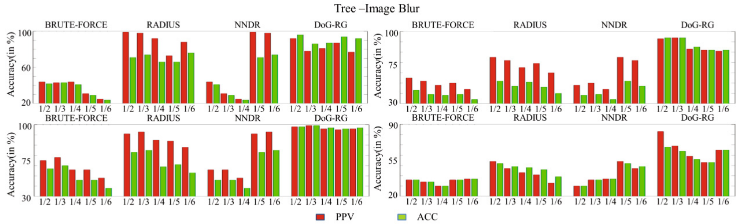


Fig. 5. PPV and ACC obtained by SIFT(Top left),SURF(Top right), ORB(Left bottom), BRISK(Right bottom) for the six image sequences of Trees

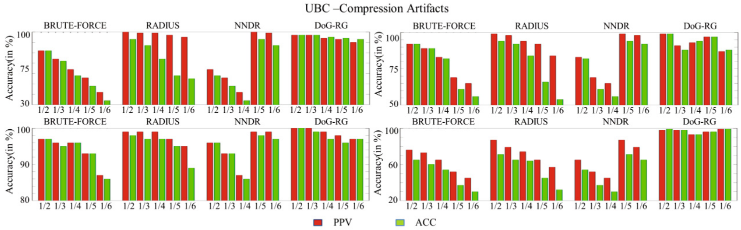


Fig. 6. PPV and ACC obtained by SIFT(Top left),SURF(Top right), ORB(Left bottom), BRISK(Right bottom) for the six image sequences of UBC

points, we use the combination of the above four algorithms to extract corresponding feature points. Correct matching enforces a one-to-one constraint so that a match is correct if two points are geometrically closet with sufficient overlap, and closest in feature space measure.

Two measures are introduced to evaluate the performances of all these methods according to the evaluation index ACC and PPV [7] and the calculating Formula 9 are listed below:

Table 1. Number of Feature Points Extracted

Sequence	Descriptor	1	2	3	4	5	6
Tree	SIFT	2613	2667	2940	3524	2163	2409
	SURF	1905	1872	1833	1637	1495	1424
	ORB	500	500	500	500	500	500
	BRISK	984	996	1033	995	818	573
Leuven	SIFT	861	682	574	521	433	349
	SURF	1313	1143	1036	937	802	650
	ORB	500	489	489	476	464	489
	BRISK	268	214	175	160	117	104
UBC	SIFT	1371	1348	1360	1418	1595	1597
	SURF	1602	1575	1620	1561	1582	1315
	ORB	500	500	500	500	500	500
	BRISK	546	558	507	503	571	774

Table 2. Threashold for matching.

Descriptor	RADIUS	NNDR	BruteForce	DoG-RG
SIFT	thr=0.24	ratio=1.0/1.2	thr=0.34	fl=0.6,DoG-level=8
SURF	thr=0.25	ratio=1.0/1.2	thr=0.35	fl=0.6,DoG-level=8
ORB	thr=65.0	ratio=1.0/1.1	thr=75	fl=0.6,DoG-level=8
BRISK	thr=145	ratio=1.0/1.1	thr=200	fl=0.6,DoG-level=8

$$\begin{aligned} Accuracy(ACC) &= \frac{TP+TN}{TP+TN+FN+FP} \\ Precision(PPV) &= \frac{TP}{TP+FP} \end{aligned} \tag{9}$$

Different feature points are extracted from sequence of test images, the points number are listed as Table 3.1.

We carefully select the thresholds for each patch-based method as shown in Table 2 by comprehensive considering of the overall performance. These settings are also used in DoG-RG's patch-based parts. Among them, fl is the feature line threshold, *thr* is the threshold of descriptors and *ratio* is used for NNDR.

Through the experiment, we find that tree structured method obtains a quite higher performance than other algorithms. The explaining is that patch-based descriptors and feature lines are local, but the tree structured feature lines are more "global", this flexible structure enables us to overcome patch myopia. This strategy is attractive because of its simplicity and flexibility. The combination of feature descriptors and tree structured feature lines can effectively suppress the unstability of accuracy in different conditions and obtain more excellent results in general.

3.2 Matching in Irregular Deformation and Repeating Pattern

In order to further the ability test on irregular transformations, we select several common image transformations in real life .

1. The ability to match the partial translations and rotations in one scene.
2. Test matching capabilities under partial irregular deformation.
3. Test matching capabilities in high repeat patterns.

In the following cases, we combine the descriptor SIFT with dynamic feature lines to complete the matching process. Matching results are shown in the mosaic: the uppers are the original images, the middles are the feature matching trees and the lowers are the connections of the corresponding points.

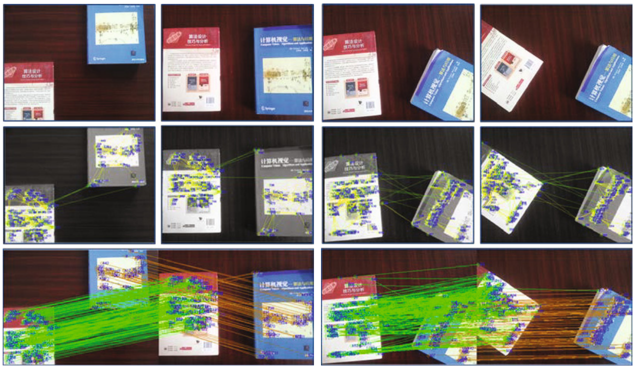


Fig. 7. (UP)Local mobile origin images(Middle)Matching trees (Bottom)Matching figure

Our algorithm can easily handle the partial inconsistent deformations are seen in the Figure 7. Explanation is as follows: by using the tree-structured searching strategy, local gentle irregular deformation can be easily cope with local tree nodes searching strategy, regional steep deformations in different transformations can be easily solved by bring more different matching trees in.

These local irregular deformations of fisheye images are come from [2]. It can be seen from the Figure 8 that feature lines from high level Difference-of-Gaussians pyramid can effectively adapt to the local irregular deformations. Generally, this combination of feature descriptors and tree structured feature lines have a quite good robustness in irregular deformations.

Seen from Figure 9, although we do not use the consistent algorithm to purify the result, this proposed method can effectively distinguish the repeat patterns effectively. Even in dense points distribution area, dynamic feature lines strategy can still automatically select appropriate connections to achieve a good matching result. At the same time, feature lines can also prevent the spread of error matches. Just as the description in the figure, very few mismatched feature

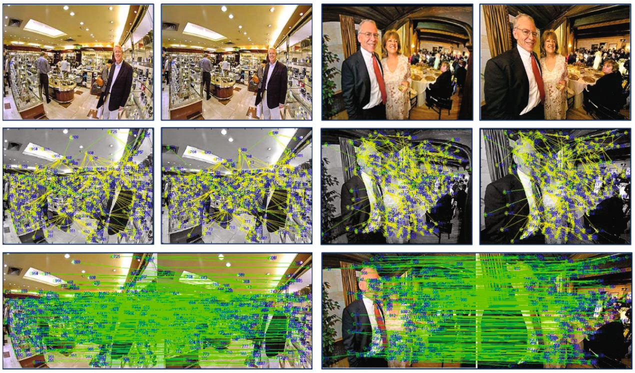


Fig. 8. (UP)Fisheye origin images(Middle)Matching trees (Bottom)Matching figure

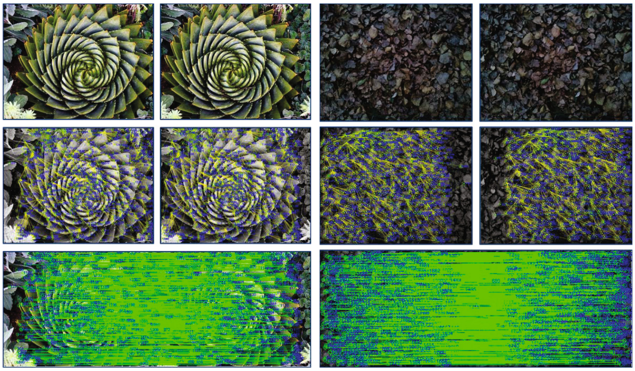


Fig. 9. (UP)Repetitive patterns (Middle)Matching trees (Bottom)Matching figure

points distributed in the border area are all isolated and these short trees will be removed in the scrub filter process.

4 Conclusion

In this paper, we propose a new image feature matching algorithm DoG-RG. By combining the feature lines with dynamic strategy and the patch-based feature descriptors, it constructs a incremental tree structured matching algorithm. The substantial benefits of this work is the good matching performance in simple calculation method and high robust ability. Experiment results show its better performance in common transformations and high local repetitive patterns. In addition, proposed methods can easily combine with various of patch-based descriptors to satisfy the needs of different matching conditions.

References

1. Comparison of feature descriptors. http://computer-vision-talks.com/2011/08/feature-descriptor-comparison-report/
2. Fisheye-hemi plug-in samples. http://imagetrendsinc.com/gallery/galleryhemi.asp
3. Visual geometry group-the oxford affine convariant regions datasets. http://www.robots.ox.ac.uk/vgg/research/affine/
4. Berg, A., Berg, T., Malik, J.: Shape matching and object recognition using low distortion correspondences. In: CVPR, vol. 1, pp. 26–33 (2005)
5. Chui, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration. Computer Vision and Image Understanding **89**(2c3), 114–141 (2003). nonrigid Image Registration
6. Duchenne, O., Bach, F., Kweon, I.S., Ponce, J.: A tensor-based algorithm for high-order graph matching. PAMI **33**(12), 2383–2395 (2011)
7. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letters **27**(8), 861–874 (2006)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
9. Bay, H., Tuytelaars, T., Van Gool, L.: Speeded-up robust features(surf). CVIU **110**(3), 346–359 (2008)
10. Izadi, M., Saeedi, P.: Robust weighted graph transformation matching for rigid and nonrigid image registration. IEEE Transactions on Image Processing **21**(10), 4369–4382 (2012)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR, vol. 2, pp. 2169–2178 (2006)
12. Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision **30**(2), 79–116 (1998)
13. Ong, E.J., Bowden, R.: Robust facial feature tracking using shape-constrained multiresolution-selected linear predictors. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(9), 1844–1859 (2011)
14. Rublee, E., Rabaud, V., Konolige, K.: Orb: an efficient alternative to sift or surf. In: ICCV, pp. 2564–2571 (2011)
15. Leutenegger, S., Chli, M., Siegwart, R.Y : Brisk: binary robust invariant scalable keypoints. In: ICCV, pp. 2548–2555 (2011)
16. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. PAMI **19**(5), 530–535 (1997)
17. Wu, Z., Goshtasby, A.: Adaptive image registration via hierarchical voronoi subdivision. IEEE Transactions on Image Processing **21**(5), 2464–2473 (2012)
18. Xu, X., Yu, C., Zhou, J.: Robust feature point matching based on geometric consistency and affine invariant spatial constraint. In: ICIP, pp. 2077–2081 (2013)
19. Yang, X., Cheng, K.T.: Local difference binary for ultrafast and distinctive feature description. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(1), 188–194 (2014)
20. Zheng, Y., Doermann, D.: Robust point matching for nonrigid shapes by preserving local neighborhood structures. PAMI **28**(4), 643–649 (2006)

Crack Detection in Tread Area Based on Analysis of Multi-scale Singular Area

Li Jinping[1,2(✉)], Hou Wendi[1,2], Han Yanbin[1,2], and Yin Jianqin[1,2]

[1] Institute of Pattern Recognition and Intelligent System, School of Information Science and Engineering, University of Jinan, Jinan 250022, China
[2] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China

Abstract. Tyre quality inspection is very important for tyre industry. In the present paper, an algorithm is proposed to detect cracks in tread area by using the idea of Hough transform and analyzing singular area obtained by multi-scale decomposition of wavelet transform. Firstly, tyre X-ray images are obtained by using X-ray beam. Secondly, a series of curves are obtained by projecting X-ray image of tyre to different angles. Thirdly, those projection curves are decomposed into multi-scale curves by wavelet transform, and the orientation and location of cracks are determined by analyzing the singularity of those multi-scale curves and the texture regularity of normal tread area images. The experimental results show that most of cracks in tread area can be detected effectively.

Keywords: Tyre tread area · X-ray image · Wavelet transform · Multi-scale analysis · Singular area · Crack detection

1 Introduction

Tyre quality inspection is very important for tyre industry, accurate detection of tyre defects can reduce economic and life loss caused by traffic accidents to a great extent.

There are many kinds of tyre, and the meridian tyre attracts the most concern due to its practical extensive applications [1]. Meridian tyre is composed of ply layer, belt layer, tread area, shoulder and steel bead, and the carcass is pinched by belt layer. In the process of manufacturing, meridian tyre has a very high requirement of technology,

Supported by National Natural Science Foundation of China (61203341), University Scientific Projects of Shandong Province (J12LN19, J14LN15).

L. Jinping–a professor of School of Information Science and Engineering (ISE) of University of Jinan, his present research interests include artificial intelligence, image processing, pattern recognition, optimization algorithms, *et al.* **H. Wendi**–a postgraduate of University of Jinan, her present research interest is pattern recognition, image processing. **H. Yanbin**–an associate professor of ISE of University of Jinan, his present research covers image processing, optimization. **Y. Jianqin**–an associate professor of ISE of University of Jinan, her present research covers image processing, machine vision, *et al.*

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 201–218, 2015.
DOI: 10.1007/978-3-662-48570-5_20

because the defects in tyre may reduce the normal performance and shorten the normal service life, and may further result in serious danger to the crew, vehicles and cargo. Therefore, the quality detection of tyre is very important.

The defects are generally inside the carcass of tyre, so special methods are often used to detect these defects. There are two basic detection approaches: the use of laser interference imaging [2] and the use of X-ray beam imaging.

The former uses vacuum load and laser interference imaging, and finds defects by analyzing the holographic images. In the state of vacuum, deformation of tyre will lead to the change of phase and light intensity of laser beam, and then we record all these changes in the holographic plate and finally obtain the holographic interference images. In the holographic interference images, dense interference stripes and sparse interference stripes indicate big deformation and small deformation, respectively, and regular distribution of interference stripe shows the normal deformation of tyre. Accordingly, by using holographic interference images, the position and the size of defects can be determined. Laser interferometer imaging, however, is not convenient in practical application, since it is costly and less intuitive, and requires professional knowledge; moreover, it can only detect such defects as bubble and wire delamination. So laser interferometer method has its inherent weakness.

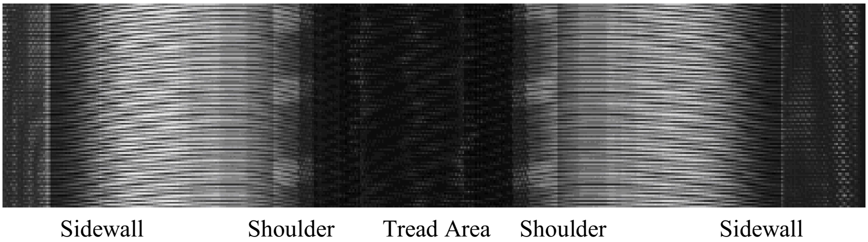


Fig. 1. Projection image of radial tyre X-ray image

The latter makes use of the difference of absorption capacity of different materials, *i.e.*, with of rotation of tyre, the projection images can be obtained when X-ray beam passes through the cross section of tyre from X-ray source to the image intensifier (Fig. 1).

Up to now, we can find many references about the manufacturing and improvement of X-ray imaging machine, but it is not easy to see any references about intelligent detection of tyre defects by using X-ray images, especially about the defect detection algorithms.

X-ray image of a tyre can reflect the basic structure of the tyre, which can be generally divided into three parts (Fig. 1): sidewall area, shoulder and tread area. Generally, it is a comparatively easier job to detect defects in sidewall area and shoulder, but it will be more difficult to detect defects in tread area, because the texture in tread area is much more complex than that in sidewall and shoulder. Currently, most of the references are discussing the detection of defects in sidewall area and shoulder [1, 3, 4], *e.g.*, abnormal array of the body cords and blister in sidewall area. The abnormal array of the body cords includes overlapped cord, open ply splices, wide spacing, crossed cord, broken cord and bending cord, *et al* (Fig. 2). It shall be pointed out here that these kinds of defects can be found in the whole tyre, i.e., we can find these defects in sidewall and shoulder, and can also find them in tread area.

Manual detection has many unfavorable factors, *e.g.*, personal subjectivity, heavy workload, eye strain, *et al*, which may result in a strong probability of misjudgment. In order to improve the detection efficiency, intelligent detection methods shall be developed.

The detection of defects from X-ray images can be divided into four basic steps: preprocessing, defect detection, feature extraction and defect classification. The main task of image preprocessing is to obtain a high-quality X-ray image, and manages to sharpen the difference between the defects and the background, the common employed methods include histogram equalization[1,3], filtering (*e.g.*, mean filter, median filter, smoothing filter) [4] and gray level adjustment [6], *et al.* It's obvious that the first two steps are the most important procedures. In most cases, X-ray images of tyre can be regarded as a kind of complex texture, and the defect detection is equivalent to the abnormal texture detection [7-8].

Thorough investigations indicate that the detection of defects in sidewall and shoulder attracts more attention whereas the researches of detection of defects in tread area are hardly seen, there are two reasons for this: one is the comparatively simpler texture background in sidewall and shoulder, the other is the more importance of sidewall and shoulder because the sidewall is thinner and weaker than tread area.

With the increasing requirement of tyre quality, the detection of defects in tread area needs the same urgent research as that in sidewall and shoulder. Crack, as an important and the most common defect in tread, seriously degrades the performance of tyre, and then are attracting more attentions gradually.

Most of defects in sidewall and shoulder can be detected by using histogram feature matching, adaptive threshold segmentation and chain-code tracking based on integral image [1], invariant moments and Fourier descriptor [1], mean image and variance image [5], frequency spectra analysis in polar coordinates and gray level co-occurrence matrix, *et al.* The detection of defects, for example, cracks, in tread area, however, is still under exploration. Some people may wonder whether the detection methods for sidewall and shoulder can apply for tread area, the answer is basically negative, the reason for this is: when employing the methods of detection of defects in sidewall and shoulder to detect the defects, e.g., cracks, in tread area, we found that the background textures in tread area are so complex that the features of those defects are drowned out by noises from the background textures. So we must find other effective methods to deal with cracks in tread area.

In the present paper, we introduce an effective and practical method to detect cracks in tread area by means of wavelet transform and multi-scale analysis of singular area. In section 2, we observe and analyze the features of cracks in tread area image, and then present detection algorithms briefly based on Hough transform and wavelet transform. For the sake of convenience and universality, in section 3, we perform an omni-directional rotating projection and then obtain a series of projection curves of tread area images. In section 4, we use multi-scale analysis of wavelet transform, and then obtain the singular curves corresponding to the part of tread area containing crack, and finally we are able to determine the position and orientation of crack in tread area image. The experimental results in section 5 demonstrate the proposed method is effective.

2 Crack Features of Tyre Tread Area and Detection Algorithm

2.1 Crack Features Analysis of Tyre Tread Area

With the knowledge of tyre tread area manufacturing process, tyre tread area is composed of regular layers of steel cords. In the ideal case, there are no defects in the tyre of good quality; In other cases, however, there may be some defects in tyre manufacturing due to the technical limitations in production technology and equipments. For example, multi-layer steel cord layers may not coincided completely or layers edge overlap, which can cause uneven layer thickness, and cracks may occur in the corresponding X-ray image.

There are many kinds of defects in tread area, *e.g.*, overlapped cord (Fig. 2.a-1), open ply splices (Fig. 2.b-1), overlap (Fig. 2.c-1), impurity (Fig. 2.d-1) and so on, Fig. 2.a-2, Fig. 2.b-2, Fig. 2.c-2 and Fig. 2.d-2 are the corresponding gray equalization images, respectively. Overlapped cord and open ply splices are more common defects, and both of them have the same line feature, we call them crack uniformly. In practice, we seldom see any cracks in the form of curved lines, so the detection of cracks with line pattern is the focus in the present paper.

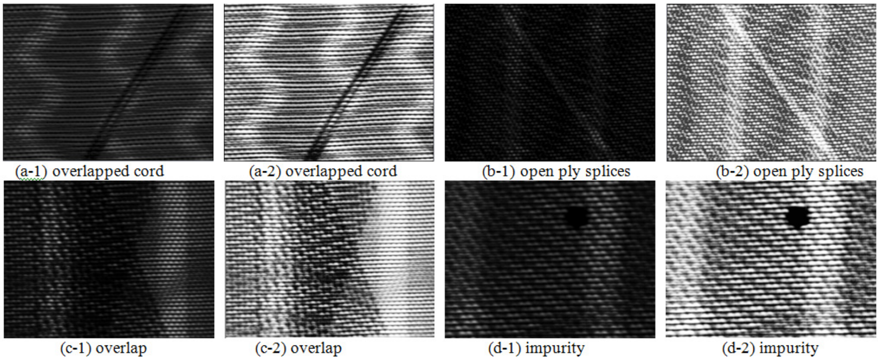

(a-1) overlapped cord (a-2) overlapped cord (b-1) open ply splices (b-2) open ply splices
(c-1) overlap (c-2) overlap (d-1) impurity (d-2) impurity

Fig. 2. Tyre defect images

2.2 Crack Detection Based on the Idea of Hough Transform

It is easy to see from the crack images that the grayscale distribution at cracks is abnormal in comparison with the rest part. So it is the first task to sharpen the local abnormal grayscale distribution (see Fig. 2.a-2, Fig. 2.b-2, Fig. 2.c-2 and Fig. 2.d-2) if we want to detect cracks in tread area effectively and accurately.

A traditional method to detect defects is Gabor transform. Up to now, Gabor filters constructed by Gabor transform have been extensively used in the detection of tyre texture defects [1, 7], and also extensively applied in the field of hot-rolled steel plate and fabric defects detection [9-13]. In essence, the application of Gabor filters is equivalent to a kind of curved surface fitting [12], and each filter determines a characteristic function. Since the texture in sidewall and shoulder is comparatively simpler and with better regularity, features can be extracted with Gabor filters preferably and easily. With the same reason, most of fabric textures are also relatively simpler and has obvious regularity, so Gabor filters are also wildly used in the detection of texture defects in fabric images [10-13]. As of tread area, the background texture is so complex (Fig. 1 and Fig. 2) that it will be a very hard job to perform the curved surface fitting, Gabor transform, consequently, is not a good candidate in the detection of defects in tread area.

Another traditional method for the detection of defects is Hough transform. As is known, Hough transform is often used to detect line patterns. From the preceding analysis, we know that the crack is generally a line pattern, so Hough transform can be obviously regarded as a good candidate for detecting cracks in tread area. How to effectively simplify and apply Hough transform is the key in practice.

As is known, Hough transform can detect specific geometric shape effectively in an image. Based on the duality between the point and the corresponding geometric shape, Hough transform converts the problem of detection of a specific geometric shape into the problem of peak search in parameter space by representing the given curves in the original space with a point in parameter space [14]. For example, the following parametric form of Hough transform can be used to detect lines,

$$\rho = x \cos\theta + y \sin\theta \tag{1}$$

The points in an image can be mapped to the parameter space ($\rho - \theta$) by using the above formula.

In Hough transform, for curves of some specific geometric shape, a point in the original coordinates can be mapped to a specific curve in the parameter coordinates, meanwhile, a specific curve in the parameter coordinates can also be mapped to a point in the original coordinates. In a sense, Hough transform is equivalent to the search of the identical or similar geometric shape in a complex image by using the same specific geometric template. From the mathematical point of view, this is a special curvilinear integral, only when the geometric shape is identical or similar to the specific given geometric template, can we obtain the greatest integral value.

Therefore we will detect cracks by using the idea of Hough transform, not merely by using the specific steps directly given in the literatures. The preceding analysis indicates the crack is generally in a line pattern, so the maximum (in the case of open ply splices) or the minimum (in the case of overlapped cord) could be obtained if the integral is along the crack when the direction of crack is different from the direction of normal texture.

To our delight, through a large number of observations, we find that most of cracks are different from the actual direction of normal texture stripe. Then we propose an effective and simple method to detect cracks in tread area based on Hough transform: for an image of tread area containing cracks, we can obtain an integral curve if the integral is performed along one direction; and when a region of the maximum value or the minimum value appears in the integration curve, we may find a candidate region of a crack.

In the current paper, the regions of the maximum value or the minimum value are called the singular region. However, we shall be aware of such a fact that the direction of a crack is not fixed even though it is generally different from the actual directions of normal texture stripe. So in order not to miss cracks in all possible directions, we adopt a full range of integral that the image of tread area is projected to all possible directions, and then a family of integral curves can be obtained.

In actual computation, for the sake of convenience, we carry out the omni-directional projection every 2 degree. For each image window, a total of 90 projection curves can be obtained (in the range of 180 degree). In this way, the analysis based on Hough transform becomes the singularity analysis of projection histogram curves. Experiments showed this approach is effective.

3 Omni-Directional Projection Histogram of Tread Area

3.1 Omni-Directional Projection of Tread Area

According to the traditional projection algorithm, we need to rotate images in a certain degree and then project the rotated image. Since the actual obtained tyre images are usually very large, the traditional projection algorithm will consume too much time and will then lead to lower detection efficiency. So we make use of a special projection method, i.e., project the image along a series of given angles (see Fig. 3).

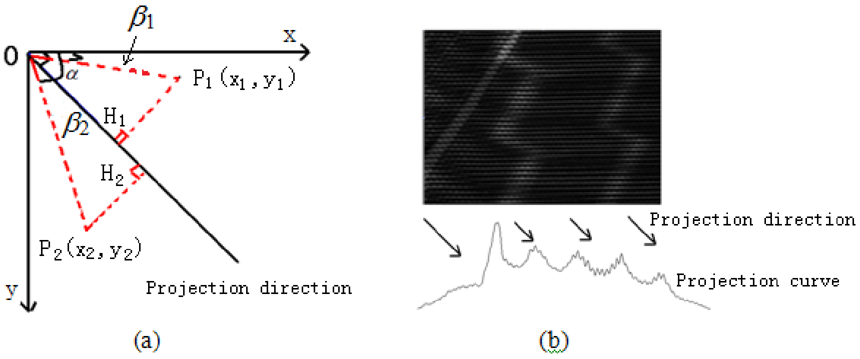


Fig. 3. The method of obtaining projection curve. (*a*) Projection principle; (*b*) An obtained projection curve.

In Fig. 3, Oxy is the image coordinate system, the projection line is shown in the Fig. 3(a). P_1 and P_2 are two pixels in the image. The angle between projection line and abscissa (X-axis) is α, the angles between P_1 , P_2 and abscissa are β_1, β_2, respectively. Draw two perpendicular lines from P_1 and P_2, the feet are H_1 and H_2, respectively. We aim to compute OH_1 and OH_2 so as to obtain the projection curve. The following formula can be used to compute OH.

$$|OH_1|=|OP_1|\cos(\alpha-\beta_1)= |OP_1|(\cos\alpha \cos\beta_1+ \sin\alpha \sin\beta_1)$$

Since $\cos\beta_1 = x_1/|OP_1|$, $\sin\beta_1 = y_1/|OP_1|$, So $|OH_1| = x_1\cos\alpha + y_1\sin\alpha$, then $|OH_2| = |OP_2|\cos(\beta_2 - \alpha) = x_2\cos\alpha + y_2\sin\alpha$.

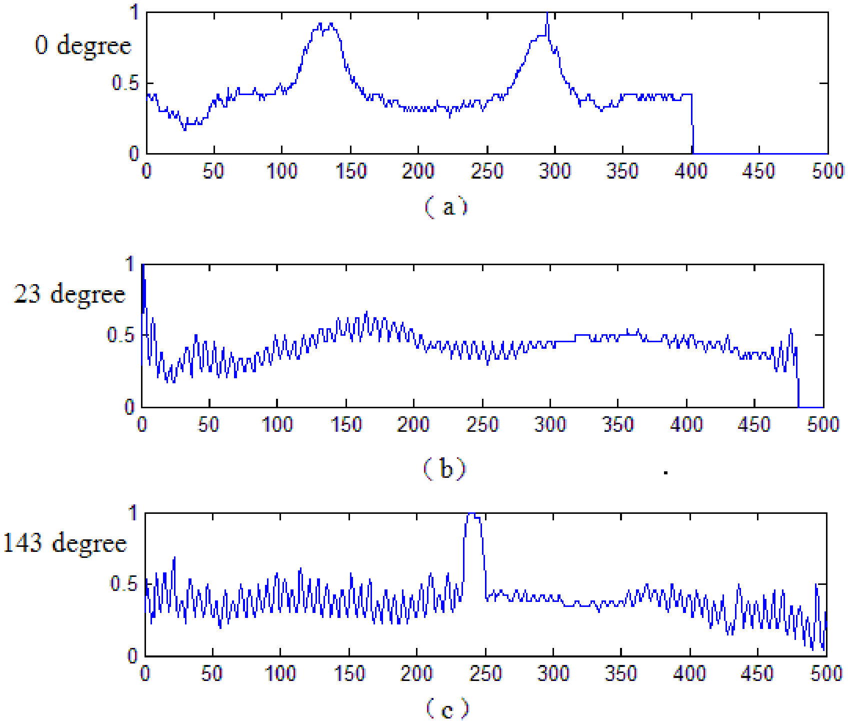
It is easy to see that the projection of each pixel is independent of the angle between the current pixel and the abscissa, and only depends on the projection direction. As of a pixel $P(x, y)$, the projection can be computed by

$$OH = x\cos\alpha + y\sin\alpha \qquad (2)$$

3.2 Analysis of Projection Curves

We perform omni-directional projection for images of tread area sequentially from 1 degree to 180 degree, and then we can obtain a series of projection curves.

If the minimum projection interval angle is a, the error for crack angle will be $a/2$. We choose 2 degree as the minimum projection angle, so the error is 1 degree. According to (2), we sequentially project the images of tread area, and then we can obtain 90 projection curves. After thorough observation and analysis, it is easy to see that there are three types of projection curves (see Fig. 4),



(a)

(b)

(c)

Fig. 4. Three types of projection curves for interface open, which is obtained from three different projection angles

From Fig. 4 (a) and (c), it can be easily seen that there are obvious protrusions, which means that the data of protrusions is much higher than the neighborhood data. In the current paper, we call these protrusions as singular area, and the projection curves containing singular areas as singular curves. From Fig. 4 (b), the curve is relatively flat with no singular area, and it shall be pointed out that most of the projection curves are of this type. Generally, it is impossible that the curves like Fig. 4 (c) can be found from the projection of tread area images in normal tyre without any defects. In the normal case, the projection curves can be classified into the following types for the normal tyre images of tread area,

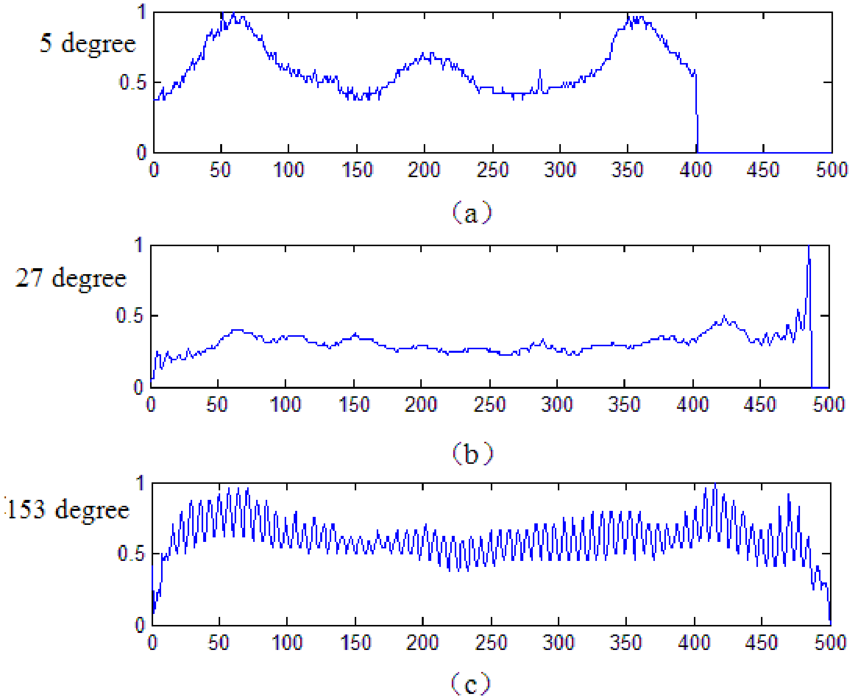


Fig. 5. The projection curves for tread area image in normal tyre

From Fig. 5, we can see that the projection curves of normal tyre image are quieter and have no significant volatility in local area except the starting position and ending position, because the peaks and valleys at the starting position and ending position are caused by image boundaries. Therefore whether there are singular areas in projection curves can be taken as a judgment for the determination of cracks in tread area, since we can seldom see any projection curves similar to Fig. 4 (c) in normal tyre image. As of the projection curves similar to Fig. 4 (a), we will point out in Sec. 4 that it is hard to detect cracks from this kind of curves because of the texture regularity.

In the current paper, the detection of cracks in tread area is consistent with the detection of singular area in projection curves.

4 Singular Area Analysis Based on Wavelet Theory

4.1 Multi-scale Decomposition of Projection Curves Based on Wavelet Transform

Wavelet transform is a powerful tool for data analysis and a common method for feature extraction. In wavelet transform, multi-scale decomposition plays an important role in signal analysis, which will be employed to analyze the projection curves of tread area image of tyre.

In the scale space, the multi-scale decomposition of a signal can be considered to be a kind of smoothing in different scales. In the process of decomposition, the fluctuations in different scales can be extracted successively. Thus, we can obtain the profile signal and the corresponding detail in different scales. With the increase of scale, the profile signal becomes clearer and clearer, which will bring convenience in the detection of the local prominent fluctuations of signal.

Suppose $f(t)$ is a one dimensional signal with limited energy, *i.e.*, $f(t)\in L^2(R)$. The wavelet transform of $f(t)$ is,

$$W_f(a,\tau) =< f(t), \varphi_{a,\tau}(t) >= \frac{1}{|a|}\int_{\infty}^{\infty} f(t)\varphi_{a,\tau}(\frac{t-\tau}{a})dt \tag{3}$$

where a and τ represent the scaling factor and the translational factor, respectively, a, $\tau \in R$, and $a\neq 0$. $\varphi(t)$ is called the mother wavelet from which a series of wavelet functions $\varphi_{a,\tau}(t)$ can be generated by expanding and contracting the signal in scale and translating the signal in time, i.e., the wavelet basis,

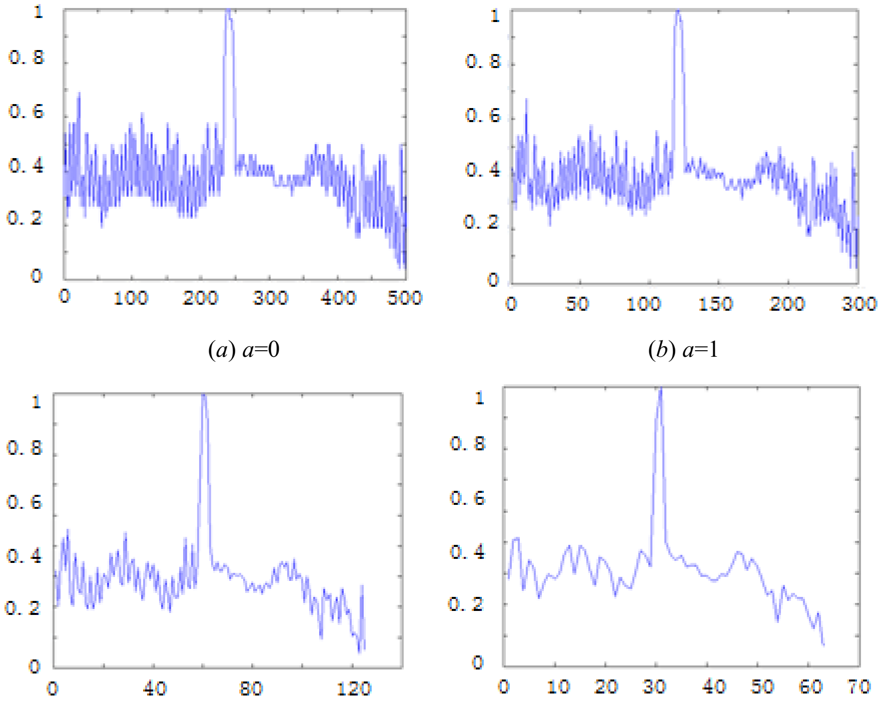
$$\varphi_{a,\tau}(t) = \frac{1}{|a|}\varphi(\frac{t-\tau}{a}) \tag{4}$$

With the change of translational factor, wavelet window moves along the time axis; with the change of scaling factor, the signal can be analyzed in different scales.

In a sense, the wavelet transform of a signal is equivalent to filtering a signal by using a series of filters. For each scale, a profile signal and the corresponding detail signal can be obtained simultaneously, and the former corresponds to the signal of low frequency and can reflect the local fluctuations easily. So we detect singular area of projection curves by using the profile signals in different scales. We employ discrete wavelet transform because the actual signals are in discrete form.

The multi-scale decompositions of Fig. 2(a-1)'s projection curve 4(c) are shown as follows :

We can easily see from Fig. 6 that the singular area and the local instability becomes more and more obvious with the increasing scaling factor, and it is the instability and abnormal change in local area that makes up the basis of the detection of defects in tread area.



(*a*) *a*=0

(*b*) *a*=1

(*c*) *a*=2

(*d*) *a*=3

Fig. 6. The multi-scale decompositions of a projection curve

4.2 Location of Curve Peaks and Valleys

For projection curves, in order to analyze features and detect abnormal local singular areas correctly, the key is to determine the location of peaks and valleys of projection curves in different scales.

In this paper, we propose an improved neighborhood comparison algorithm to determine peaks and valleys. The traditional algorithm determines the maximum or the minimum of local area by means of the comparison of neighborhood data, which may cause non-extreme points near the peaks and valleys (see Fig. 7, the non-extreme points are marked in red circles),

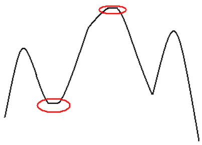


Fig. 7. Non-extreme points

In the case of non-extreme points, the following method is designed to determine peaks and valleys:

$$\mathrm{X}_{\max}=\max\,(X_{i-r}, X_{i-r-1}, \ldots, X_i, \ldots, X_{i+r-1}, X_{i+r}) \tag{5}$$

$$\mathrm{X}_{\max}=\min\,(X_{i-r}, X_{i-r-1}, \ldots, X_i, \ldots, X_{i+r-1}, X_{i+r}) \tag{6}$$

The above two formulae are used to determine whether $X_{\max}$ or $X_{\min}$ is the extreme value in the local area centered at X_i, r is the neighborhood radius of the local area. If the current obtained the extreme value equals to the previous one, the current one shall be regarded as a new extreme point; otherwise, it shall be discarded. Experimental results show that this method can greatly reduce the non-extreme points and retain unrepeated extreme points.

4.3 Acquisition of Local Energy Curves

From the preceding discussion, we know the determination of cracks in tread area is consistent with the determination of singular area in projection curves, so how to sharpen the characteristics of singular area is the key of the defects defection.

After the location of peaks and valleys of curve, how to describe the local singularity becomes a key to detect defects. In this paper, peak density is defined as the number of peaks (valleys) per unit distance. High peak density means fast change and small interval of peaks (valleys), so a narrow window is enough to contain all the local changing data according to the relationship between the scaling factor of wavelet transform and the time-domain analysis window. On the contrary, low peak density means slow change and large interval of peaks (valleys), so a wide window is necessary to contain all the local changing data. Thus, peak density can well measure the fluctuation of local data and can be used to determine width of the analysis window of local data (see Fig. 8).

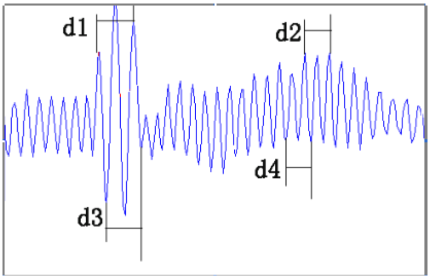


Fig. 8. Analysis window of peak density

As is shown in Fig. 8, we take the distance of three adjacent peaks (valleys) as a measure, obviously, $d_1>d_2$, $d_3>d_4$. The curve marked by d_1 fluctuates slowly and has a lower peak density, so we choose a wide window; meanwhile the curve marked by d_2 fluctuates fast and has a higher peak density, so we choose a narrow window. The relationship between the window and the peak density can be described as follows,

$$a = \frac{n}{\rho} \tag{7}$$

Where a is the analysis window, ρ is the peak density, n is the number of chosen adjacent peaks (valleys).

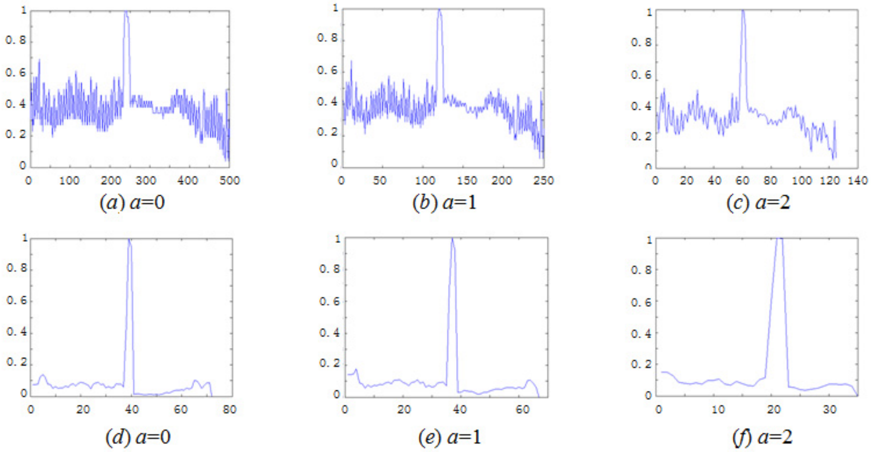
It can be easily seen from the above analysis that the width of local analysis window changes with the peak density in real time: when the peak density increases, the window becomes narrow, whereas when the peak density decreases, the window wide. Now the question is how to reflect the local instability. Since the second moment can describe fluctuations of local data around its mean and can especially reflect the characteristics of singular area, *e.g.*, the instability of local data, thus the second moment is taken to measure the changes of local data.

For one-dimensional discrete signal, the second moment can be computed as follows (suppose there are K elements in the analysis window),

$$\bar{X} = \sum_{i=0}^{K} X_i \tag{8}$$

$$E = \frac{1}{K}\sum_{i=0}^{K}(X_i - \bar{X})^2 \tag{9}$$

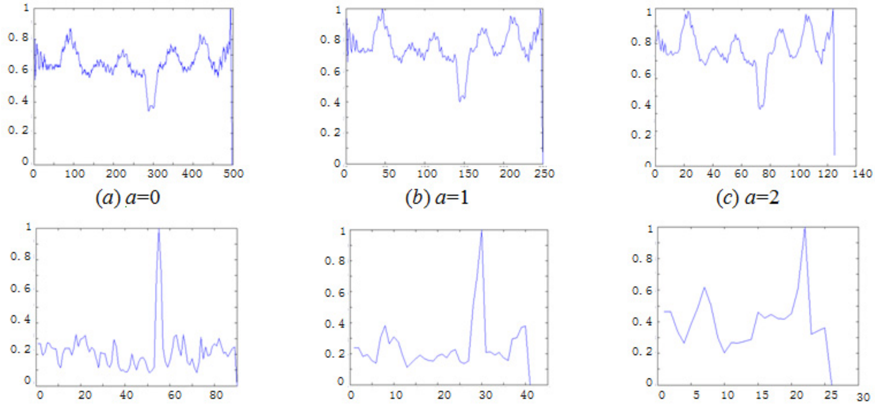
From (9), it is easy to see that the second moment is equivalent to the local energy, so we regard (9) as a local energy definition. Let n=3 in (7), then the local energy curves in different scales of image 2(a-1) and 2(b-1) can be computed by using the (8) and (9) (see Fig. 9 and Fig. 10),



(a) a=0 (b) a=1 (c) a=2

(d) a=0 (e) a=1 (f) a=2

Fig. 9. Projection curves in different scaling factors (a)-(c) and local energy curves in different scaling factors (d)-(f)



(*a*) *a*=0 (*b*) *a*=1 (*c*) *a*=2

(*d*) *a*=0 (*e*) *a*=1 (*f*) *a*=2

Fig. 10. Projection curves in different scaling factors (*a*)-(*c*) and local energy curves in different scaling factors (*d*)-(*f*)

From the above figures, apparent changes can be easily found in different scales at some specific locations in the local energy curves, *i.e.*, the energy at these locations is much higher than the other locations. This indicates that there exist singular areas at these locations.

Fig. 11 shows projection curves and local energy curves of a normal image in scaling factor *a*=0 and 1. From Fig. 11, we can see that these curves change stably and smoothly and there is no intense fluctuation (except the starting position and ending position, the reason is that the peaks and valleys at the starting position and ending position are caused by image boundaries). So we can draw the conclusion that the proposed method in this paper can be used to determine the singular area in projection curves and then further to detect cracks in tread area.

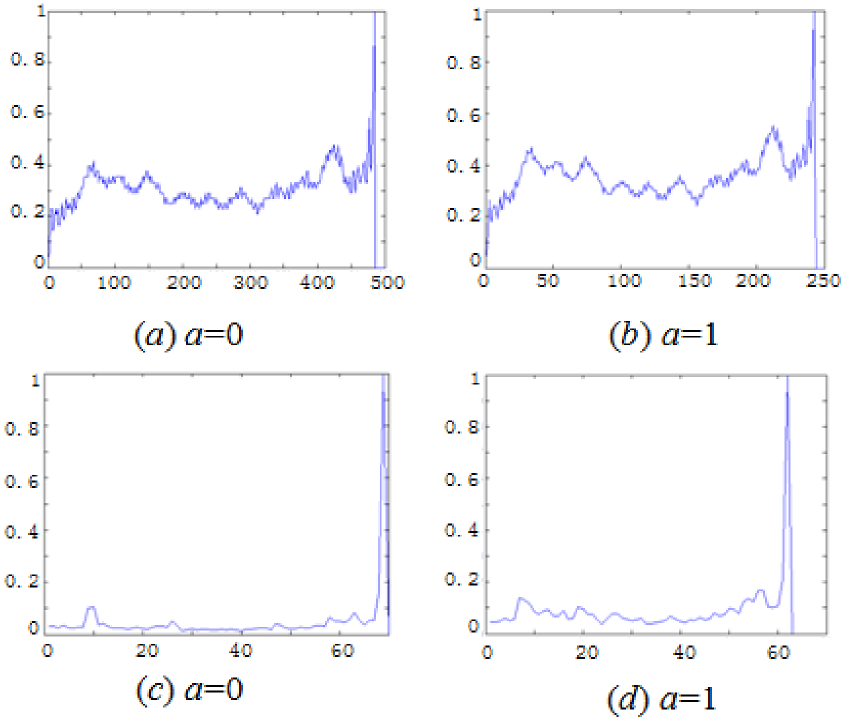
4.4 Singular Area Location in Local Energy Curve

In order to further determine singular area, we make use of a synthetic energy curve by combining the energy curves in different scales (from 0 to 2) according to (10),

$$y_i = \frac{y_{0i} + y_{1i} + y_{2i}}{3} \tag{10}$$

where y_{0i}, y_{1i} and y_{2i} represents the energy curves with length normalized in different scales, and y_{i} represents the synthetic energy curve. Then the presence of singular area on the synthetic energy curve can be confirmed according to the energy information in multiple scales.

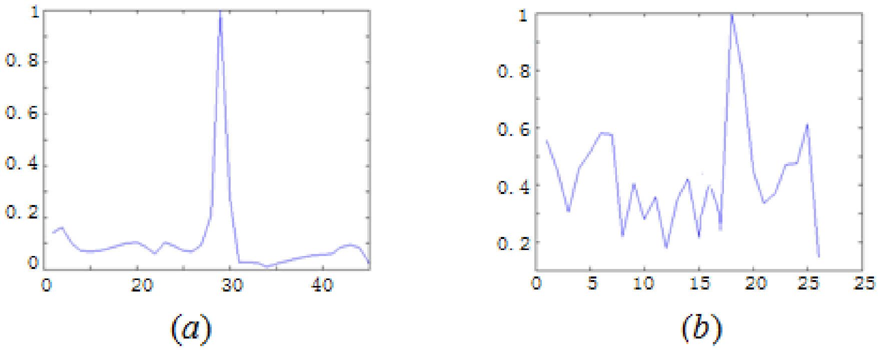
The synthetic energy curves of Fig. 2(*a*-1) and Fig. 2(*b*-1) are shown in Fig. 12.



(*a*) *a*=0 (*b*) *a*=1

(*c*) *a*=0 (*d*) *a*=1

Fig. 11. Projection curves in different scaling factors (*a*)-(*b*) and local energy curves in different scaling factors (*c*)-(*d*)



(*a*) (*b*)

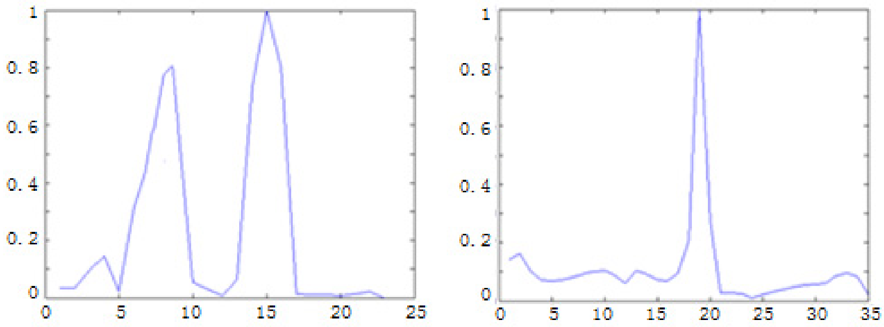
Fig. 12. The synthetic energy curves of Fig. 2(*a*) and Fig. 1(*b*)

From the synthetic energy curves, we find that the difference was very obvious between the maximum and the second maximum; the rest areas in the curves change stably and have no intense fluctuations. From the synthetic energy curve, it is easy to see that the area with the maximum value corresponds to the so-called "singular" area.

In this paper, the absolute difference between the maximum and the second maximum value in the synthetic curve is taken as a standard for determining a singular area, and the threshold is set to be 0.4, *i.e.*, if the difference is greater than 0.4, the area corresponding to the maximum energy is taken as a singular area, otherwise, not.

5 Crack Detection Based on Texture Regularity

After the analysis of Omni-directional projection curves of tread area, we find that the curve with singular area can be detected by using the multi-scale decomposition of wavelet transform and the computing of local energy curves when the corresponding image contains crack, the orientation and the location of crack can be then easily extracted from the singular curve. The detected singular curves can be divided into two types (see Fig. 13),



(*a*) Many peaks (*b*) One peak

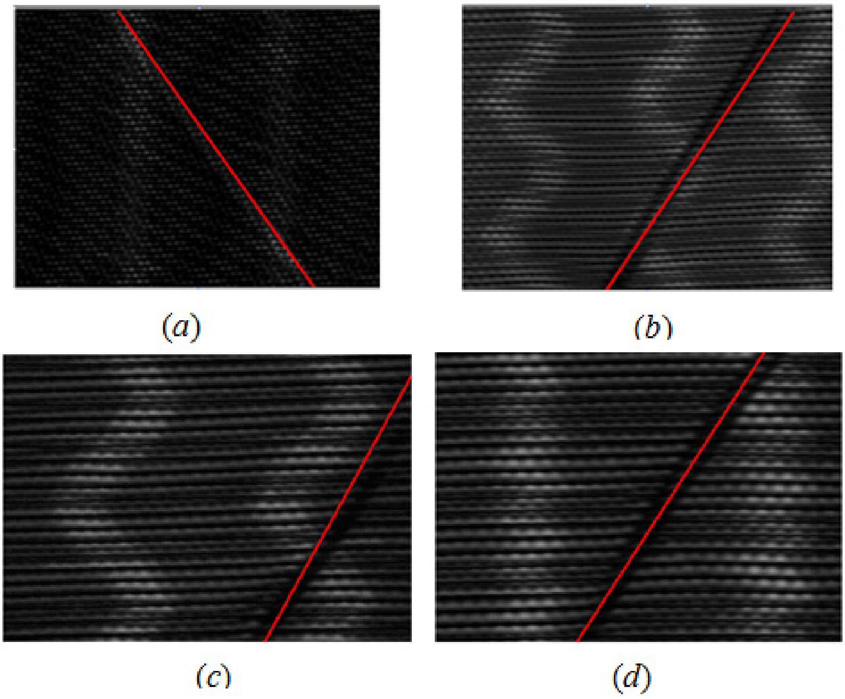
Fig. 13. Two types of projection curves containing singular area

In Fig. 13, we can see that there are two peaks in (a) and there is only one peak in (b). From (a), we can easily deduce that there are at least two crack-like singular areas along the projection angle, and from (b), there is only one singular area along the projection angle. Generally, we can also see that the height and the width of the peaks in (a) are different from that in (b), so we use the following formula to describe the difference:

$$\sigma = h/w \tag{11}$$

where h and w represent the height and width of peak, respectively. Since there are two peaks in Fig. 13 (a), we can obtain $\sigma_{a1}=h_{a1}/w_{a1}$ and $\sigma_{a2}=h_{a2}/w_{a2}$; there is only one peak in Fig. 13 (b), we obtain $\sigma_b=h_b/w_b$. Obviously, $\sigma_{a1}<\sigma_b$, $\sigma_{a2}<\sigma_b$ and $w_\mathrm{a}> w_\mathrm{b}$. It is easy to see that the width of singular area represent the width of crack and number of peaks represent the number of cracks. *For actual tyre image, however, the probability that there are more than two parallel cracks is rather low*. Generally, the width of a crack is within 1cm~2cm (of course this depends on the actual resolution of images). In order to get good performance, we set the threshold σ=5 just by experience, i.e., only the peak with σ≥5, we take it as a real singular crack, otherwise, not. Therefore, the singular areas shown in Fig. 13 (*a*) are in general not caused by cracks but by the texture of images, *e.g.*, in Fig. 14 (a), two vertical white texture patterns may lead to the curve pattern corresponding to Fig. 13 (a). As a matter of fact, this method to detect real singular crack makes use of texture regularity of a tyre, *i.e.*, we suppose the probability that there are more than two parallel cracks is rather low.

Projection curves with singular areas caused by image texture can be excluded by using the *characteristic of texture regularity*, and only singular curves caused by real cracks will be retained. By the projection angle and the range of singular area in projection energy curves, we can easily find the orientation and location of a crack in the corresponding image. So cracks can be marked in tread area image with fairly high accuracy (see Fig. 14).

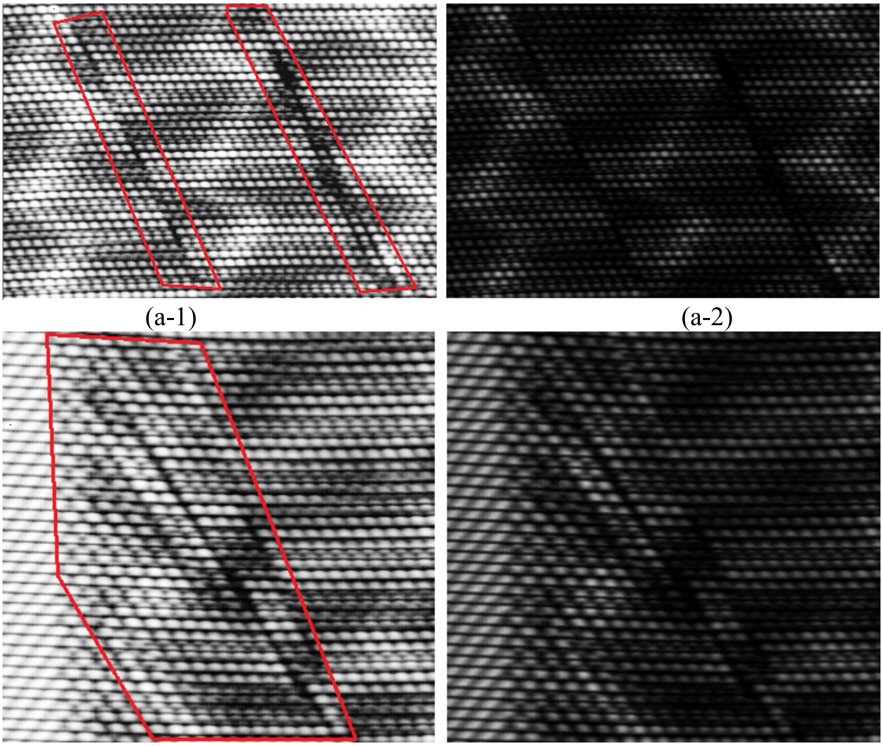

(*a*) (*b*)

(*c*) (*d*)

Fig. 14. Result of crack detection in four images of tread area.

At present we cannot find any image dataset of tread area for defects detection, and all the images used for defects detection are provided by Shandong Linglong Tire Co., Ltd. More than 800 images are tested in our experiment, and the results showed this method can detect cracks of tread area at an accuracy of 90%.

The following cracks are the two main failure cases (Fig. 15), the first one is parallel cracks (see a-1 and a-2), the second one is divergence (see b-1 and b-2). The first case does not obey our assumption, the second may display weak peaks. These will be studied in our future research.



(a-1) (a-2)

(b-1) (b-2)

Fig. 15. Two main failure cases.

6 Conclusions

We propose a method for the detection of cracks in tread area by analyzing singular area based on Hough transform and multi-scale decomposition of wavelet transform. Firstly, we make full use of the abnormal distribution characteristics of grayscale at cracks in tread area image, and obtain a series of projection curves based on Hough transform and Omni-directional projection, and then manage to convert the detection of abnormal distribution of grayscale into the detection of singular area.

Secondly, we analyze the projection curves in different scales by means of wavelet transform. The key point is to compute the energy curves in different scales after the location of peaks and valleys in those curves and the automatic adjustment of window width according to the peaks density dynamically. The abnormal fluctuations of the energy curves in different scales are used to locate the singular area.

Finally, we exclude the projection curves caused by normal regular texture, even though these curves contain singular areas. Only curves caused by real cracks are detected for the determination of orientation and location of cracks. The experimental results demonstrated the proposed approach is effective to detect cracks in tread area. However, this method needs the operation of image omni-directional projection, which may lead to high computational complexity. So it is very important to reduce complexity and improve detection efficiency in the future.

References

1. Xia, F.: Tire defects inspection based on digital image processing. Master's dissertation of Shandong University, China, May 2011
2. Qilin, Z., Lei, X., Zhiqiang, Z., Weibin, H., Tiejun, Ma.: Lossless tyre detector based on laser speckle. Tyre Industry (in Chinese) **28**(2), 113–116 (2008)
3. Maoqiang, Z.: Research on detection method and system design of radial tire's defects. Master's dissertation of Shandong University, China (2014)
4. Yan, Z.: Research on nondestructive tire defect detection using computer vision methods. Ph.D. dissertation of Qingdao University of Science and Technology, China (2014)
5. Zhanhua, H., Zheng, L., Meng, Z., Huaiyu, C., Yinxin, Z.: Defects on-line detection of tire textures based on statistical features. Optical Technique (in Chinese) **35**(1), 60–62 (2009)
6. Yue, Z., Wenyao, L., Ye, Y., Fangchao, L., Jinjiang, W.: A defect extraction and segmentation method for radial tire X-ray image. Journal of Optoelectronics Laser (in Chinese) **21**(5), 758–761 (2010)
7. Zheng, L.: Tyre X-ray inspection and multi-textural image analysis technology. Ph.D. dissertation of Tianjin University, China, December 2008
8. Chuanhai, Z., Xuemei, L., Guo Qiang, Yu., Xichen, Y., Caiming, Z.: Texture-Invariant Detection Method for Tire Crack. Journal of Computer-Aided Design & Computer Graphics **06**, 809–816 (2013)
9. Wu Xiuyong, X., Jinwu, K.X.: Automatic recognition method of surface defects based on Gabor wavelet and kernel locality preserving projections. Acta Automatica Sinica (in Chinese) **36**(3), 438–441 (2010)
10. Jing, S., Xuezhi, Y.: A new method for the fabric defect defection based on texture watershed. Journal of Image and Graphics (in Chinese) **14**(10), 1997–2003 (2009)
11. Kumar, A.: Computer Vision-based Fabric Defect Detection: A Survey. IEEE Transactions on Industrial Electronics **55**(1), 348–363 (2008)
12. Mak, K.L., Peng, P., Yiu, K.F.C.: Fabric defect detection using morphological filters. Image and Vision Computing **27**, 1585–1592 (2009)
13. Ngan, H.Y.T., Pang, G.K.H., Yung, N.H.C.: Automated fabric defect detection—A review. Image and Vision Computing **29**, 442–458 (2010)
14. Gonzalez, R.C., et al.: Digital Image Processing (2nd Version). Publishing House of Electronic Industry (China) (2002)

Multiple Scaled Person Re-Identification Framework for HD Video Surveillance Application

Hua Yang[1,2](✉), Xinyu Wang[1,2], Wenqi Ma[1,2], Hang Su[1,2], and Ji Zhu[1,2]

[1] Institue of Image Communication and Network Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
[2] Shanghai Key Lab of Digital Media Processing and Transmission,
Shanghai 200240, China
{hyang,xinyuwang}@sjtu.edu.cn, mawenqi@sjtu.org,
{suhangss,jizhu1023}@gmail.com

Abstract. Person re-identification is an important problem in automated video surveillance. It remains challenging in terms of extraction of reliable and distinctive features, and matching of the features under different camera views. In this paper, we propose a novel re-identification strategy for person re-identification based on multiple image scaled framework. Specifically, global features and local features are extracted separately in different image scales. These two-scaled processing are constructed in a cascaded system. We use semi-supervised SVM to obtain a similarity function for global features and a similarity function combining the spatial constraint and salience weight for local features. Experiments are conducted on two datasets: ETHZ and our dataset with high resolution. Experimental results demonstrate that the proposed method outperforms the conventional method in terms of both accuracy and efficiency.

Keywords: Person re-identification · Multiple scaled framework · Distance metrics

1 Introduction

Person re-identification across different views of cameras is a fundamental task in automated video surveillance.Despite best efforts have been made in computer vision area in the past years, person re-identification problem remains largely unsolved. This is due to a number of reasons. First, the resolution of the current monitored cameras is not high enough so that person verification relying upon biometrics is infeasible and unreliable. Second, as the transition time between disjoint cameras varies greatly from individual to individual with uncertainty, it is hard to impose accurate temporal and spatial constraints. Third, the visual appearance features, which are extracted mainly from the clothing and shapes of people, are intrinsically indistinctive for matching people.

To solve the re-identification problem, discriminative and reliable signature for the person is needed. The image can be described by color[1], shape[2, 3],

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 219–228, 2015.
DOI: 10.1007/978-3-662-48570-5_21

texture[3, 4, 5, 6], Haar-like representations[7], edges[3], interest points[8, 9, 10] and image patches[4]. Since a single type of features is not powerful enough to capture the subtle differences of all pairs of objects, multiple features are combined here to make the person signatures more discriminative and reliable. Bazzani et al.[1] and Cheng et al. [11]combined MSCR descriptors with weighted Color Histograms, achieving state-of-the-art results on several widely used person re-identification datasets.There are also some other research works on person re-identification have been done to learn reliable and effective mid-level features. Li et al. [12] proposed a deep learning framework to learn filter pairs, which encode photometric transforms across camera views for person re-identification. Zhao et al. [13] proposed a method to automatically learn discriminative mid-level features without annotation of human attributes.

Conventional methods attend to seek the discriminative and reliable signature, after feature extraction, these methods simply choose a standard distance measure such as L1 norm and L2 norm. However, under severe changes in viewing conditions, extracting a set of features that are both distinctive and reliable is extremely hard. Moreover, given that certain features could be more reliable than others under a certain condition, applying a standard distance measure is undesirable as it essentially treats all features equally without discarding bad features selectively in certain matching circumstances. To overcome these shortcomings, recent years researchers focus on the distance metric of person re-identification. That is, given a set of features of each person image, they seek to quantify and differentiate these features by learning the optimal distance measure that is most likely to give correct matches. Gray et al. [4] combined spatial and color information in an ensemble of local features by boosting. Prosser et al. [5] formulated person re-identification as a ranking problem, and used ensembled RankSVMs to learn pairwise similarity.Zheng et al. [14] formulate person re-identification as a relative distance comparison learning problem in order to learn the optimal similarity measure between a pair of person images.

This paper focuses on effectively using the benefits of high resolution images and reducing the complexity in the foundation of improving the accuracy of re-identification. To perform re-identification under the HD monitor cameras, we propose a re-identification framework, in which global features and local features are extracted respectively in different image scales based on their scale behaviours. Specifically, the global features are represented by the histogram whose matching performance is not related to the image scale directly while the local features perform better on the higher image scale since they need more feature details to match. It is worth mentioning that our approach is performed under special application scenarios, that is the scenarios monitored by HD cameras. So we do not compare with the most advanced methods.

The contributions of this framework can be summarized in three-folds: First, we propose a novel multiple scaled framework in which different features can be extracted in proper image scales on their behaviors, so that the benefits of HD monitor cameras can be exploited. Second, we use a cascaded system to improve the efficiency of the system. Third, we use a new matching algorithm based on feature points, adding the color feature into the descriptor and combing the spatial constraint and salience weight, which improves the accuracy of person re-identification.

The rest of the paper is organized as follows: Section 2 describes the details of the proposed framework which we refer to as Multiple Scale Re-identification Framework (MSRF). Section 3 illustrates and analyzes the experimental results. Finally, the main conclusions are summarized in Section 4.

2 Multiple Scale Re-identification Framework

Figure 1 shows a re-identification system under the HD monitor circumstances. There are four steps in our framework. First, the global features are extracted on low image scale. Second, we use semi-supervised SVM to get a match result and choose the top $k\%$ as the filtered candidate. Third, local features are extracted on high image scale. Forth, a local feature points based algorithm is used to get another match result. Last,the match results obtained in the two-scaled processing are added together to get the final ranking. In our experiments, the low image scale is obtained with the down sampling factor 2 while the large scale with the sampling factor 1.

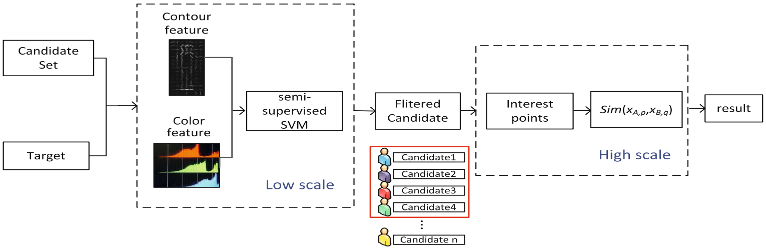


Fig. 1. System structure of Multiple Scale Re-identification

2.1 Re-identification on Low Image Scale

The appearance of objects is usually characterized in three aspects, color, contour and texture. Since a single type of feature is not powerful enough to capture the subtle differences of all pairs of objects, color and contour are combined here to make the person signatures more discriminative and reliable.

Color. Color histograms of the whole image region are widely used as global features to match objects across camera views because they are robust to the variations of poses and viewpoints. However, they also have the weakness that they are sensitive to the variations of lighting conditions and photometric settings of cameras and that their discriminative power is not high enough to distinguish a large number of objects. Various color spaces such RGB, Lab, HSV and Log-RGB have been investigated and compared in [15]. Removing the lightness component in the HSV color space can reduce the color variation across camera views and we use this method to obtain color feature in our experiment.

Contour. Histogram of Oriented Gradients (HOG)[15]characterizes local shapes by capturing edges and gradient structures. It is robust to small translations and rotations of object parts.

In our experiments, the color feature and contour feature are concatenated to form a new representation.The dimension of the histogram in each color channel is 128 and the dimension of HOG histogram is 3528.

After the feature histograms have been extracted, the standard distance measure such as L1 norm could be applied. The distance learning method like RankSVM [5], RDC [14] could also be used according to the application circumstances. Here we use semi-supervised SVM to get a match result and choose the top $k\%$ as the filtered candidate.Here is a brief introduction of the principle of semi-supervised SVM. In order to exploit the unlabeled data, Bennett[16] created a method to classify the unlabeled data based on the original support vector machine. It is assumed that the unlabeled points are classified as Category 1, and the classification accuracy is calculated. Then, the points are classified as Category 2. Select the class that has the high classification accuracy.

The choice of k relates to the accuracy and complexity of the algorithm. If k is too small,the number of samples for the following processing is small,which will reduce the accuracy of the algorithm.If k is too large, the complexity will be increased. Considering the accuracy and complexity, we choose k=30 here.

2.2 Re-identification on High Image Scale

Local features perform better on high image scale since they need more feature details to match.Under the HD monitor circumstances, the image details could be exploited, which will benefit the matching performance based on the texture interest points. Traditional methods just extracted the texture feature of the interest points, which were less discriminative and reliable. Here this paper proposed an improved re-identification method based on the interest points, which we called Local Salience Feature (LSF) method.

There are four steps in the re-identification. First, the interest points will be extracted by the SURF operator. Second, in the center of each interest point, a patch is extracted. Then the color histogram will be extracted from the patch. Color histogram and contour histogram are concatenated to form a new representation. Third, the location of each point is considered to make the matching process more effective. Finally, the salience weight of each point is learned to make the re-identification more reliable.

Feature Extraction. In the center of each interest point, a patch will be extracted. A LAB color histogram is extracted from each patch. For the purpose of combination with other features, all the histograms are L2 normalized. To handle viewpoint and illumination changes, SURF descriptor[17]is used as a complementary feature to color histograms, which are also L2 normalized. In our experiments, the parameters of feature extraction are as follows: patches of size 10x10 pixels.128-bin color histograms are computed in L, A, B channels respectively. And in each channel, SURF features produces a 128 feature vector for each interest points. In a summary, each patch is finally represented by a discriminative descriptor vector with length 128x3+128 =512.

Spatial Constrain. The distance of pairwise person could be converted to compute the distance of pairwise interest point set. The greedy algorithm [18] will be applied in our experiment. For each point of the target, we will find the corresponding one which has the shortest distance with it in the candidate point set. For each point pair, in order to deal with the misalignment in the matching process, we also compute the distance of the locations of the features with the Euclidian distance. The final ground distance between two interest points is shown in Eq.1.

$$D\left(x_A, x_B\right) = FD\left(x_A, x_B\right) + \alpha \times ED\left(x_A, x_B\right) \tag{1}$$

Where α is a weighting parameter, FD is the distance of the feature vector, L1 norm is applied in our experiments, ED is the Euclidean distance, and x is the location of the centroid of the point. It is worth mentioning that several distance measures were considered, and experiments showed the effectiveness of the proposed combination of distances.

As suggested in [15], aggregating similarity scores is much more effective than minimizing accumulated distances, especially for those misaligned or background points which could generate very large distances during matching. By converting to similarity, their effect could be reduced. We convert distance value to similarity score with the Gaussian function:

$$s\left(x_A, x_B\right) = exp\left(-\frac{D\left(x_A, x_B\right)^2}{2}\right) \tag{2}$$

Salience Weight. Each interest point of a person has certain information, so different point has different identify power in the matching process. According to [19], KNN method could be applied to learning the salience weight of the interest points. We could get the following function:

$$X_{nn}\left(x^i_{A,p}\right) = \left\{ x | arg \max_{x^j_{B,q} \in \{B_q\}} s\left(x^i_{A,p}, x^j_{B,q}\right), q = 1, 2, ..., N_B \right\} \tag{3}$$

Where the interest point in target image is represented as $x^i_{A,p}$, where (A, p) denotes the p-th image in camera A and i denotes the point index in set. The interest point in candidate image is represented as $x^j_{B,q}$, where (B, q) denotes the q-th image in camera B and j denotes the point index. $\{B_q\}$ means the candidate set under camera B. s is the similarity score function in Eq.2. N_B is the candidate number.

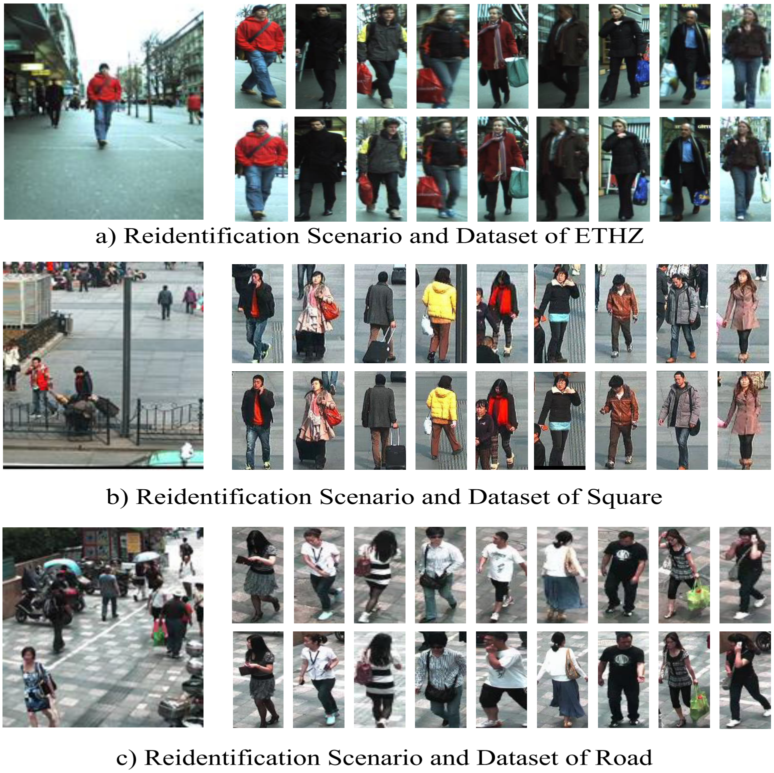
We apply a similar scheme in [19] for each test point, and the KNN distance is utilized to define the salience score:

$$w\left(x^i_{A,p}\right) = D\left(x^i_{A,p}, X_{nn}\left(x^i_{A,p}, k\right)\right), k = \frac{N_B}{2} \tag{4}$$

Where D denotes the distance of the k-th nearest neighbor. If the distribution of the reference set well reflects the test scenario, the interest point could only find limited number of visually similar neighbors. More details about the salience learning method could be found in [19][2].

Then we could get the similarity of two point sets by the following equation.

$$sim\left(x_{A,p}, x_{B,q}\right) = \sum_{i=1}^{|x_{A,p}|} w\left(x^i_{A,p}\right) \cdot s\left(x^i_{A,p}, x^j_{B,q}\right) \tag{5}$$



a) Reidentification Scenario and Dataset of ETHZ

b) Reidentification Scenario and Dataset of Square

c) Reidentification Scenario and Dataset of Road

Fig. 2. Example images of different datasets used in our evaluation.The first column denotes the scenario and the rest columns denote image pairs of the same person.**a)** ETHZ,**b)** our dataset on square,**c)** our dataset on road

3 Experimental Results

3.1 Experiment 1

In this experiment, we evaluated the accuracy of the proposed strategy. We evaluated our approach on the public ETHZ dataset. The results are shown in standard Cumulated Matching Characteristics (CMC) curve [19]. A rank r matching rate indicates the percentage of the p images with correct matches found in the top r ranks against the p gallery images. Rank 1 matching rate is thus the correct matching/recognition rate. Note that, in practice, although a high rank 1 matching rate is critical, the top r ranked matching rate with a small r value is also important because that the top matched images will normally be verified by a human operator. We also apply our method on two real HD monitor circumstances. Both ETHZ and real dataset are shown in Fig. 2.

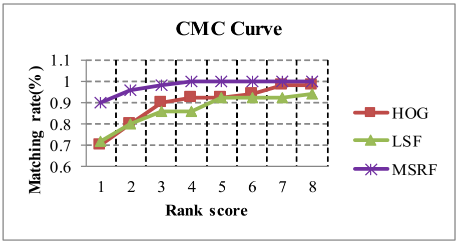


Fig. 3. Re-identification Result of ETHZ Dataset

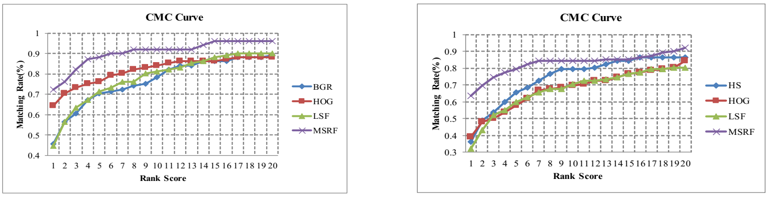
Table 1. Matching rate(%) of Different Methods on ETHZ Dataset

	ETHZ Dataset				
Methods	r=1	r=2	r=3	r=4	r=5
MSRF	90.00	96.00	98.00	100.00	100.00
BGR[15]	62.00	76.00	82.00	88.00	92.00
HS[15]	62.00	88.00	92.00	96.00	100.00
LAB[15]	66.00	86.00	92.00	98.00	100.00
HOG[18]	70.00	80.00	90.00	92.00	92.00
SIFT[18]	46.00	56.00	58.00	62.00	64.00
LSF	72.00	80.00	86.00	86.00	92.00

ETHZ Dataset. This dataset contains video sequences captured from moving cameras. It contains a large number of different people in uncontrolled conditions. With these video sequences, we get 50 pairwise people images for evaluation. All image samples are normalized to 128x64 pixels. Traditional methods like appearance-based methods[15]are compared here.

As shown in Fig. 3 and Table 1, our approach outperforms other methods based on single feature because that the MSRF method exploits the benefits of the HD images and multiple features are combined in our framework. The ETHZ is not a very challenging dataset,so we evaluate our method on two real monitor circumstances with different challenges.

Square and Road Dataset. The square datasets were captured from a railway station by two non-overlapping cameras. We collected 101 pairwise people images under each monitor circumstance for evaluation. All image samples are normalized to 128x64 pixels. Traditional methods like appearance-based methods[15]are compared here.



a) Scenario of Square

b) Scenario of Road

Fig. 4. Re-identification Result of Real Monitor Dataset

Table 2. Matching rate (%) of Different Methods on Different Datasets

	Square scenario				*Road scenario*			
Methods	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
MSRF	72.28	88.12	92.79	96.04	63.72	77.45	84.31	92.15
BGR[15]	45.54	70.30	78.22	89.11	22.55	47.06	66.67	78.41
HS[15]	44.55	63.37	74.26	83.17	36.27	59.80	79.41	86.27
LAB[15]	44.55	68.32	78.21	89.11	21.57	50.00	67.64	79.41
HOG[18]	64.36	76.24	84.16	88.12	39.21	53.92	69.61	84.31
SIFT[18]	55.45	65.35	73.27	82.18	24.72	36.49	50.21	71.78
SURF[18]	56.44	70.30	79.21	83.17	24.90	40.69	74.51	75.00
LSF	57.56	71.28	81.19	90.10	32.35	54.90	70.59	80.39

As shown in Fig. 4 and Table 2, our approach outperforms other methods based on single feature, especially when r is small. This is because the MSRF method exploits the benefits of the HD images and multiple features are combined in our framework. Images have a range of variations in human appearance and illumination under the real monitor circumstances, which made single feature less discriminative and reliable. In our framework, multiple features will have more identify power which benefited the re-identification result. Moreover, the local feature extracted on the high scale can get more feature details, which made the match result more reliable.

3.2 Experiment 2

In this experiment, we evaluated the high efficiency of the proposed strategy. Table 3 gives the cost of different algorithms. The hardware platform is Intel i7, 3.4GHz, 4GB RAM. Each algorithm is conducted on 101 pairwise people images and there are 10210 comparison operations in our experiments. It can be seen that the algorithms based on statistical characteristics have low complexity because the number of distance calculations is small. While the local feature algorithms need to conduct matching operation for each feature point. For there are M feature points extracted from one pedestrian image and N feature points from another pedestrian image, the computation cost is *O(MN)*. In the proposed strategy, firstly we use the statistical characteristics based algorithm to obtain the selected candidates. And then we use the local feature points based algorithm to recognise the selected candidates. These two steps can reduce the complexity.

Table 3. Time Cost Result of Different Methods

Methods	COLOR	CONTOUR	SURF	LSF	MSRF
Cost Time(ms)	14992	71480	842046	842311	307806

4 Conclusions

In this paper, we propose a new re-identification framework based on the multiple scaled framework to perform re-identification under the HD monitor cameras. Global features and local features are extracted separately in different image scales based on their scale behaviours. Specifically, the global features are represented by the histogram whose matching performance is not related to the image scale directly, while the local features perform better on the higher image scale since they need more feature details to match. In our framework, firstly we use the statistical characteristics based algorithm to obtain the selected candidates. And then we use the local feature points based algorithm to recognise the selected candidates. Experimental results demonstrate that the proposed method outperforms the conventional method in terms of re-identification accuracy and efficiency.

Acknowledgments. This research is supported by STCSM (No.10231204002, No.11231203102),NSFC (No.61171172) and the Innovation Research Group Project of National Natural Science Foundation of China(No.61221001).

References

1. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Comput. Vis. Image Underst. **117**(2), 130–144 (2013)

2. Oreifej, O., Mehran, R., Shah, M.: Human identity recognition in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2010)
3. Schwartz, W., Davis, L.: Learning discriminative appearance based models using partial least squares. In: Brazilian Symposium on Computer Graphics and Image Processing (2009)
4. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European Conference on Computer-Vision, pp. 262–275 (2008)
5. Prosser, B., Zheng, W., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference (2010)
6. Zhang, Y., Li, S.: Gabor-LBP based region covariance descriptor for person re-identification. In: International Conference on Image and Graphics, pp. 368–371 (2011)
7. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using haar-based and DCD-based signature. In: Proceedings of International Workshop on Activity Monitoring by Multi-camera Surveillance Systems (2010)
8. Gheissari, N., Sebastian, T., Tu, P., Rittscher, J., Hartley, R.: Person reidentification using spatiotemporal appearance. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition **2**, 1528–1535 (2006)
9. Kai, J., Bodensteiner, C., Arens, M.: Person re-identification in multi-camera networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio Workshops, pp. 55–61 (2011)
10. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: Proceedings of British Machine Vision Conference (2009)
11. Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of British Machine Vision Conference (2011)
12. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
13. Zhao, R., Ouyang, W., Wang, X.: Learning midlevel filters for person reidentification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
14. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 653–668 (2013)
15. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: Proceedings of the IEEE International Conference on Computer Vision (2007)
16. Bennett, K., Demiriz, A.: Semi-supervised support vector machines. In Advances in Neural Information Processing Systems 11 (1998)
17. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Proceedings of the European Conference on Computer Vision, pp. 404–417 (2006)
18. Doretto, G., Sebastian, T., Tu, P.H., Rittscher, J.: Appearance-based person reidentification in camera networks: problem overview and current approaches. J. Ambient Intell. HumanizedComput. **2**(2), 127–151 (2011)
19. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2013)
20. http://www.vision.ee.ethz.ch/aess/dataset/

Aesthetic Image Classification Based on Multiple Kernel Learning

Ningning Liu[1], Xin Jin[2(✉)], Hui Lin[1], and De Zhang[3]

[1] School of Information Technology and Management,
University of International Business and Economics,
Beijing 100029, People's Republic of China
ningning.liu@uibe.edu.cn, linhuivicky@foxmail.com
[2] Department of Computer Science and Technology,
Beijing Electronic Science and Technology Institute,
Beijing 100070, People's Republic of China
jinxinbesti@foxmail.com
[3] Department of Automation,
Beijing University of Civil Engineering and Architecture, Beijing 100044, China
zhangde@bucea.edu.cn

Abstract. Aesthetic image classification aims at predicting the aesthetic quality of photos automatically, *i.e.* whether the photo elicits a high or low level of affection in a majority of people. To solve the problem, one challenge is to build features specific to image aesthetic perceptions, and another one is to build effective learning models to bridge the "semantic gap" between the emotion related concepts and the extracted visual features. In this paper, we present an approach for aesthetic image classification based on Multiple Kernel Learning (MKL) method, which seeks for maximizing the classification performance without explicit feature selection steps. The experiments are conducted on a large diverse database built from online photo sharing website, and the results demonstrated the advantages of MKL in terms of feature selection, classification performance, and interpretation, for the aesthetic image classification task.

Keywords: Aesthetic quality · Image classification · Multiple kernel learning

1 Introduction

Aesthetics is a sub discipline of philosophy and axiology dealing with the nature of beauty, art, and taste. The assessment or prediction of aesthetic value in images is considered to be of subjectivity and universality. The subjective feature suggests that the judgment relies on individual personal feelings, and there is no single agreement on what it exactly belongs to. In contrast, the universality indicates that certain features in photographic images are believed to please humans more than others. In conclusion, though the evaluation of beauty and other aesthetic qualities of photographs is highly subjective, still they have certain stability and generality across different people and cultures as a universal validity to classify images in terms of aesthetic quality [2]. Figure 1 shows two photos from an online website, and according to the ratings by web users, it is confirmed that the photograph (b) can inspire higher aesthetic feelings than the left one (a) for most people.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 229–236, 2015.
DOI: 10.1007/978-3-662-48570-5_22

There could be many applications making use of an algorithm for photo quality assessment. For example, a search engine can merge a photo aesthetic factor into its ranking stage to get most relevant and better looking photos. An advertiser can make a choice referring to the most beautiful photos selected by the aesthetic quality assessment tools. Photo management solutions, like Picasa and iPhoto, can analyze the quality of one’s holiday snapshots and automatically present the best ones.

Research in the field of aesthetic image classification focuses on designing representation from various aspects, e.g., color, composition, lighting, and subjects. Recently, the impressive work made by R. Datta [2], Y. Ke [3] and M.Nishiyama et al. [4] have made a progress to this important issue. R. Datta [2] proposed 56 features based on the 'rules of thumb in photography'. Classification and linear regression on a community-based database showed that there is a significant correlation between various visual properties of photographs and their aesthetics ratings. Y. Ke [3] firstly proposed high level features based on a group of principles, including simplicity, realism and basic photographic technique, and the test provided a classification rate of 72% on a database. M. Nishiyama assess the aesthetic quality of a photo based on color harmony feature, namely ‘bags-of-color-patterns, and their results show that the performance of aesthetic image classification is improved by combining our color harmony feature with blur, edges, and saliency features. Meanwhile, there also other people [5, 6] simply employed the traditional low-level color, shape and texture features. Above works have designed various visual representations to characterize beauty in the form of photo art, but without considering the classifier or combination at all. For example, the authors in [2] use 5 cross-validation SVM accuracy score to rank and then select the top 15 descriptive features from the 56 proposed feature set, which requires explicit cross-validation steps for selecting features while optimizing the classifier parameters, and thus suffers from heavy computational complexities.

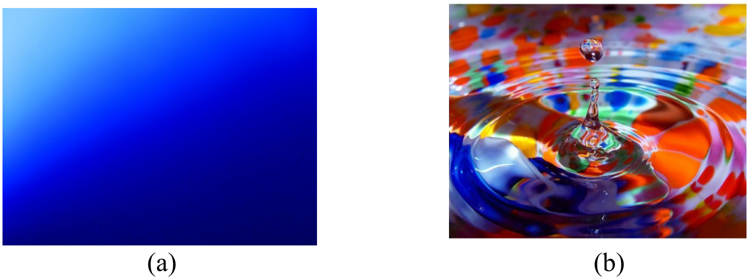

(a) (b)

Fig. 1. Example photos (a) and (b) received an average aesthetic rating of 2.4 from 222 votes and of 8.6 from 137 votes from a photo sharing website [6] respectively.

In this paper, we study the aesthetic image classification by applying multiple kernel framework, which can learns the feature representation weights and corresponding classifier in an intelligent way simultaneously. The main contributions of this paper included: (1) we investigate and implement visual features related to aesthetics, and also propose mid-level features to describe the dynamism and harmony in a photo; (2) we build a MKL scheme to perform aesthetic image classification, and received a good performance compared to the state-of-the-arts.

The rest of this paper is organized as follows: Section 2 introduces the image features used in this paper. Section 3 introduces our MKL framework for the aesthetic image classification. In Section 4, the experimental setup and results are reported. Finally, the conclusion and future work are presented in Section 5.

2 Image Features for Aesthetic Classification

Image feature extraction is a key issue for concept recognition in images. Features should be designed to carry sufficient information to be able to recognize the different concepts. In this paper, we complement low-level visual features based on color, texture and shape with higher level features such as color harmony and dynamism. Moreover, we make use of features based on aspects of a photograph appealing from a population and statistical standpoint [2], as well as representations based on perceptual factors that distinguish between professional photos and snapshots [3], and the aesthetic features based on color harmony [4]. The list of the features is given in Table 1.

2.1 Color, Texture and Shape

Studies have shown that the HSV (Hue, Saturation, and Value) color space is more related to human color perception than others such as traditional RGB. Moreover, different colors have different emotional meanings. Indeed, red is associated with happiness, dynamism and power whereas its opposite color, green, is associated with calmness and relaxation [7]. In this paper, different methods based on HSV color space are employed to describe color contents in images such as moments of color, color histograms.

The spatial gray-level difference statistics, known as co-occurrence matrix, can describe the brightness relationship of pixels within neighborhoods, and the local binary pattern (LBP) descriptor is a powerful feature for image texture classification. In this paper, these texture features are employed to contribute to aesthetic quality assessment.

Studies on artistic paintings have brought to the fore semantic meanings of shape and lines, and it is believed that shapes in a picture also influence the degree of aesthetic beauty perceived by humans [7]. Therefore, we make use of the Hough transform to build a histogram of line orientations in 12 different orientations.

2.2 Mid-Level

According to Itten's color theory [7], color combinations can produce effects such as harmony, non-harmony, calmness and excitation. Indeed, visual harmony can be obtained by combining hues and saturations so that an effect of stability on human eye can be produced. Itten has proposed to organize colors into a chromatic sphere where contrasting colors have opposite coordinates according to the center of the sphere. In case of harmony, color positions on Itten sphere are connected thanks to regular polygons. Therefore, by projecting the dominant image colors into the sphere and by comparing the distance between the polygon center and the sphere center, a value characterizing the image harmony can be obtained. At last, we extract the harmony features in 11 parts by dividing the image in (1, 2x2, 1x3, 3x1) sunblock's and concatenate them into one feature vector, by this way, it can include the spatial information.

Table 1. Summary of the features in this work.

Category	*Feature name*	#	*Short Description*
Color	Color moments	144	Three central moments (Mean, Standard deviation and Skewness) on HSV channels.
	Color histogram	64	4^3 = 64 bin histogram is created based on each HSV channel.
Texture	Grey level Co-occurrence matrix	16	GLCM, described by *Haralick* (1973), defined over an image to be the distribution of co-occurring values at a given offset.
	Local binary pattern(LBP)	256	A compact multi-scale texture descriptor analysis of textures with multiple scales by combining neighborhoods with different sizes.
Shape	Histogram of line orientations	12	12 different orientations by using Hough transform
Mid-level	Harmony	11	Try to describe color harmony of images based on Itten's color theory [7].
	Dynamism	11	The ratio of oblique lines against horizontal and vertical ones. Indeed, oblique lines communicate dynamism and action whereas horizontal or vertical lines rather communicate calmness and relaxation.
Others	Y. Ke	5	Features by Y. Ke [3] were chosen to measure criteria including: spatial distribution of edges, color distribution, hue count, blur, contrast and brightness.
	R.Datta	44	Features by R. Datta [2] including: exposure of light and colorfulness, saturation and hue, the rule of thirds, familiarity measure, wavelet-based texture, size and aspect ratio, region composition, low depth of field indicators, shape convexity. Note that we implement most of the features (44 of 56) except those (some from familiarity measure and Region composition) that are related to IRM (integrated region matching) technique [4].
	M. Nishiyama	200	Features by M. Nishiyama [4] namely "bags-of-color-patterns" based on the photo's color harmony the sum of color harmony scores computed from the local regions of a photograph is closely related to its aesthetic quality.

Lines also carry important semantic information in images: oblique lines communicate dynamism and action whereas horizontal or vertical lines rather communicate calmness and relaxation. To characterize dynamism in images, the ratio is computed between the numbers of oblique lines with respect to the total number of lines in an image. At last the dynamism features are obtained by extracting in 11 parts just as the harmony feature.

3 MKL for Image Aesthetic Classification

MKL refers to set methods that learn an optimal linear or non-linear combination of a predefined set of kernels. The reasons we build our image aesthetic classification based on MKL include: a) the ability to select an optimal kernel and parameters from a larger set of kernels, without an explicit feature selection step and b) combining data from different types of feature (e.g. color and texture) that have different notions of similarity and thus require different kernels. Moreover, instead of creating a new kernel, multiple kernel algorithms can be used to combine kernels which are already established for each individual features. All of these can improve the classification performance and makes the interpretation of the results straightforward. MKL has earlier been applied for visual object classification in [9], and we are the first to introduce it into image aesthetic classification. Our experimental results demonstrate the advantages of the MKL framework in image aesthetic classification.

According to the works [10, 12], we employ the Lasso MKL as our kernel learning method for it's simple and efficient. The algorithm formulates an alternating optimization method and updates the kernel weights $\eta_m$as follows:

$$\eta_m = \frac{\|\omega_m\|^2}{\sum_{h=1}^{P} \|\omega_h\|^2} \tag{1}$$

where $\|\omega_m\|^2 = \eta_m^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathrm{K}_m(x_i^m x_j^m)$ is from the duality conditions. K_m denotes the kernel function calculated on the mth feature representation. P is the number of kernels or feature representations (P = 10 in our case), and $\sum_{m=1}^{P} \eta_m$= 1.

After updating the kernel weights in equation (1), the algorithm then solves a classical SVM problem by maximizing SVM dual formulation with the combined kernel $\mathrm{K} = \sum_{m=1}^{P} \eta_m \mathrm{K}_m$ as follows:

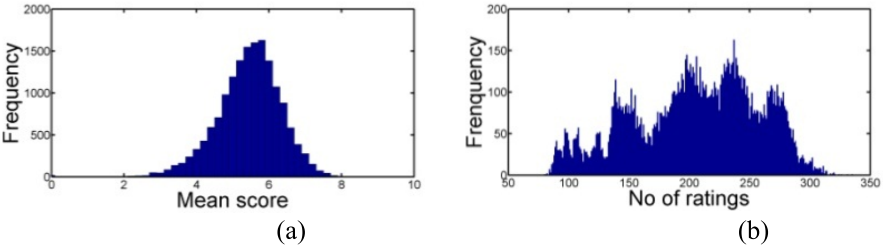
$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathrm{K}(x_i, x_j) \tag{2}$$

subject to the constraints: $0 \le \alpha_i \le C$ for all $\mathrm{i} = 1, \dots, \mathrm{N}$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$, where C is the regularization parameter and y_i is the label (±1) of training sample x_i. The two steps alternate until convergence.

4 Experiments and Results

4.1 Database

Previously, due to copyright restrictions, there is few public available database for photo aesthetic quality analysis. An exception is the preliminary work in [2] where photos have been collected from three Web-based sources [5, 6], in which photos have been rated by users of its community. Unfortunately, becasue photographs have been removed, it is hard to collect the same dataset as R. Datta [2] (about 15% has changed). Therefore, we have chosen to build a large and diverse training and testing database based on the Web source DPChallenge.com [6], which was created in January 2002 by Drew Ungvarsky and Langdon Oliver. To date, 180,255 users have submitted 318,599 photographs to 2086 challenges. Thus, we have collected a total of 60000 photographs by random crawling. Each photo is rated by at least 115 users with a mean average of 185 users, and the mean scores of all images are 5.6 with a std. dev. of 0.72. Figure 2 shows the distribution of average score and number of ratings. In order to reduce noise in the experiments, the top 10% and bottom 10% mean score of the photos were chosen and assigned as high and low aesthetic quality photo set respectively. From each set, half of the photos (3000) were used for training and the other half for testing. Some of the photos, especially the high quality ones, contain borders which we removed using a simple color counting algorithm in order to reduce bias in our results.



(a) (b)

Fig. 2. The distribution of mean score (a) and number of ratings (b).

4.2 Results

Experimental Setup. Our experiments are conducted as follows: Firstly, for each set, half of the photos (3000) were used for training and the other half for testing. To obtain the ground truth labels for the SVM classifier, we adopt the top 10% photos as the positive class, whereas those with bottom 10% are treated as the negative class. We conduct experiments in order to (1) compare visual features that show correlation with community-based aesthetics scores. For this, SVM is run 20 times per feature, and using a 5-fold cross-validation; (2) build a classification model based on MKL such that there is no need to select features and generalization performance is near optimal. For the MKL parameters, we set the regularization parameter C as $C = 1$, the kernel width $s = 2\sqrt{D}, D$ is the feature dimension size and the alternating iterations for inference as 20 times.

Results. Figure 3 shows us the accuracy performance of different features. We see that the features from R. Datta [2] receive 65% accuracy as the first place among the 10 features described in Section 2. But considering the size of feature vector, the feature from Y.Ke [3] with 5 dimensions belongs to the most effective one. The color and texture-based features achieve better results compared to the shape ones (dynamism and line histogram). SVM_all simply concatenates all the 10 features to a single feature, and receives a result around 70% better than other single features. This confirms our belief that by employing fusion method, we can improve the accuracy of aesthetic classification as it can provide complementary information to represent photo aesthetics.

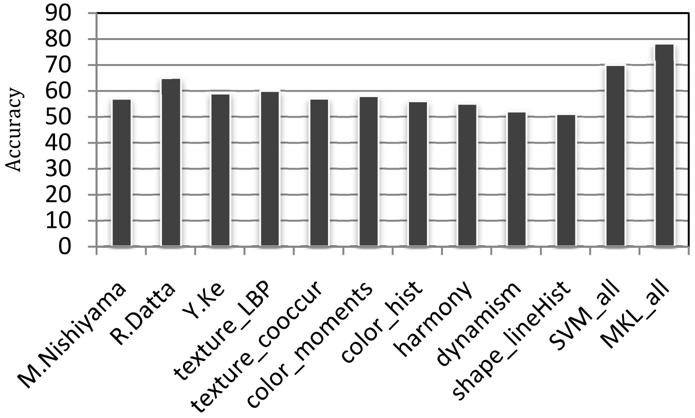


Fig. 3. The performance of different features.

Table 1 shows the comparison with other works. We can see that our method based on MKL scheme received the best performances. One should be noted that our database and Y. Ke's [3] are different, but are collected from the same web source and with the same training setting. Considering the nature of this problem, these classification results are indeed promising.

Table 2. The comparison with other works.

	Data size	*Selection / combine method*	*Performance*
R.Datta[2]	3581	filer and wrapper	70.12%
Y. Ke[3]	≈ 6000	naïve Bayers classifier	76%
Our method	**6000**	**MKL**	**78.3%**

Note: The data source [2] is from Photo.net [5].

5 Conclusion and Future Work

In this paper, we have presented an approach for aesthetic image classification based on MKL, which can make use of different feature representations simultaneously such that it jointly learns the feature weights and the corresponding classifier, by seeks for maximizing the classification performance without explicit feature selection steps. The experiments are conducted on a large diverse database built from online photo sharing website, and the results demonstrated the advantages of MKL in terms of feature selection, classification performance, and interpretation, for the aesthetic image classification task.

In future works, we believe that following effort can further enhance the performance: (1) proposing higher level visual features by combing visual saliency information, which indicated the region of interesting (ROI) in the image, (2) investigating the photograph metadata such as exposure time, aperture and ISO, and (3) introducing effective combination or regression techniques such as, the evidence theory and sparse logistic regression methods.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (No. 0803603) (No. 61402021), Science and Technology Program of the State Archives Administration (2015-B-10), Beijing Nature Science Foundation (No. 4144070) (No. 4144072) and supported in part by the Fundamental Research Funds for the Central University in UIBE (14QD21).

References

1. Wang, J.Z., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries. IEEE Trans. on PAMI **23**(9), 947–963 (2001)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
3. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: Proceedings of CVPR (2006)
4. Nishiyama, M., Okabe, T., Sato, I., Sato, Y.: Aesthetic quality classification of photographs based on color harmony. In: CVPR (2012)
5. Photo.net. http://photo.net
6. DPChallenge. http://www.dpchallenge.com
7. Columbo, C., Del Bimbo, A., Pala, P.: Semantics in visual information retrieval. IEEE Multimedia **6**(3), 38–53 (1999)
8. Datta, R., Li, J., Wang, J.: Algorithmic inferencing of aesthetics and emotion in natural images: an exposition. In: Proceedings of ICIP (2008)
9. Bucak, S.S., Jin, R., Jain, A.K.: Multiple Kernel Learning for Visual Object Recognition: A Review. T-PAMI (2013)
10. Zhang, H., Yang, Z., Gönen, M., Koskela, M., Laaksonen, J., Honkela, T., Oja, E.: Affective abstract image classification and retrieval using multiple kernel learning. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP 2013, Part III. LNCS, vol. 8228, pp. 166–175. Springer, Heidelberg (2013)
11. Yu, et al.: L2-norm multiple kernel learning and its application to biomedical data fusion. BMC Bioinformatics **11**, 309 (2010)
12. Bach, F.: Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research **9**, 1179–1225 (2008)

1000 Fps Highly Accurate Eye Detection with Stacked Denoising Autoencoder

Wei Tang, Yongzhen Huang, and Liang Wang(✉)

Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
wangliang@nlpr.ia.ac.cn

Abstract. Eye detection is an important step for a range of applications such as iris and face recognition. For eye detection in practice, speed is as equally important as accuracy. In this paper, we propose a super-fast (1000 fps on a general PC) eye detection method based on the label map of the raw image without face detection. We firstly produce the label map of a raw image according to the coordinates of its bounding box . Then we train a stacked denoising autoencoder (SDAE) which is specifically designed to learn the mapping from the raw image to the label map. Finally, through an effective post-processing step, we obtain the bounding boxes of two eyes. Experimental results show that our method is about 2,500 times faster than the deformable part-based model (DPM) while maintaining a comparable accuracy. Also, our method is much better than the popular LBP+Cascade model in terms of both accuracy and speed.

Keywords: Eye detection · Autoencoder · Label map

1 Introduction

As a challenging problem in computer vision, eye detection has attracted increasing attention in recent years due to its importance in some real applications such as iris and face recognition. Eye detection aims to solve the problem of getting the accurate position of eyes in a given image. Great achievements have been made on the accuracy of object detection over the past years [2] [3] [7] and these methods could be directly utilized to eye detection.

However, when facing truly practical problems, we find that few methods can run at a fast speed and keep a high accuracy at the same time. On one hand, despite of the great accuracy achieved by many recently proposed methods such as DPM [2] and RCNN [3], they usually rely on tools of high performance computing (HPC) for the demand of real-time detection. Sometimes, even though the HPC technology is adopted, the speed still cannot meet the requirements in applications such as on the embedded devices. On the other hand, traditional methods like LBP+Cascade can run in real time, but their detection accuracy is usually not satisfactory.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 237–246, 2015.
DOI: 10.1007/978-3-662-48570-5_23

In this paper, we propose a novel method based on the label map to address fast and accurate eye detection. To obtain great acceleration, we adopt SDAE [14] to learn the mapping from raw image to label map image, which can be very fast in testing because SDAE needs only a few times of matrix multiplication.

Label map has been proposed in [8] for face parsing. Our method differs from that in two aspects. Firstly, the method in [8] deals with the face parsing problem, so the label map needs segmentation for each pixel. However, our task is specific object detection and the label map with the location of the bounding box is enough, which means traditional object detection datasets can be directly utilized to train our model. Secondly, face parsing in [8] needs the face detection results as the input. In our approach, to avoid the use of face detector, we adopt a different strategy called *patch based label map training*. This strategy ensures getting the whole label map with a high accuracy, at the same time at a super-fast speed.

The major contributions of this paper are summarized as follows:

1) We propose a novel method based on the label map for reliable eye detection, which avoids the time-consuming face detection. It is robust to illumination changes, non-rigid deformation, incomplete object and partial occlusion.
2) We specifically design a SDAE model to learn the mapping from a raw image to the label map, and propose a patch based label map training strategy to effectively train the SDAE model.
3) Our approach is about 2,500 times faster than DPM while maintaining a similar accuracy and is much better than the LBP+Cascade model in terms of both accuracy and speed, which gives the potential for our approach to be used in computing-limited scenarios such as the embedded mobile devices.

2 Related Work

The work proposed in this paper is related to object detection and deep learning. We simply introduce some related work as follows.

Object Detection. Object detection has long been studied and attracted increasing attention in recent years. As for dealing with real-time tasks such as face detection in videos, LBP+Cascade [7] might be one of the most mature methods and has been proved very effective in common use. Deformable part based model (DPM) proposed in [2] is another milestone because of its high detection accuracy. Both LBP+Cascade and DPM in essence are the sliding window based methods.

These methods first turn an image into a large amount of image windows by slidingly sampling windows in the image pyramid. Then features (LBP in LBP+Cascade and HOG in DPM) are extracted from each window and classified by a category specific classifier (Cascade in LBP+Cascade and Latent SVM in DPM). Finally through a post-processing method named non-maximum suppression (NMS) [2], a bounding box surrounding the specific object can be obtained. Sliding window based methods turn out to be effective in some object detection tasks such as face detection [7], pedestrian detection [2]. A main problem for sliding window based methods is the large amount of image windows. Especially, when objects in an image have a large range of size variance, to guarantee a high recall, the number of layers in the image pyramid should be larger and the stride between two windows should be smaller, which will produce a huge number of image windows. Classifying these image windows will be quite time-consuming. Our method adopts a label map based measure which avoids constructing the image pyramid, and thus significantly reduces the number of image windows (Section 3).

Deep Learning. The deep learning technology, with its strong representation learning capacity, has been utilized to deal with various computer vision problems and great success has been achieved in many areas such as image classification [6] [11] and object detection [3]. There are different deep learning models that have been proposed, such as DBN [5], Autoencoder [10], Convolutional Neural Network (CNN) [6]. Among all these models, CNN may be the most widely used, but its high computing cost is the biggest obstacle in real-time scenarios. Autoencoder (AE) is trained in an unsupervised manner by setting the output equal to the input. Lots of variants of AE have been developed in recent years such as denoising AE(DAE)[12], dropout AE [10], sparse AE [9] and stacked AE (SAE) [13]. AE has its own advantage that all it needs is just a few times of matrix multiplication, which leads to a super-fast forward propagating speed when testing.

3 Our Method

In this section, we detail our method. The whole framework includes three parts: data preparation, SDAE training and bounding boxes acquisition. We explain each part one by one below. More implementation details will be further introduced in Section 4.

3.1 Data Preparation

Two key problems we should solve before training a SDAE are 1) how to get the label map from traditional object detection datasets and 2) how to ensure we have enough data to effectively train the SDAE.

Getting Label Map. Unlike the method in [8], we do not need accurate label map with the segmentation for each pixel during training. Traditional object detection datasets can be directly used in our framework. In particular, we just utilize the size of the input image and the labeled bounding box to create a binary image with the same size. As shown in Fig. 1, we set the value of pixels in the bounding box to one and other pixels to zero, and thus get the label map used in our method.

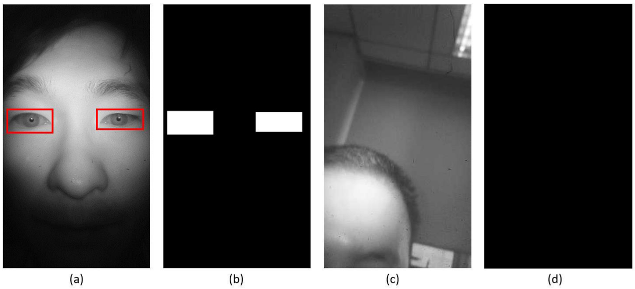


Fig. 1. (a) is an original image with bounding boxes and (b) is the corresponding label map. (c) is an image without the target object and (d) is the corresponding label map.

Patch Based Label Map Training. In our method, instead of first performing face detection as some other methods do, we directly perform eye detection on the whole input image. But training a SDAE with the whole image as input is proved to be unreasonable. For example, although the input image is 480x268, we experimentally find that the size of 80x44 is almost the limit to guarantee good performance. A simple AE with such a size will have ten millions of parameters. Optimizing so many parameters needs huge amount of training data and the optimized model will be very large so that it is hard to be loaded to a normal PC memory.

To solve this problem, we propose a training strategy called *patch based label map training*. In this strategy, instead of using the whole image as the input, we randomly crop image patches from the whole image with a reasonable size as the training data.

Another key point in this strategy is the cleaning of the generated label map patches before training. Let us define the response rate of a label map patch as:

$$\mathrm{r} = \frac{\sum_{i=1,j=1}^{i=N,j=N} I(i,j)}{N^2} \tag{1}$$

where I denotes the label map patch, N denotes the size of I, $I(i,j) \in \{0,1\}$ denotes the pixel value at (i,j) in I. We find that, if the training data contains label map patches with a small r, the output label map will have many small noisy response regions, which will be a severe problem in the post-processing procedure. Therefore if r of a patch is smaller than a predefined threshold, we set the label map of this patch to 0.

Because of the above strategy we propose, we can easily sample thousands of image patches from one original image. Meanwhile, the model can be smaller so that fewer parameters need to be optimized. Fig. 2 shows the procedure of our strategy. To increase the contrast, the red border is added for sampled patches. The r value of the upper patch in Fig. 2 is above the threshold, so no further processing is needed. But the r value of the lower patch in Fig. 2 is smaller than the threshold, so all the values in this patch are set to 0.

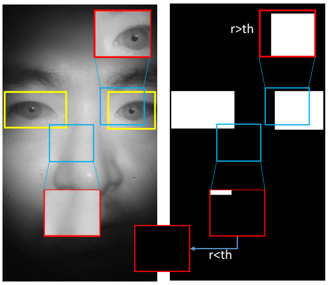


Fig. 2. The procedure of the training strategy described in Section 3.1. (Best viewed in color)

3.2 SDAE Training

The whole training procedure includes two parts: pre-training and fine tuning. The whole model architecture is shown in Fig. 3. (a) and (b) are two DAEs used for pre-training. (c) is the SDAE used in our framework.

Pre-training. Pre-training is an unsupervised layer-wise initialization procedure for deep network to avoid getting stuck in local minima or plateaus [1]. In our framework, we choose a variant of the conventional AE called denoising AE (DAE) as the building block of SDAE. DAE learns to recover a data sample from its corrupted version, which means it can learn more robust features than conventional AE. The architecture of DAE is shown in Fig. 3(a).

Suppose there are N training samples. Let i_k denote the kth image patch and $\tilde{i}_k$ denote the corrupted i_k, where corruption can be Gaussian or salt-and-pepper noise. Let W^1 and W^2 denote the weights (including the bias) for the encoder and decoder respectively. A DAE is learned by solving the following optimization problem:

$$\min_{W^1,W^2} \sum_{k=1}^{N} ||i_k - \hat{i}_k||_2^2 + \lambda(||W^1||_F^2 + ||W^2||_F^2) \tag{2}$$

where

$$h_k = f(W^1\tilde{i}_k) \tag{3}$$

$$\hat{i}_k = f(W^2 h_k) \tag{4}$$

Here λ is a parameter that balances the reconstruction loss and weight penalty terms, $||\cdot||_F^2$ denotes Frobenius norm, and $f(\cdot)$ is a nonlinear activation function which is typically a sigmoid function or hyperbolic tangent function. As shown in Fig. 3, two DAEs are trained in our framework, thus leading to three hidden layers in SDAE.

Specifically Designed SDAE and Fine-Tuning. In the fine-tuning procedure, we utilize the two pre-training DAEs to build a SDAE. The weights of SDAE are all initialized with the weights from the pre-training stage as shown in Fig. 3(c) except that the weights between the last hidden layer and the output

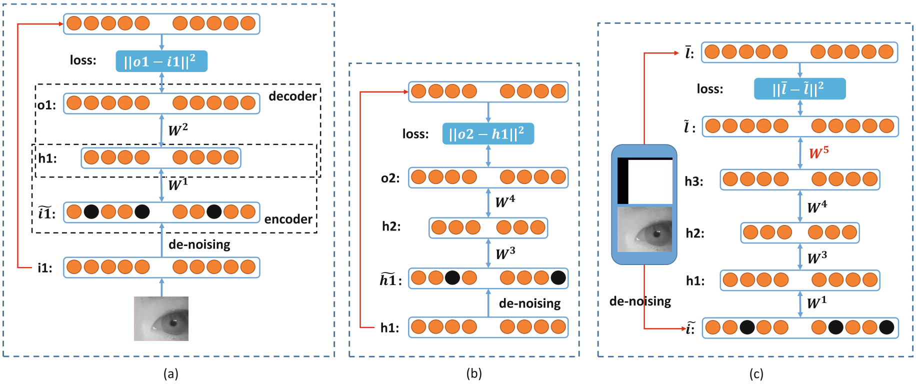


Fig. 3. (a) and (b) are two DAEs used for pre-training.(c) is the SDAE used in our framework. The weights in (c) are initialized with weights from (a) and (b) except that the weights between $h3$ and $\tilde{l}$ are randomly initialized.

layer are randomly initialized. These weights in conventional SDAE should be set to W^2, because the output is just a re-construction of the input. But in our method, we expect the output to produce the corresponding label map of the input image patch, so we initialize these weights with a random matrix denoted W^5 aiming to let the optimizing algorithm search for the optimal solution in a larger scope.

Other Strategies Adopted in Training. Overcomplete filters are used in the hidden layer of DAE. In conventional AE, the hidden layer is always in a "bottleneck" style. Overcomplete filters demand that the number of units in the hidden layer is larger than that of the input layer because it has been found that an overcomplete basis can usually capture better image structure [16].

To further learn more meaningful features, we also adopt sparsity constraints [4] imposed on the hidden units. If a sigmoid activation function is used, the output of each neural unit can be regarded as the probability of being active. Let ρ_i denote the target sparsity level of the ith unit and $\hat{\rho}_i$ denote its average empirical activation rate. The cross-entropy of ρ_i and $\hat{\rho}_i$ can then be introduced as an additional penalty term to Eqn.(2):

$$\sum_{j=1}^{s_2} \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \qquad (5)$$

where s_2 is the number of hidden units.

3.3 Bounding Boxes Acquisition

Once we obtain many label map patches of the input image, we can merge these patches to form the whole label map according to their original positions. Example merging result is shown in Fig. 4(a) and we usually binarize the label

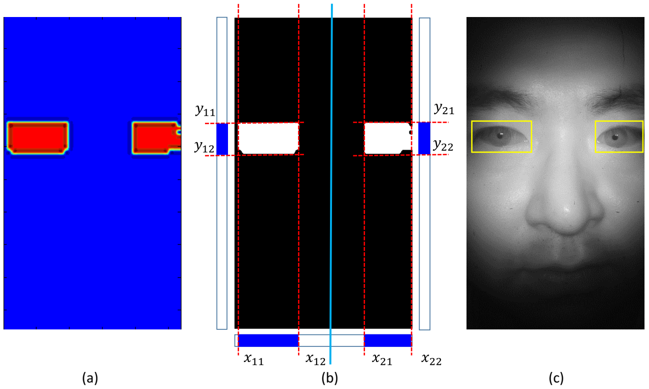


Fig. 4. (a) is the label map image obtained by merging all the label map patches from SDAE. (b) is the binary label map image, showing how to produce bounding boxes from label map. Please see the text for details. (c) is the result we get finally.

map as shown in Fig. 4(b). To get the bounding box from such a label map, we introduce a simple but very effective method in our eye detection task.

This method is first to add the matrix of the binary image along y axis to get one or two longest continuous non-zero sequence, which corresponds to the x coordinates of the eyes. Then we can separate the binary image into two parts according to the x coordinates. Adding the matrix of each part along x axis will get the corresponding y coordinates. We show this procedure in Fig. 4(b).

When response areas in the label map cannot be separated by x or y axis, we can also adopt some methods such as finding contours to get the bounding box from the label map.

4 Experimental Results

In this section, we first introduce the implementation details of our experiments. Then we present our experimental results including the results of different strategies, comparison with other methods and the visualization of our detection results.

4.1 Experimental Setting

Dataset. We collect 2,732 near-infrared eye images as the dataset used in our experiments for the background of this task is the Asian iris recognition. We randomly choose 2,182 images for training and the remaining 550 images for testing. The image size is 480x268. In our experiments, we resize all the images to 80x44. The patch size is 36x36. In testing, we sample patches with a stride of 18, just the half of the patch size.

Implementation Details. We empirically set the threshold of r to 0.02. For DAE and SDAE, we set the denoising ratio to 0.5, the sparsity target to 0.05,

Table 1. Effectiveness comparison of different strategies.

F1	h1 = 1024, h2 = 512	h1 = 2048, h2 = 1024
$r = 0$	—	0.879
$r = 0.02$	0.892	0.906

the weight penalty to 0.0001, the batch size to 100 and the learning rate to 0.2 in DAE and 0.1 in SDAE respectively. The sizes of $h1$ and $h2$ are 2,048 and 1,024 respectively. For DPM, the settings suggested by the paper [15] are used.

Evaluation Metrics. For accuracy, we use the same criterion as in [2] that the predicted bounding box is valid if its overlap ratio with the ground truth is bigger than 50%. Because our method does not return a score for a predicted bounding box, we use $F1$ value as the performance metric instead of the ROC curve. Let P denote precision and R denote recall rate, $F1$ can be defined as follows:

$$F_1 = \frac{2PR}{P+R} \tag{6}$$

From the formula, we can see that $F1$ value is a balance between precision and recall rates.

For speed, we use frames per second (fps) as the performance metric.

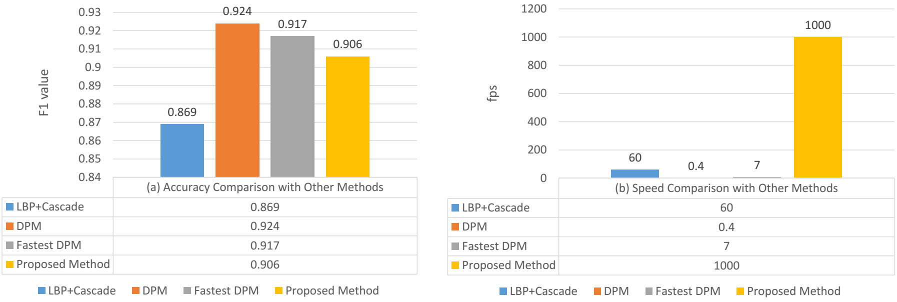
4.2 Basic Results

Effectiveness of Cleaning Small Response Area. We set r to 0 and 0.02 separately. Results show that when r is equal to 0, the label map will have some noisy regions which results in the overlap between the predicted bounding box and ground truth being less than 50%. In our test, this will reduce the accuracy by about 2.7% compared with $r = 0.02$.

Effectiveness of Overcomplete Filters. Except for setting $h1$ to 2,048 and $h2$ to 1,024, we also use the conventional "bottleneck" hidden layer with $h1 =$ 1,024 and $h2 = 512$. It shows that by adopting overcomplete filters, the accuracy can be improved by about 1.4%. All these results are shown in Tab. 1.

4.3 Comparison

We compare our method with other methods from two aspects: speed and accuracy. As for the methods we choose to compare with, LBP+Cascade is the most widely used object detection method and DPM has achieved excellent results on the challenging PASCAL VOC dataset. The accuracy and speed comparisons are shown in Fig. 5. We can see that our method can get a surprising 1,000 fps on a general PC with CPU (i7-3770 in our experiments), which is 17 times faster than LBP+Cascade, 140 times faser than the fastest DPM [15] and 2,500 times faster than original DPM [2]. The accuracy of our method is slightly lower than DPM but obviously better than LBP+Cascade. Overall, our method achieves a super-fast speed while maintaining a high accuracy.



	(a) Accuracy Comparison with Other Methods
🟦 LBP+Cascade	0.869
🟧 DPM	0.924
⬜ Fastest DPM	0.917
🟨 Proposed Method	0.906

	(b) Speed Comparison with Other Methods
🟦 LBP+Cascade	60
🟧 DPM	0.4
⬜ Fastest DPM	7
🟨 Proposed Method	1000

Fig. 5. Comparison with other methods on accuracy and speed.

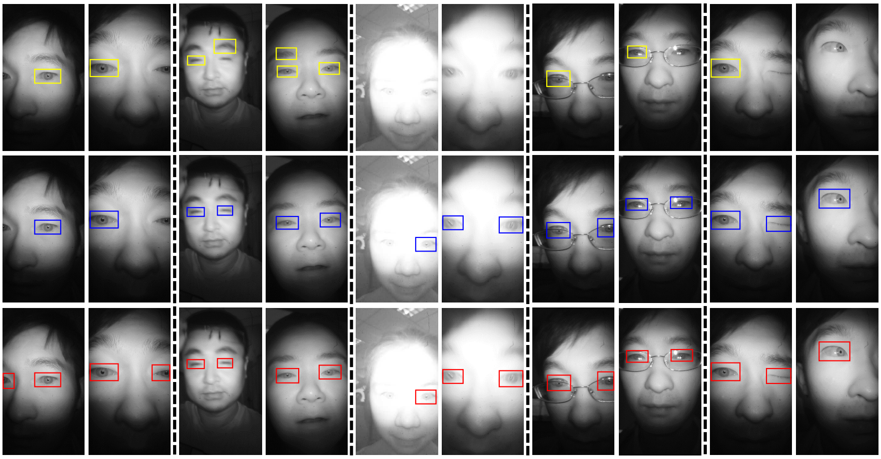


Fig. 6. Example results produced by LBP+Cascade(top), DPM (middle) and our approach (bottom). From left to right, we can see that our method is robust to the incomplete object, background disturbance, changes of illumination, occlusion and non-rigid deformation (best viewed in color).

4.4 Visualization

We show some detection results in Fig. 6. From left to right, we can see that our method can effectively deal with the incomplete object (such as the incomplete eyes), background disturbance (such as the disturbance of eyebrows), changes of illumination, occlusion (such as wearing glasses) and non-rigid deformation (such as the non-rigid deformation of eyes).

5 Conclusion

In this paper, we have presented a label map based eye detection method. By training a specifically designed SDAE, the label map can be accurately produced. The resulting method is 2,500 times faster than DPM while maintaining a comparable accuracy, which shows the great potential of our method to be used for real-time applications and in computing-limited scenarios.

Acknowledgments. This work is jointly supported by National Basic Research Program of China (2012CB316300), National Natural Science Foundation of China (61135002, 61420106015, 61203252), SAMSUNG GRO Program, CCF-Tencent Open Fund and 360 OpenLab Program.

References

1. Bengio, Y.: Learning deep architectures for AI. Foundations and trends in Machine Learning **2**(1), 1–127 (2009)
2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI **32**(9), 1627–1645 (2010)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
4. Hinton, G.: A practical guide to training restricted boltzmann machines. Momentum **9**(1), 926 (2010)
5. Hinton, G.E.: Deep belief networks. Scholarpedia **4**(5), 5947 (2009)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
7. Liao, S.C., Zhu, X.X., Lei, Z., Zhang, L., Li, S.Z.: Learning multi-scale block local binary patterns for face recognition. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 828–837. Springer, Heidelberg (2007)
8. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: CVPR. IEEE (2012)
9. Ng, A.: Sparse autoencoder. CS294A Lecture notes 72 (2011)
10. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
11. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. arXiv preprint arXiv:1409.4842 (2014)
12. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103. ACM (2008)
13. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research 11 (2010)
14. Wang, N., Yeung, D.Y.: Learning a deep compact image representation for visual tracking. In: NIPS, pp. 809–817 (2013)
15. Yan, J., Lei, Z., Wen, L., Li, S.Z.: The fastest deformable part model for object detection. In: CVPR. IEEE (2014)

Motion Compensation Based Fast Moving Object Detection in Dynamic Background

Wei Zhang, Chenglong Li, Aihua Zheng(✉), Jin Tang, and Bin Luo

Key Lab of Intelligent Computing and Signal Processing of Ministry of Education,
AnHui University, Hefei 230601, China
{wzhang1127,lcl1314}@foxmail.com, {ahzheng214,tj}@ahu.edu.cn,
luobinahu@163.com

Abstract. This paper investigates robust and fast moving object detection in dynamic background. A motion compensation based approach is proposed to maintain an online background model, then the moving objects are detected in a fast fashion. Specifically, the pixel-level background model is built for each pixel, and is represented by a set of pixel values drawn from its location and neighborhoods. Given the background models of previous frame, the edge-preserving optical flow algorithm is employed to estimate the motion of each pixel, followed by propagating their background models to the current frame. Each pixel can be classified as foreground or background pixel according to the compensated background model. Moreover, the compensated background model is updated online by a fast random algorithm to adapt the variation of background. Extensive experiments on collected challenging videos suggest that our method outperforms other state-of-the-art methods, and achieves 8 fps in efficiency.

Keywords: Fast object detection · Random algorithm · Dynamic background · Motion compensation

1 Introduction

Moving object detection with dynamic background is to detect moving objects under a moving camera, and has a broad prospect of application and research value in the intelligent transportation, medical diagnosis, security monitoring, and many other industries. However, due to the high complexity of the existing method which are unable to meet the time demand of many applications, it is still a challenging subject in computer vision.

Aimed at overcoming this limitation, this paper proposes a fast moving object detection framework in dynamic background, in which the motion compensation algorithm is utilized to accommodate the dynamic background, and the background model is updated online in a probability way to adapt the variation of background. Specifically, the background model of each pixel consists of a set of pixels, which are initialized by its location and neighbors. When new frame arriving, the optical flow algorithm, based on edge-preserving patch matching

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 247–256, 2015.
DOI: 10.1007/978-3-662-48570-5_24

is employed to compensate the motion of each pixel and propagate their background models from previous frame to current one. Then, every pixel can be classified as the foreground or background pixel by the matching score with their background models. Furthermore, the background models are updated in an online fashion to adapt the variation of background.

To the best of our knowledge, it's the first time to develop a near real-time moving object detection in dynamic background. The key contributions of this paper are summed up in three aspects. Firstly, a general framework is proposed for robustly and fast detecting moving objects in dynamic background, in which the detection speed can reach near real-time. Secondly, a robust background model based on motion compensation is developed and updated online by a random algorithm to adapt the motion and variation of background over time. Thirdly, 10 challenging videos are collected in dynamic background from different scenes to comprehensively evaluate our approach against other state-of-the-art approaches. Extensive experiments on the collected challenging video sequences suggest that our method outperforms other state-of-the-art methods in accuracy, and achieves 8 fps in efficiency.

2 Related Works

Generally, moving object detection methods can be divided into two categories, *i.e.*, static background and dynamic background. At present, moving object detection in static background has become an increasingly mature technique and many related technologies have been successfully applied to real life. Stauffer et al. [11] proposed an adaptive background mixture models for real-time tracking, in which each pixel was modelled as mixture of Gaussian while using an online approximation to update the model. Some improved approaches on Gaussian Mixture Model (GMM) had proposed to address different issues, such as parameters initialization [7], model updating [8] and the number of Gaussian components [18]. Although these approaches achieved nearly real-time, it was still difficult to apply them to many applications unless with some parallel optimizations. Barnich et al. [5],[14] presented a simple background modelling method to detect the moving object with high accuracy and efficiency. The background model of each pixel consisted of a set of values taken in the past at the same location or in the neighborhood and randomly updated from the last pixel at same location or its neighbors. Although lots of progress has been made on moving objects detection in static background, there still exists many critical issues in dynamic background. Zhou et al. [17] proposed a moving object detection framework DECOLOR to address several complex scenarios, such as non-rigid motion and dynamic background. They assumed that the transformation between consecutive frames was linear and thus utilized the 2D parametric transforms [12] to model translation, rotation, and planar deformation of the background. DECOLOR can achieve state-of-the-art performance, but it was time-consuming and only processed the video in a batch fashion. Therefore, we aim at finding a kind of better way to solve some mentioned problems in dynamic background.

3 Our Approach

The details of our approach are described in this section. We utilize motion estimation algorithm to adaptively maintain a robust background models in the dynamic background. Fig. 1 shows the flowchart of our framework.

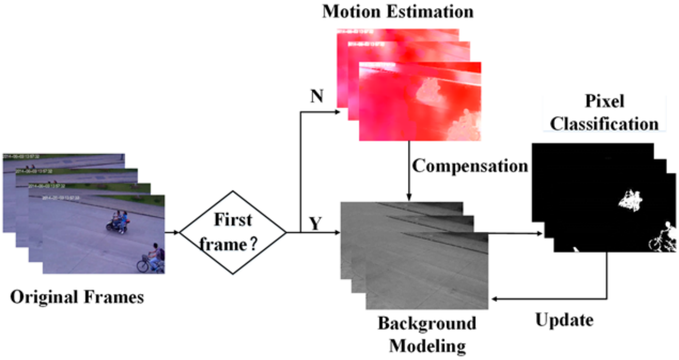


Fig. 1. Flowchart of our framework.

3.1 Motion Estimation

In this paper, the motion of each pixel will be accurately estimated to propagate to their background model to accommodate the motion of the camera. Most of existing methods on dense optical flow are time-consuming and computationally inefficient [1]. On the other hand, a fast optical flow algorithm based on edge-preserving PatchMatch is recently proposed by Bao et al. [3] with high accuracy and efficiency. Therefore, we employ the edge-preserving PatchMatch optical flow to estimate the motion of background in this work, and briefly review it as follows.

The edge-preserving PatchMatch optical flow is a fast algorithm that employs approximate the nearest neighbor field [6] to handle the large displacement motions and consists of four steps: matching cost computation, correspondence approximation, occlusions and outliers handling, and subpixel refinement.

(1) Matching Cost Computation. The edge-preserving PatchMatch optical flow follows the traditional local correspondence searching framework [10]. To make the nearest neighbor field preserve the details of the frame, it employs bilateral weights [16] into matching cost calculation, and can be defined as

$$d(a,b) = \frac{1}{W}\sum_{\Delta} \omega(a,b,\Delta)C(a,b,\Delta), \tag{1}$$

where a and b denote two pixels, $\varDelta$ indicates the patches center on a and b , W is a normalization factor, $\omega(\bullet)$ is the bilateral weighting function and $c(\bullet)$ is the robust cost between a and n. More detailed definitions please refer to [3].

(2) Correspondence Approximation. To produce high-quality flow fields, this optical flow method utilizes self-similarity propagation and a hierarchical matching scheme to approximate the exact edge-preserving Patch Match[4]. Firstly, self-similarity propagation algorithm is based on the fact that adjacent pixels tend to be similar to each other. Specifically, for each pixel, a set of pixels from its surrounding region is randomly selected and stored into a self-similarity vector in the order of their similarities to the center pixel. Then, its adjacent pixels' vector is merged into its own vector from top-left to bottom-right. This process is reversely repeated. Thanks to the propagation between adjacent pixels, the algorithm can produce reasonably good approximate results in a much faster speed. Secondly, a hierarchical matching scheme is employed to further accelerate the algorithm and similar with SimpleFlow method [13].

(3) Occlusions and Outliers Handling. The edge-preserving PatchMatch optical flow explicitly performs the forward-backward consistency check [9] between the two nearest neighbor fields to detect occlusion regions. Moreover, a weighted median filtering is performed [2] on the flow fields to remove the outliers.

(4) Subpixel Refinement. The edge-preserving PatchMatch optical flow produces subpixel accurately with a more efficient technique - paraboloid fitting, which is a 2D extension from the 1D parabola fitting [15].

3.2 Background Modeling

Compared with the background models of a static background, the background modeling in dynamic background is difficult to maintain online since the background pixels are also moving. Although estimated optical flow can compensate the background motion, the background model is still sensitive to noises, due to incorrect optical flow estimation. Thus, a robust pixel model of background is proposed in this paper to adaptively detect the objects in the dynamic background. The two main components of the proposed background model can be described as follows.

Initialization. For each input video, the first frame is selected to initialize the background model. The background model of each pixel is a set of pixel values, and can be represented as

$$B(p) = \{I(p_1), I(p_2), \cdots, I(p_n)\}, \tag{2}$$

where $p_i \in N(p)$, and $N(\bullet)$ indicates the neighbors of pixel p. $I(\bullet)$ denotes the pixel value. For each pixel, n samples are selected from itself and its neighboring pixel values to initialize its background model.

Update. In this section, we assume that each pixel has been accurately classified by the background model (the details are discussed in next section) when new frame arriving. Thus, the background model of each pixel can be updated online by randomly selecting the classified background pixels at the same location or its neighbors. Specifically, for one classified background pixel p_b, two robust background model updating strategies are adapted to obtain its background model $B(p_b)$.

Firstly, one element from $B(p_b)$ is selected in a uniform probability way to replace p_b. Secondly, one pixel value is heuristically taken from its neighbors $N(p_b)$, and substituted by the element randomly selected in $B(p_b)$. Herein, we assume that if one pixel belongs to its background model, its distance to all the values of the background model should be as close as possible. This assumption will be helpful to suppress the effect of the noises. Thus, the selected probability of pixel p_b^i from $N(p_b)$ is defined as

$$q_i = \frac{1}{Q}\exp\{-\frac{1}{n}\sum_{j=1}^{n} D(I(p_b^i), B_j(p_b))\}, \tag{3}$$

where $D(\bullet, \bullet)$ denotes the Euclidian distance function, and Q is a normalization factor.

In addition, to accommodate the change speed of the background, the updating probability, called as updating factor and denoted as η in this paper, is introduced to determine whether the above updating is carried out or not.

3.3 Pixel Classification

Given the background model of previous frame, it can be propagated to the current frame by employing the motion estimation algorithm. Then, every pixel of current frame can be classified as the foreground or background pixel according to the matching scores with their corresponding background model.

For one pixel p, the matching score with background $B(p)$ is defined as

$$M(p) = \sum_{i=1}^{n} \delta(D(I(p), B_i(p)) > R), \tag{4}$$

where $\delta(\bullet)$ denotes the indicator function, and R indicates the adaptive threshold of matching cost, which is determined by the variation σ of $B(p)$. Herein, σ indicates the complexity of the background, and, R is defined as

$$R = \begin{cases} 20, \sigma/2 \leq 20, \\ \sigma/2, 20 < \sigma/2 < 40, \\ 40, \sigma/2 \geq 40. \end{cases} \tag{5}$$

Then, p can be classified by

$$U(p) = \begin{cases} 0, M(p) \geq T, \\ 1, M(p) < T, \end{cases} \tag{6}$$

where 0 and 1 indicate the background and foreground, respectively. T denotes the threshold of matching score.

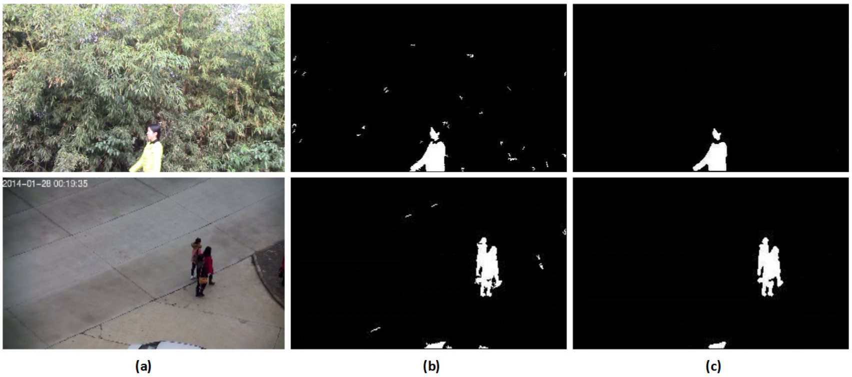


Fig. 2. Illustration of the noises produced by pixel classification and the results by morphological opening operation. (a) Denote the original frames, (b) Denote detection results with noises, and (c) Denote detection results post-processed by morphological opening operation.

3.4 Postprocessing

Due to the pixel-level modelling and classification, the proposed moving object detection may introduce some errors, which usually are isolated points. Therefore, the morphological opening operation is further utilized, in which the structural element is defined as 3 × 3, to remove these errors. Fig. 2 illustrates this process.

4 Experimental Result

In this section, our approach is evaluated on 10 collected challenging video sequences comparing with other state-of-the-art approaches, followed by the discussion of the efficiency analysis of our approach.

4.1 Evaluation Setting

The test videos are the real-life videos recorded from the university security monitoring system by PTZ cameras and hand-held cameras with resolution of 320 × 180 and frame rate 25fps. The evaluation is performed on 10 challenging video containing 4000 frames in total with vary moving objects, including pedestrians, cars, motorcycles and bicycles in dynamic background of the road or the playground, which take into account of the size and the type of moving objects as well as the camera movement and can comprehensively evaluate the performance of the proposed detection algorithm with others.

To make the comparison more comprehensive, the parameters are empirically fixed as $\{n, \eta, T\} = \{20, 0.2, 2\}$ in all evaluations.

4.2 Comparison Results

We compare our approach with two state-of-the-art moving object detection approaches, including DECOLOR [17] and ViBe [5]. Tab. 1 illustrates the average Recall (R), Precision (P) and F-measure on 10 collected video sequences while the detailed R and P values on each video sequence are shown in Fig. 3 and Fig. 4. We can conclude that our method can significantly outperforms other state-of-the-art in Precision and F-measure, although worse than others in Recall.

Table 1. The average R, P and F-measure values on 10 collected video sequences

	R	P	F-measure
DECOLOR	87.9%	49.1%	51.4%
ViBe	82.8%	22.7%	31.9%
Ours	70.1%	74.5%	70.6%

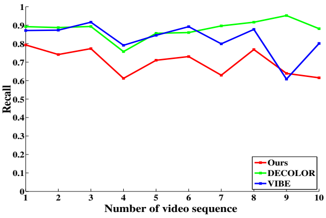


Fig. 3. Recall.

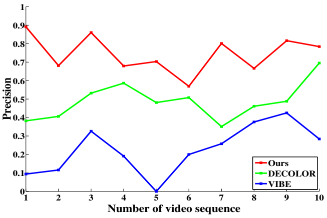


Fig. 4. Precision.

To demonstrate the performance of our proposed detection method against other two methods, we present some typical detection examples with different objects or backgrounds, as shown in Fig. 5. DECOLOR segments moving object in image sequence using a framework that detects the outliers to avoid complicated calculation, and uses low rank model to deal with complex background. It's easier to detect relatively dense and continuous region from the group. However, due to the smooth assumption of DECOLOR, more than one closed objects, especially in some occlusions, usually are detected as one single object (the second and third rows of the second column). ViBe produces ghost and has many noises (the second, the third and the fifth rows of the third column). From the comparative experiments, we can see that the proposed method outperforms DECOLOR in the details of objects, especially in the case of multiple objects, and is robust to the background interference compared with ViBe.

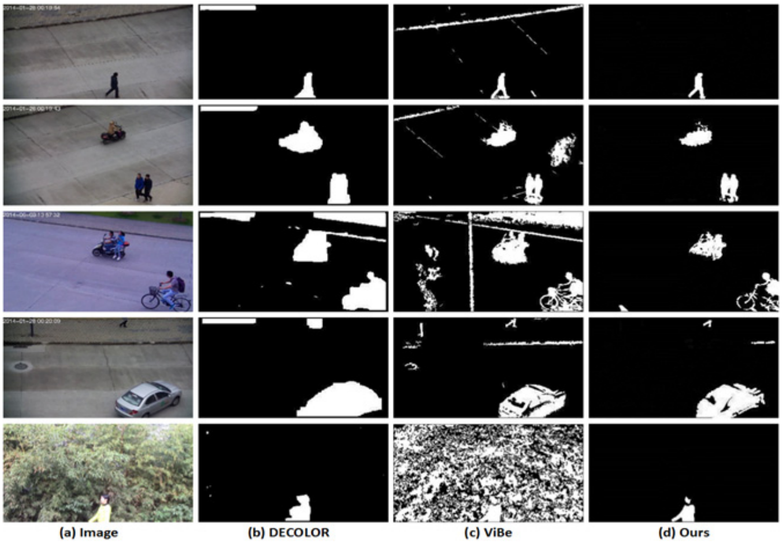


Fig. 5. Detection examples by our method comparing with other two methods, DECOLOR [17] and ViBe [5], with different objects under different dynamic backgrounds. The first column presents the sample frame of each type of video and the rest 3 columns present the detection results by DECOLOR, ViBe and the proposed method, respectively.

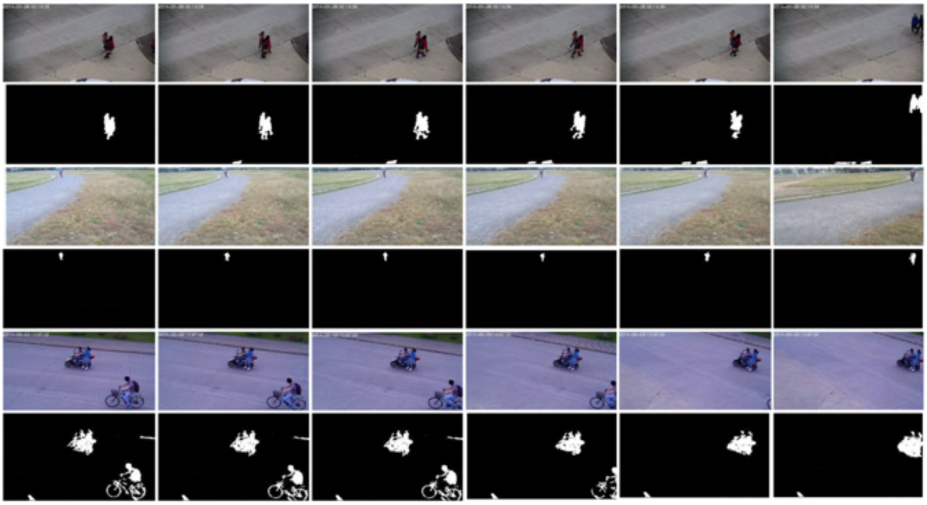


Fig. 6. The detection results of our proposed algorithm on 3 videos in every 5 frames. The 3 odd rows are the frames from 3 test videos respectively and the other 3 even rows are the corresponding detection results by our proposed method.

Fig. 6 presents the detection results of our proposed algorithm on every 5 frames of three videos. From Fig. 6 we can see that the proposed method can achieve superior performance in different surroundings with different types of objects in dynamic background.

4.3 Efficiency Analysis

The experiments are carried out on a desktop with an Intel i7 3.4GHz CPU and 32GB RAM, and implemented on C++ platform without any optimization. In the above experiments, the average runtime of proposed method is 0.12 second per frame while DECOLOR is 20 second per frame. Therefore, the proposed method are substantially faster than DECOLOR. In addition, our method is online while DECOLOR is a batch method. ViBe costs 0.02 second per frame, but it can only handle weak jitter problem of the camera, and is not suitable for the situation of dynamic background.

5 Conclusions

In view of the problems of moving object detection in dynamic background, this paper proposed a fast object detection method based on motion compensation. The background model of each pixel is initialized according to the first frame and is propagated to current frame by employing the edge-preserving optical flow algorithm to estimate the motion of each pixel. Each pixel can be finally classified as foreground or background pixel according to the compensated background model which is updated online by the fast random algorithm. The comparisons with DECOLOR and ViBe demonstrated the effectiveness of the proposed method, particularly in dynamic background. Moreover, the speed of proposed method achieved 8 fps. In future works, we will focus on developing more robust moving object detection approaches in real-time way to meet other applications.

Acknowledgments. Our thanks to the support from the National Nature Science Foundation of China (61502006, 61472002), the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2014A015) and the Natural Science Foundation of Anhui Province (1508085QF127).

References

1. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision **92**(1), 1–31 (2011)
2. Bao, L., Song, Y., Yang, Q., Ahuja, N.: An edge-preserving filtering framework for visibility restoration. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 384–387. IEEE (2012)

3. Bao, L., Yang, Q., Jin, H.: Fast edge-preserving patchmatch for large displacement optical flow. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3534–3541. IEEE (2014)
4. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.: Patchmatch: a randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics-TOG **28**(3), 24 (2009)
5. Barnich, O., Van Droogenbroeck, M.: Vibe: a universal background subtraction algorithm for video sequences. IEEE Transactions on Image Processing **20**(6), 1709–1724 (2011)
6. He, K., Sun, J.: Computing nearest-neighbor fields via propagation-assisted kd-trees. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 111–118. IEEE (2012)
7. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Video-Based Surveillance Systems, pp. 135–144. Springer (2002)
8. Lee, D.-S.: Effective gaussian mixture learning for video background subtraction. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(5), 827–832 (2005)
9. Rhemann, C., Hosni, A., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3017–3024. IEEE (2011)
10. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47**(1–3), 7–42 (2002)
11. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE (1999)
12. Szeliski, R.: Computer vision: algorithms and applications. Springer Science & Business Media (2010)
13. Tao, M., Bai, J., Kohli, P., Paris, S.: Simpleflow: a non-iterative, sublinear optical flow algorithm. In: Computer Graphics Forum, vol. 31, pp. 345–353. Wiley Online Library (2012)
14. Van Droogenbroeck, M., Paquot, O.: Background subtraction: experiments and improvements for vibe. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 32–37. IEEE (2012)
15. Yang, Q., Yang, R., Davis, J., Nistér, D.: Spatial-depth super resolution for range images. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2007, pp. 1–8. IEEE (2007)
16. Yoon, K.J., Kweon, I.S.: Adaptive support-weight approach for correspondence search. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(4), 650–656 (2006)
17. Zhou, X., Yang, C., Yu, W.: Moving object detection by detecting contiguous outliers in the low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(3), 597–610 (2013)
18. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: Proceedings of the 17th International Conference on Pattern Recognition. ICPR 2004, vol. 2, pp. 28–31. IEEE (2004)

A Discriminative Framework for Hashing

Lijie Ding, Xiumei Wang(✉), and Xinbo Gao

School of Electronic Engineering, Xidian University, Xi'an 710071, China
dinglijiedlj@163.com, wangxm@xidian.edu.cn,
xbgao@mail.xidian.edu.cn

Abstract. Hashing methods have been widely applied for fast retrieval and efficient data storage. However, most existing hashing methods have not taken the discriminative features into account in nearest neighbors search which leads to unsatisfied retrieval accuracy, especially for high-dimensional dataset. In order to take best use of discriminative information, we introduce an effective feature extraction framework for hashing that can get high retrieval accuracy in this paper. Firstly, the divergence between two classes is represented by the ratio of signal to noise. Then the discriminative features of high-dimensional data are obtained through the generalized eigen-decomposition. Finally, we exploit the discriminative feature into hashing methods to generate compact binary code. Experimental results on data sets show that the proposed framework can reach better results in comparison with state-of-the-art methods.

Keywords: Approximate nearest neighbor search · Hashing · Discriminative features · Generalized eigenvectors · Precision-recall

1 Introduction

In recent years, with the rapid development of information technology and widespread use of digital multimedia, the amount of global data comes into the era of ZB. Moreover, the dimensionality of data is very high, leading to the problem of curse of dimensionality in many real applications. The high-dimensionality and the big sample size make the data storage and retrieval very challenging. How to extract useful information from these high-dimensional, massive and complex data becomes a core problem. Among large-scale data processing technologies, Approximate Nearest Neighbor (ANN) search[1] is widely used in image retrieval and can retrieve the query sample with sub-linear, logarithmic, or even constant query time.

There has drawn wide attention in mapping image data onto binary codes for ANN. The goal of binary embedding is to well use Hamming distance approximate the input distance so that efficient learning and retrieval can happen in the binary space. It is important to note that related area called hashing becoming popular for efficient retrieval and learning on massive data sets in a large number of application. Hashing creating hash tables make similar data points that are fall in the same (or nearby) bucket with high probability so that yielding a dramatic increase in search speed and the dimension of the hash codes is much lower so that can save memory space greatly. To obtain better hash

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 257–265, 2015.
DOI: 10.1007/978-3-662-48570-5_25

algorithm requires the hash function should ensure that they can quickly calculate the hash codes of new query point and maps similar images to similar binary codes. Especially, Hashing can preserve much more construct information and get better performance while make use of similarity between data and class label information.

Some classical hash methods have been proposed for ANN search. We summarize them into unsupervised, semi-supervised, and supervised methods. For unsupervised hashing, Locality Sensitive Hashing (LSH)[2]-[3] can embedded similar samples into same bucket with high probability by random projects. But LSH needs long hash codes to maintain high precision, which make recall is low and cost storage. The other strategy to obtain effectively hash function based learning scheme. Iterative Quantization (ITQ)[4] minimizing the quantization error through learning the rotation function to get hash function. Some hash techniques get hash codes based on product quantization [5]. Such as K-means hashing (KMH)[6] can learn binary codes through partition the feature space by k-means quantization and estimates the Euclidean distance by Hamming distance of each cluster indexes. In addition, Inductive Hashing on Manifolds(IMH) [7] embedding the original data into a low dimensional space and preserving the inherent neighborhood structure for learning effective hash codes by non-parametric manifold learning. What's more, a semi-supervised hashing (SSH) [8] framework was proposed that minimizes empirical error over the labeled set and an information theoretic regularizer over both labeled and unlabeled set. Besides, for supervised hashing, Supervised Hashing with Kernels (KSH)[9] learning effective hash by utilizing the equivalence between optimizing the code inner products and the Hamming distances.

Although many hash algorithms can get effective retrieval performance, the most hashing methods are not taken the discriminative features into account in nearest neighbors search so that ignore structure information of data and unable to obtain better retrieval accuracy. In this paper, we introduce a kind of feature vector extraction algorithm under signal-to-noise ratio (SNR) framework. We construct conditional second moment matrices for every cluster data points to construct signal and noise vectors, then to get discriminative features of original data through solve the ratio of signal to noise using generalized eigen-decomposition, which is favorable for feature extraction. The algorithm using the data of local neighbor to redefine the relationship between signal and noise can get the discriminative features of data and make each dimensional feature of extracted contains discriminative information as more as possible. Then, we using discriminative features into hashing and confirming its higher accuracy compared with the unsupervised, and supervised hashing methods.

The rest of the paper is organized as follows: The basic principle of hashing algorithm and extract discriminative features are presented in Section 2. Section 3 shows the experimental results. Finally, we conclude the paper in Section 4.

2 The Proposed Method

2.1 Hashing

Given n data points $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$, where each sample is a d-dimensional column vector. The aim of the hash method is to determine r hashing functions

$$H(x) = [h_1(x), h_2(x), \ldots, h_r(x)] \tag{1}$$

where $h_i(x) \in \{0,1\}, i = 1, \ldots, r$, hashing function can transform the samples into r-bit binary hash codes. Based on hash functions $H(x) = \{h_k(x)\}_{k=1}^{r}$, given a sample $x_i, i = 1, 2, \ldots, n$, its hash codes can be obtained by

$$B_{x_i} = H(x_i) \tag{2}$$

2.2 The Framework of Discriminative

Given a labeled data set $\{x_i, y_i\}_{i=1}^{n}$, and $(x_i, y_i) \in \mathbb{R}^d \times [k]$, where each sample is a d-dimensional column vector, there are n training samples and k classes. The aim of the proposed method is to extracting discriminative feature between data of every two classes.

Towards cluster pairs $(i, j) \in T = \{(i, j) \mid i, j \in [k], i \neq j\}$, we expect that the data distribution contains much more information than that contained in the first moment statistics in usual discriminative setting, inspired by [10], we defined signal and noise are the two cluster's conditional second moment matrices $\mathbb{E}[(w^T x)^2 \mid y = i]$ and $\mathbb{E}[(w^T x)^2 \mid y = j]$ respectively. w was the direction of feature projection. As demonstrated in flowchart of the proposed framework in Fig. 1, the ratio of signal to noise about cluster pairs was:

$$R_{ij}(w) = \frac{\mathbb{E}[(w^T x)^2 \mid y = i]}{\mathbb{E}[(w^T x)^2 \mid y = j]} = \frac{w^T \mathbb{E}[x^T x \mid y = i] w}{w^T \mathbb{E}[x^T x \mid y = j] w} = \frac{w^T C_i w}{w^T C_j w} \tag{3}$$

whose local maximizers are the generalized eigenvectors solving

$$C_i w = \lambda C_j w \tag{4}$$

From the view of intuitive, we maximize the ratio of signal to noise is to make the data of class i with bigger covariance and make the data of class j with smaller covariance. We estimate $\mathbb{E}[x^T x \mid y = c]$ using sample covariance of the class c [11]

$$\hat{C}_c = \frac{\sum_{i=1}^{n} \mathbb{I}[y_i = c] x_i x_i^T}{\sum_{i=1}^{n} \mathbb{I}[y_i = c]} \tag{5}$$

Where $\mathbb{I}[y_i = c]$ represent $\mathbb{I}[y_i = c] = 1$ if and only if y_i belongs to class c, otherwise $\mathbb{I}[y_i = c] = 0$. Considering the number of training samples is limited, may not be able to meet the condition of full rank, the commonly solution [12] method is add the appropriate regularization matrix

$$R_{ij}^{\gamma}(w) = \frac{w^T \hat{C}_i w}{w^T \left(\hat{C}_j + \gamma I\right) w} \tag{6}$$

where $\gamma > 0$ is regularization parameter. we will have $w^T \hat{C}_i w = \lambda$ if we assume that each eigenvector w is scaled such that $w^T \left(\hat{C}_j + \gamma I\right) w = 1$, i.e. on average, the squared projection of an example from class i on w will be λ while the squared projection of an example from class j will be 1. We will be able to discriminate the two classes by simply using the magnitude of the projection while λ is far from 1. Then, we extracted generalized eigenvectors for each class pair by solving

$$\hat{C}_i w = \lambda \left(\hat{C}_j + \frac{\gamma}{d} Trace\left(\hat{C}_j\right) I \right) w \tag{7}$$

Generalized eigenvalues are useful for feature selection. Aim to extract the top few eigenvectors, we can get r eigenvectors are associated with the r largest eigenvalues, or make eigenvalues are bigger than a threshold. To all class pairs $(i, j \neq i) \in \{1, \ldots, k\}^2$, According to the guiding of the generalized eigenvalue choose corresponding eigenvectors as the final projection matrix

$$W = \{ w_{ij} \mid \hat{C}_i w_{ij} = \lambda \left(\hat{C}_j + \frac{\gamma}{d} Trace\left(\hat{C}_j\right) I \right) w_{ij},\ i, j \in [k] \} \tag{8}$$

Embedding the original data points to the new coordinates with discriminative information by final projection matrix

$$Y = W^T X \tag{9}$$

The goal about our hashing method is to learn binary codes such that neighbors in the input space are mapped to similar codes in the Hamming space. Then, the compact hash codes for training data are obtained by hash method processing Y and thresholding at zero:

$$B = \left\{ h_k \left(W^T X \right) \right\}_{k=1}^{r} = \left\{ h_k (Y) \right\}_{k=1}^{r} \tag{10}$$

The flowchart of the proposed framework is illustrated as Fig. 1.

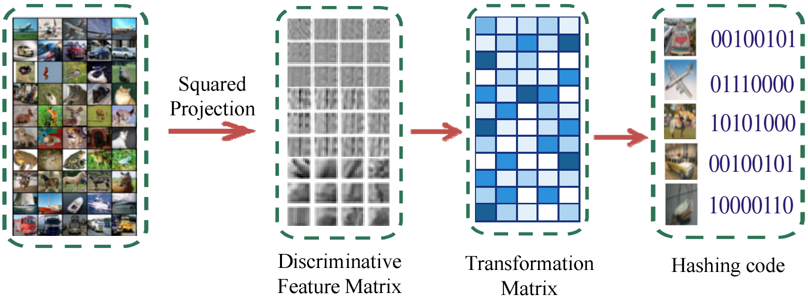


Fig. 1. Flowchart of the proposed framework

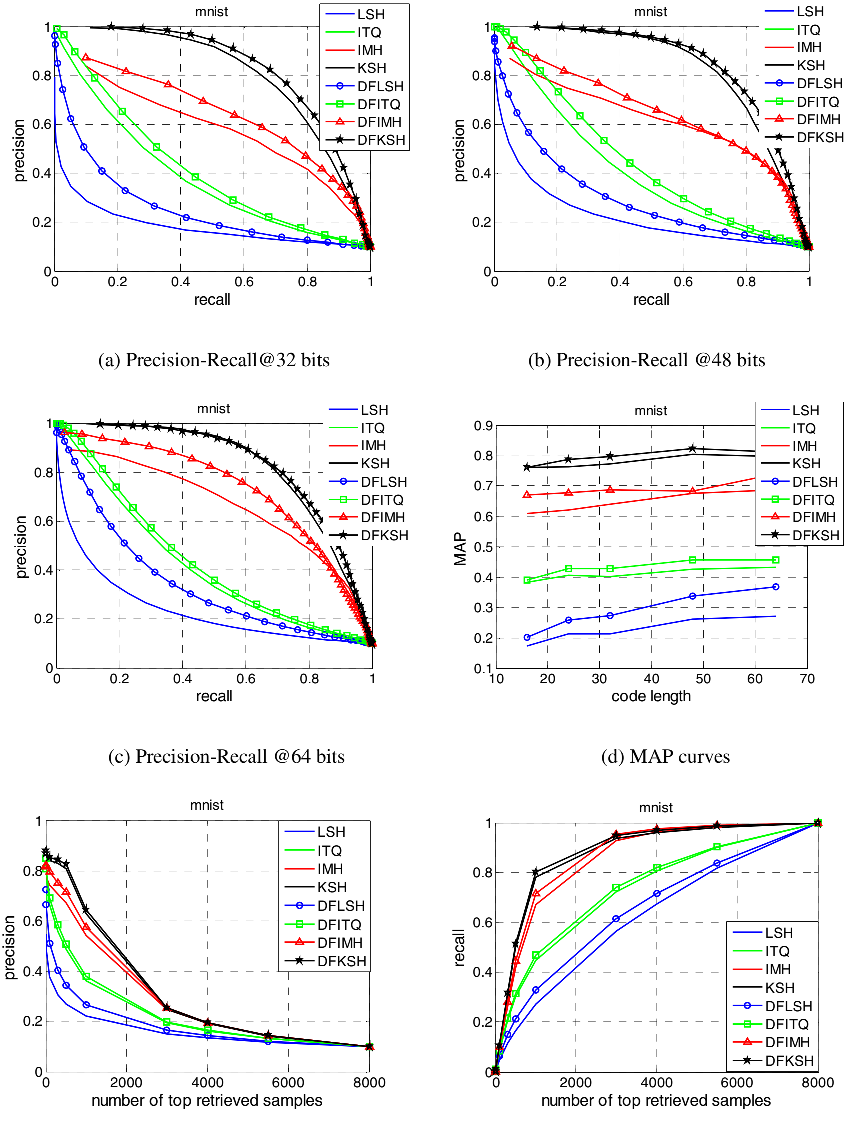
3 Experimental Results

We evaluate the effectiveness of hashing algorithm with discriminative features on CIFAR10 [13] and MNIST [14] two databases. The CIFAR10 database including 60000 color images and is divided into 10 categories with 6000 samples for each class. Each image is represented by a 512-dimensional GIST feature vector. The MNIST database consists of 70000 images of handwritten digits from ‘0’ to ‘9’, each digit image have 28×28=784 pixels.

Besides, we apply discriminative features to Locality Sensitive Hashing, Iterative Quantization, Supervised Hashing with Kernels and Inductive Hashing on Manifolds (DFLSH, DFITQ, DFIMH and DFKSH) and compare with original hash algorithms (LSH, ITQ, IMH, and KSH). For unsupervised hash algorithm LSH, ITQ and IMH, the original data points without any supervision information. In order to obtain the label information, we get the category information for each data point using K-means clustering algorithm.

3.1 MNIST

The precision-recall curves at 32 bits, 48 bits and 64 bits as shown in Fig. 2(a), Fig. 2(b) and Fig. 2(c) on MNIST, respectively. For unsupervised hash algorithm DFLSH, DFITQ and DFIMH gain the biggest can reach about 8%, 4%, 6% in precision-recall curves over the corresponding LSH, ITQ and IMH. For supervised hash algorithm DFKSH gain the biggest can reach about 4% in precision-recall curves over the corresponding KSH. Fig. 2(d) is the mean average precision(MAP) curves, as we can see from the figure, the MAP curves of hash algorithms with discriminative features DFLSH, DFITQ and DFIMH and DFKSH have about 3%, 5%, 3% and 7% over the LSH, ITQ and IMH and KSH have not discriminative features. Fig. 2(e) and Fig. 2(f) are the precision curves and the recall curves with the number of top retrieved samples at 32 bits respectively, we can see DFLSH, DFITQ and DFIMH and DFKSH performs better than the corresponding hashing approach have not discriminative features.



(a) Precision-Recall@32 bits

(b) Precision-Recall @48 bits

(c) Precision-Recall @64 bits

(d) MAP curves

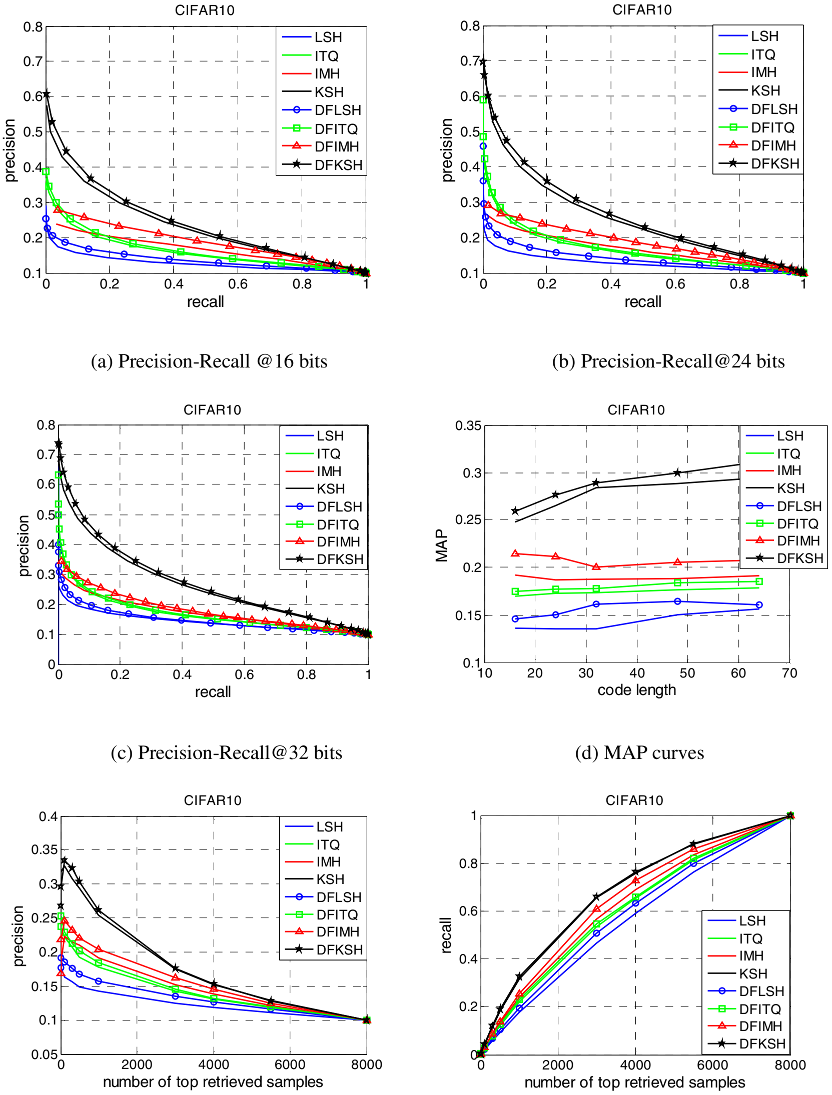
(e) Precision @32 bits

(f) Recall @32 bits

Fig. 2. The results on MNIST database

3.2 CIFAR-10

Fig. 3(a), Fig. 3(b) and Fig. 3(c) shows the evaluation results of precision-recall curves at 16 bits, 24 bits and 32 bits on CIFAR-10. The precision decreases when the recall increases and hashing approaches with discriminative features DFLSH, DFITQ, DFIMH and DFKSH outperforms associated with all hashing approaches have not



(a) Precision-Recall @16 bits

(b) Precision-Recall@24 bits

(c) Precision-Recall@32 bits

(d) MAP curves

(e) Precision@16 bits

(f) Recall@16 bits

Fig. 3. The results on CIFAR10 database

discriminative features LSH, ITQ, IMH and KSH. Fig. 3(d), Fig. 3(e) and Fig. 3(f) are the mean average precision(MAP) curves, the precision curves and the recall curves with the number of top retrieved samples at 16 bits respectively, as shown in the figure, DFLSH, DFITQ, DFIMH and DFKSH has much better performance than LSH, ITQ, IMH and KSH.

4 Conclusion

In this paper, we introduce an effective framework of signal-to-noise ratio to extract discriminative feature of the original data. We first construct signal and noise vectors and further showed that this construct is feasibility when distinguish information between the data of two classes. Moreover, our framework get discriminative features of original data through generalized eigen-decomposition are applied in the hash algorithm can obtain higher precision and recall with short binary codes. However, our techniques are not a panacea, due to the data will not discriminative when the directions are very similar for all classes, but it is very easy and simple to apply and understand. Experimental comparison showed that our framework apply to hash methods achieved very promising hashing performance over the original hashing methods. What's more, conditional second moment matrices has better discriminant feature extraction ability than three layer or multilayer algorithm, because the multilayer algorithm structure can cause over-fitting phenomenon, which lead to the extraction of low dimensional characteristics excessively depend on the training data, and losing its universal applicability.

Acknowledgement. This research was supported by the National Natural Science Foundation of China (61472304, 61125204, 61432014 and 61201294), Program for Changjiang Scholars and Innovative Research Team in University (No.IRT13088).

References

1. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proc. of ACM Symposium on Theory of Computing, pp. 604–613 (1998)
2. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: Proc. of 25th International Conference on Very Large Data Bases, pp. 518–529 (1999)
3. Datar, M., Immorlica, N., Indyk, P.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proc of ACM Symposium on Computational Geometry, pp. 253–262 (2004)
4. Gong, Y., Lazebnik, S.: Iterative quantization: a procrustean approach to learning binary codes. In: Proc of IEEE Conference on Computer Vision and Pattern Recognition, pp. 817–824 (2011)
5. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Trans. on Pattern Analysis and Machine Intelligence **33**(1), 117–128 (2011)

6. He, K., Wen, F., Sun, J.: K-means hashing: an affinity-preserving quantization method for learning binary compact codes. In: Proc of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2938–2945 (2013)
7. Shen, F., Shen, C., Shi, Q., Anton, H., Tang, Z.: Inductive hashing on manifolds. In: Proc of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1562–1569 (2013)
8. Wang, J., Sanjiv, K., Chang, S.: Supervised Hashing for Large Scale search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2393–2406 (2012)
9. Liu, W., Wang, J., Ji, R., Jiang, Y., Chang, S.: Supervised hashing with kernels. In: Proc of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2074–2081 (2012)
10. Svante, W., Michael, S.: SIMCA: a method for analyzing chemical data in terms of similarity and analogy. In: Chemometrics: Theory and Application, vol. 52, pp. 243–282 (1977)
11. Nikos, K., Paul, M.: Discriminative features via generalized eigenvectors. In: Proc of International Conference on Machine Learning, pp. 494–502 (2014)
12. John, C., Toutanova, K., Yih, W.T.: Translingual document representations from discriminative projections. In: Proc of Conference on Empirical Methods in Natural Language Processing, pp. 251–261. Association for Computational Linguistics (2010)
13. http://www.cs.toronto.edu/~kriz/cifar.html
14. http://yann.lecun.com/exdb/mnist

Image Quality Assessment Based on Local Pixel Correlation

Hongqiang Xu, Wen Lu(✉), Yuling Ren, and Lihuo He

School of Electronic Engineering, Xidian University, Xi'an 710071, China
{xuhongqiang,renyuling}@stu.xidian.edu.cn,
{luwen,lhhe}@mail.xidian.edu.cn

Abstract. The available image quality assessment methods are mostly based on statistical characteristic and consider very little the change of pixel correlation in conjunction with the quality assessment, which induces the quality assessment metric to be limited in the degradation of image quality caused by the change of pixel correlation. However, the pixel correlation change has a big effect on the image quality, so a novel image quality assessment based on the pixel correlation is proposed in this paper. Firstly image is parted based on mutual information, and then, three kinds of mutual information between the pixel intensity and the image patches are extracted to catch the variation of the pixel correlation. Finally the machine learning is utilized to learn the mapping from these differences space to image quality. The experimental results show that the proposed framework has good consistency with subjective perception.

Keywords: Image quality assessment · Mutual information · Pixel correlation

1 Introduction

Image acts as the carrier of information, which conveys the vital information to everyone and becomes a ubiquitous part of modern life. However, the impairment of image effects the substantial information contained in the image, which will bring down the satisfaction of human perceived. It is essential to build image quality evaluation metrics for various image applications [1-2]. Subjective methods perceive image quality by many participators, which is expensive and time consuming. So we move to objective measurements which accomplish the image quality assessment task automatically.

The state-of-the-art image quality assessment methods can be divided into two broad classes in which kind of information it used. The first is the image pixel domain based paradigm, where the pixel value changes between the reference and distorted signals is predicted as the image quality. Mean Squared Error [3] and Peak Signal-to-Noise Ratio are the most widely used objective quality metrics due to their convenience and clear physical meaning. Zhou Wang et al. [4] propose a method based on Structural Similarity which measures the structure variation between the reference and distorted image. Corresponding to the pixel domain based methods, are the image

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 266–275, 2015.
DOI: 10.1007/978-3-662-48570-5_26

transform domain based methods, where the transform domain coefficients information error between the reference and distorted signals is predicted as the image quality. Sheikh et al. [5] proposed a method named Visual information fidelity based on the Gaussian scale mixtures in the wavelet domain. Wang et al. [6] propose a method using a natural image statistic model in the wavelet domain, which measures the wavelet coefficients histogram difference between the reference and distorted signals to get image quality score. Ding et al. [7] propose a method using mutual information of Gabor features.

The statistics is widely used in both pixel domain and transform domain. The statistical model parameters provide a good approximation to pixel difference, but not to the weakening of pixel correlation. Three kinds of information of a pixel in different patch were introduced by Rigau et al. [8], which are sensitive to the changes of image pixel values and the correlation between pixels. The first component is the specific information to express how much image content information is conveyed in a particular pixel. The second component is defined as saliency information to express how significant the pixel is. The last one is the entanglement information, which can evaluate the pixel correlation information in a region. Based on the proposed information, a new IQA algorithm is proposed. First, a non-overlapping segmentation set is acquired to build the relation with image pixels. Then based on the mutual information of the pixels, the specific information of a pixel in different patch are extracted and the mean and variance of it are calculated to catch the pixel value changes, then accordingly the saliency and entanglement information of a pixel in different patch were measured to catch the pixel correlation change, finally these differences are mapped to the image quality

The remainder of the paper is organized as follows. In Section 2 presents the details of the proposed IQA algorithm framework. Experimental results are presented in Section 3, and Section 4 concludes.

2 The Proposed Image Quality Assessment

In this section, we propose an image pixel domain based image quality assessment algorithm after building the relation between the image pixels and the image patches. First, a non-overlapping segmentation set is acquired to build the relation with image pixels. Then based on the mutual information of the pixels and the patches, the specific information of a pixel in different patch and calculated the mean and variance of it to catch the pixel value changes, then accordingly the saliency and entanglement information of a pixel in different patch were measured to catch the pixel correlation change, finally these differences are mapped to the image quality. The algorithm framework shown in Fig. 1:

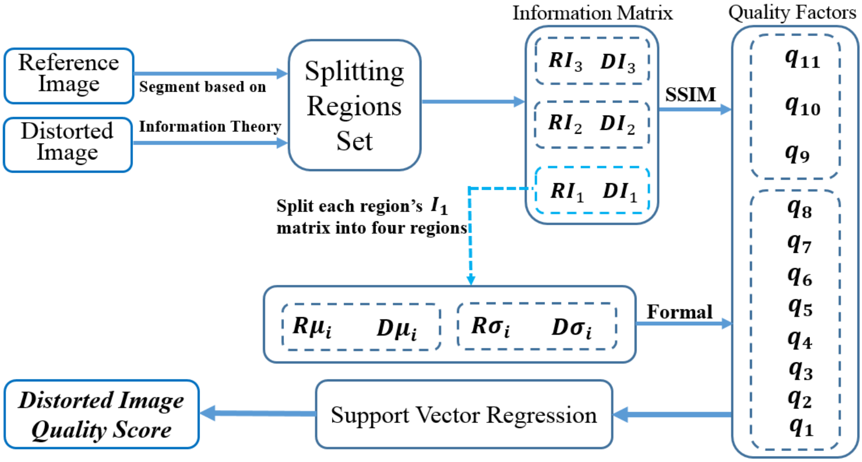


Fig. 1. The proposed image quality assessment algorithm framework

2.1 Image Partition

Composition and luminosity are the most basic two elements of a photograph. In order to build their correlation, an information channel between the luminance histogram and the regions of the image is established by image partition, here, the image partition by the method introduced by Rigau et al. [9]. The portioning algorithm progressively splits the image by extracting the maximum information at each step. In the partition procedure, it takes image luminance histogram as the input and the set of regions *R* of the image as outputs and this process is defined as the information channel *B*→*R* and corresponding the information channel *R*→*B* is defined by taking the set of regions *R* of the image as the input. Here use the Fig. 2 being a description and use the Fig. 3 to explain the relation between two information channels.

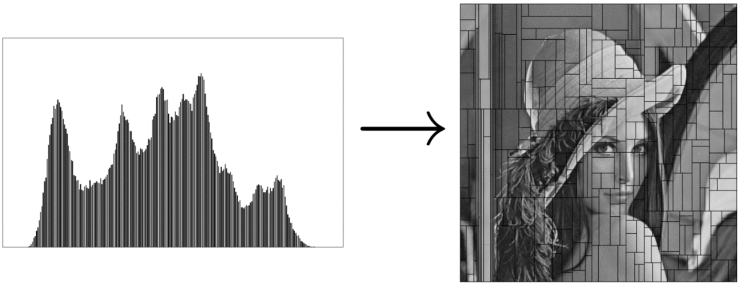


Fig. 2. The information channel *B*→*R*, the left is input and the right is output

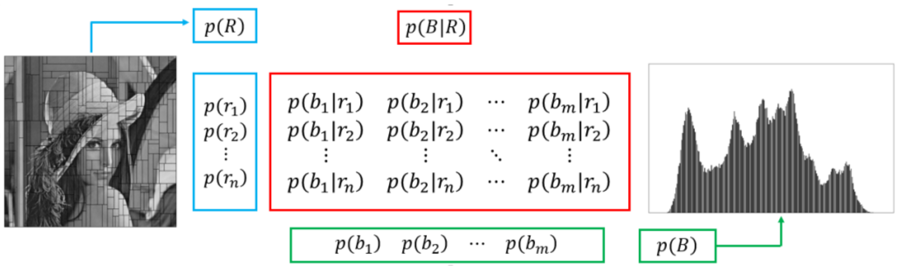

$p(R)$
$p(B|R)$
$p(r_1)$ $p(r_2)$ ⋮ $p(r_n)$
$p(b_1|r_1)$ $p(b_2|r_1)$ ⋯ $p(b_m|r_1)$
$p(b_1|r_2)$ $p(b_2|r_2)$ ⋯ $p(b_m|r_2)$
⋮ ⋮ ⋱ ⋮
$p(b_1|r_n)$ $p(b_2|r_n)$ ⋯ $p(b_m|r_n)$
$p(b_1)$ $p(b_2)$ ⋯ $p(b_m)$
$p(B)$

Fig. 3. The transformation relation between two information channels.

2.2 Measure the Variation of Pixel Difference

Given an image I of N pixels, where $R=\{r_1,r_2,\dots,r_t\}$ represent the set of region R with the number of the regions is t, and $B=\{b_1,\dots,b_i\}$, $i=0,1,\dots,255$ represent the set of image pixel value, which b_i means a pixel value is i. N_b is the frequency of bin b ($N{=}\sum_{b\in B}N_b$) and N_r is the number of pixels of region r ($N=\sum_{r\in R}N_r$), according to the mutual information (MI) formula, the MI between B and R is given by

$$
\begin{aligned}
I(B;R) &= \sum_{b\in B} p(b) \sum_{r\in R} p(\mathrm{r}\,|\,b)\log\frac{p(r\,|\,b)}{p(r)} \\
&= H(R)-H(R\,|\,B) \\
&= H(B)-H(B\,|\,R)
\end{aligned}
\tag{1}
$$

Where $p(b)=N_b/N$, $p(r\,|\,b)=N_{b,r}/N_b$ and $p(r)=N_r/N$ in (1), and the MI represents the shared information or correlation between B and R.

After dividing the image, the set of regions R, the information channel B→R and R→B are got. We defined an information of a pixel in different patch which is sensitive to the changes of image pixel values which have been introduced in [9]. Specific process as following:

The Specific Information I_1 : From (1), in the information channel B→R, mutual information can be expressed as

$$
I(B;R)=H(R)-H(R\,|\,B)=\sum_{b\in B}p(b)[H(R)-H(R\,|\,b)]=\sum_{b\in B}p(b)I_1(b;R) \tag{2}
$$

Where

$$
I_1(b;R)=H(R)-H(R\,|\,b)=-\sum_{r\in R}p(r)\log p(r)+\sum_{r\in R}p(r\,|\,b)\log p(r\,|\,b) \tag{3}
$$

Following a similar process for the reversed channel R→B, the specific information associated with region r is given by

$$I_1(r;B) = H(B) - H(B \mid r) = -\sum_{b \in B} p(b) \log p(b) + \sum_{b \in B} p(b \mid r) \log p(b \mid r) \tag{4}$$

Then we measure the mean and variance of the specific information in the patches to catch the pixel value changes quality factors q_i where i = {1, 2, … , 8}. Specific process as following: Firstly, we need to divide a region of the image into four regions averagely, and then calculate the mean of their specific information μ_i and the variance of their specific information σ_i where i = {1, 2, 3, 4}. The splitting process is shown in Fig. 4.

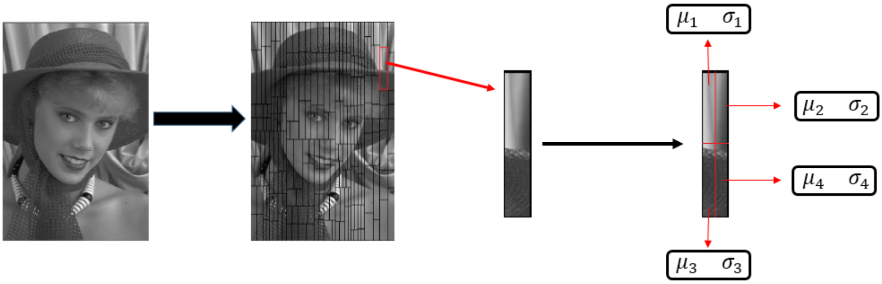


Fig. 4. The process of dividing a region into four and statist the mean and variance.

For the region r_j of the image, we can get the mean of its specific information μ_{ij} and the variances σ_{ij} where i={1, 2, 3, 4} and j={1, 2, … , n}. The mean and variance of the reference image is

$$\begin{aligned} R\mu_i &= [R\mu_{i,1}, R\mu_{i,2}, \ldots, R\mu_{i,n}] \quad i = 1,2,3,4 \\ R\sigma_i &= [R\sigma_{i,1}, R\sigma_{i,2}, \ldots, R\sigma_{i,n}] \quad i = 1,2,3,4 \end{aligned} \tag{9}$$

And the mean and variance of the distorted image is

$$\begin{aligned} D\mu_i &= [D\mu_{i,1}, D\mu_{i,2}, \ldots, D\mu_{i,n}] \quad i = 1,2,3,4 \\ D\sigma_i &= [D\sigma_{i,1}, D\sigma_{i,2}, \ldots, D\sigma_{i,n}] \quad i = 1,2,3,4 \end{aligned} \tag{10}$$

The quality factors can been defined as

$$q_t = \frac{2 \times (R\mu_t \bullet D\mu_t) + 1}{n \times (R\mu_t \bullet R\mu_t + D\mu_t \bullet D\mu_t + 1)} \quad where\ t = 1,2,3,4 \tag{11}$$

$$q_s = \frac{2 \times (R\sigma_{s-4} \bullet D\sigma_{s-4}) + 1}{n \times (R\sigma_{s-4} \bullet R\sigma_{s-4} + D\sigma_{s-4} \bullet D\sigma_{s-4} + 1)} \quad where\ t = 5,6,7,8 \tag{12}$$

2.3 Measure the Variation of Local Pixel Correlation

For getting a good approximation to the weakening of pixel correlation, we defined two information of a pixel in different patch which is sensitive to and the correlation between pixels. Specifically, the first component is defined as saliency information to express how significant the pixel is. The second one is the entanglement information, which can evaluate the pixel correlation. Then we use the three information defined in this paper to measure the image pixel correlation changes. The saliency information and the entanglement information are defined as following:

The Saliency Information I_2 **:** From (1), the MI between luminance and regions can be expressed as

$$I(B;R)=\sum_{b\in B}p(b)\sum_{r\in R}p(\mathrm{r}\,|\,b)\log\frac{p(r\,|\,b)}{p(r)}=\sum_{b\in B}p(b)I_2(b;R) \tag{5}$$

Where

$$I_2(b;R)=\sum_{r\in R}p(r\,|\,b)\log\frac{p(r\,|\,b)}{p(r)} \tag{6}$$

as the surprise associated with the luminance *b* and can be interpreted as a measure of its saliency. High values of $I_2(b;R)$ express a high surprise and identify the most salient luminance.

The Entanglement Information I_3 **:** From (1), in the information channel *B*→*R*, the entanglement information is defined as

$$I_3(b;R)=\sum_{r\in R}p(r\,|\,b)I_1(r;B) \tag{7}$$

A large value of $I_3(b;R)$ means that the specific information $I_1(r;B)$ of the regions that contain the luminance *b* are very informative.

After transforming image into the three information above, we can get three information matrixes which have the same size as the image, respectively are the saliency information matrix, the specific information matrix and the entanglement information matrix RI_i from the reference image, and DI_i from the distorted image where i=1,2,3. As it has been said, the essence of image quality descended is the loss of visual perceptive information, the more the image quality descended, the lesser the visual perceptive information kept, that means that we can measure image quality by calculating image information matrix discrepancy between RI_i and DI_i. Here, we take RI_i and DI_i. as the input of the SSIM model and take the output of SSIM as quality factor q_i. where i=1,2,3. We can represent it as

$$q_{i+8}=SSIM(RI_i,DI_i)\quad i=1,2,3 \tag{8}$$

After finding out the image quality factors q_i where i={1, 2, … ,11}, Support Vector Regression (SVR) is used to learn the mapping from quality factor space to image quality. Since each image is represented by a single quality factor vector, the image quality assessment problem can be solved as a regression problem. A lot of regression techniques such as SVR and random forest can be used to learn the mapping. Here we use SVR with RBF kernel to do the regression.

3 Experimental Results

In this section, we compare the performance of the proposed framework with standard RR-IQA methods, i.e., RRVIF [10], FEDM [11], WNISM [12], RRED [13], LBPs [14], RR-PCA [15] and RRIQV [16], based on the following experiments: the consistency experiment, the rationality experiment, and the influence of quality factor's number on the final experimental result experiment. At the beginning of this section, we first brief the image database for evaluation.

The laboratory for image & video engineering (LIVE) database [17] has been recognized as the standard database for IQA measures performance evaluation. This database contains 29 high-resolution 24 bits/pixel RGB color images and 175 corresponding JPEG and 169 JPEG2000 compressed images, as well as 145white noisy (WN), 145 Gaussian blurred (GB), and 145 fast-fading (FF) Rayleigh channel noisy images at a range of quality levels.

3.1 Consistency Experiment

In this subsection, we compare the performance of the proposed IQA framework with RRVIF, FEDM, WNISM, RRED, LBPs, RR-PCA and RRIQV. The PLCC and RMSE results for all IQA methods being compared are given as benchmark in Table 1 and Table 2.

Table 1. PLCC values the newest RR methods and the proposed method on whole LIVE Database and five data sets of different distortion categories

		JP2K	**JPEG**	**WN**	**GB**	**FF**	**ALL**
PLCC	RRVIF	0.932	0.895	0.957	0.955	0.944	0.725
	FEDM	0.921	0.875	0.925	0.902	0.875	---
	WNISM	0.924	0.876	0.890	0.888	0.925	0.710
	RRED	0.930	0.831	0.926	0.953	0.922	0.764
	LBPs	0.927	0.894	**0.990**	0.966	**0.950**	0.935
	RR-PCA	0.934	0.904	0.980	0.777	0.923	---
	RRIQV	0.944	0.909	0.886	0.882	0.919	---
	Proposed	**0.950**	**0.967**	0.971	**0.970**	0.948	**0.954**

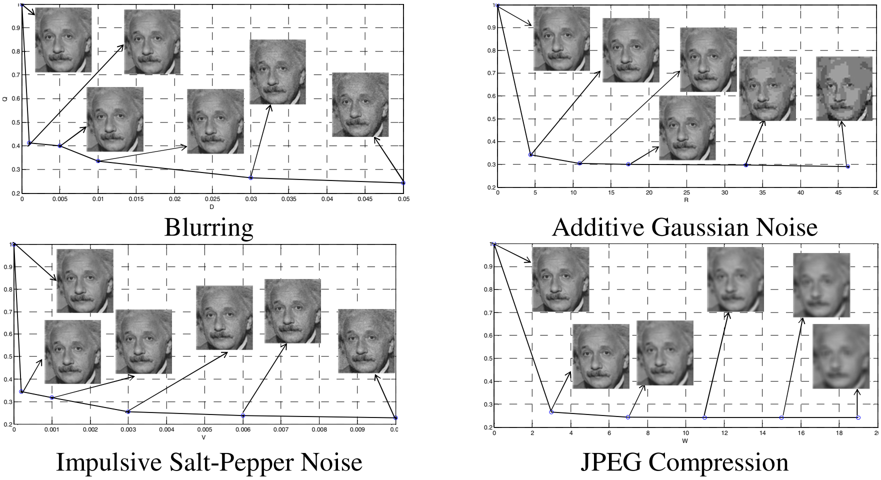
Table 2. RMSE values the newest RR methods and the proposed method on whole LIVE Database and five data sets of different distortion categories

		JP2K	**JPEG**	**WN**	**GB**	**FF**	**ALL**
RMSE	RRVIF	**5.88**	7.14	4.65	4.66	**5.42**	17.1
	FEDM	6.31	7.73	6.06	6.78	7.96	---
	WNISM	6.17	7.71	7.28	7.22	6.24	18.4
	RRED	9.28	17.7	10.5	5.62	11.3	17.6
	LBPs	5.97	6.94	**2.25**	**3.98**	4.86	9.59
	RR-PCA	9.38	17.6	5.33	13.8	14.7	---
	RRIQV	9.26	15.8	14.0	8.22	11.0	---
	Proposed	6.63	**5.72**	5.56	4.68	7.10	**5.88**

3.2 Rationality Experiment

To verify the rationality of the proposed framework, we choose the Einstein image with different distortions, which are blurring (with smoothing window of *W*×*W*), additive Gaussian noise (mean=0, variance=*V*), impulsive Salt-Pepper noise (density=*D*), and JPEG compression (compression rata=*R*).

Figs. 5 (all of the images are 8 bits/pixel; cropped from 512 × 512 to 128 × 128 for visibility) shows the Einstein image with different types of distortions and the metrics prediction trend to the corresponding image, respectively. It is found that the proposed framework prediction trend to image drop with the increasing intensity of different types of image distortions. It is consistent well with the tendency of the decreasing image quality in fact. So the results demonstrate the rationality of the proposed framework.



Blurring

Additive Gaussian Noise

Impulsive Salt-Pepper Noise

JPEG Compression

Fig. 5. Trend plots of Einstein with different types of distortion using the proposed framework.

3.3 The Influence of Quality Factor's Number on the Final Experimental Result Experiment

The perception changes caused by image quality descended not only reflect on each pixel of the image, also on a higher image scale. For catching the perception changes in a higher image scale, we design other eight quality factors q_4 .... q_{11} . Here we show that these quality factors do have positive influence on the final image quality result. When we use Support Vector Regression (SVR) to learn the mapping from quality factor space to image quality, if we changed the number of the quality factors, it would have a different image quality result, the more the quality we put in, the higher the image quality score is. This can been shown in Table 4.

Table 3. The Influence of Quality Factor's Number on the Final Experimental Result

		JP2K	JPEG	WN	GB	FF	ALL
PLCC	$q_1 \sim q_3$	0.954	0.927	0.961	0.967	**0.961**	0.924
	$q_1 \sim q_7$	**0.960**	0.963	0.96.	0.965	0.940	0.952
	$q_1 \sim q_{11}$	0.950	**0.967**	**0.971**	**0.970**	0.948	**0.954**

4 Conclusions

A novel general image quality assessment is proposed in image pixel domain. An information channel is established by a mutual information based image partition, which is aimed to analyze the correlation between the luminance histogram and the composition of the image, In this channel, three kinds of mutual information between the pixel intensity and the image patches are extracted to catch the variation of the pixel correlation. Then the differences between these information are utilized to learn the mapping from the differences space to image quality. Experimental results illustrate that the proposed framework have good consistency with subjective perception values and the objective assessment results can well reflect the visual quality of images.

Acknowledgements. This research was supported partially by the National Natural Science Foundation of China (No. 61125204, No.61372130, No.61432014), the Fundamental Research Funds for the Central Universities (No. BDY081426, No.JB140214, No. XJS14042), the Program for New Scientific and Technological Star of Shaanxi Province (No.2014KJXX-47), the Project Funded by China Postdoctoral Science Foundation (No. 2014M562378).

References

1. Wang, Z., Bovik, A.C.: Modern image quality assessment. Synthesis Lectures on Image, Video, and Multimedia Processing **2**(1), 1–156 (2006)
2. Park, H., Har, D.H.: Subjective image quality assessment based on objective image quality measurement factors. IEEE Trans. Consumer Electronics **57**(3), 1176–1184 (2011)

3. http://en.wikipedia.org/wiki/Information_theory
4. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing **13**(4), 600–612 (2004)
5. Sheikh, H.R., Bovik, A.C.: A visual information fidelity approach to video quality assessment. In: The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, pp. 23–25. IEEE (2005)
6. Wang, Z., Simoncelli, E.P.: Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In: Electronic Imaging 2005. International Society for Optics and Photonics, pp. 149–159 (2005)
7. Ding, Y., Zhang, Y., Wang, X., et al.: Perceptual image quality assessment metric using mutual information of Gabor features. Science China Information Sciences **57**(3), 1–9 (2014)
8. Rigau, J., Feixas, M., Sbert, M.: Image information in digital photography. In: Koch, R., Huang, F. (eds.) ACCV 2010 Workshops, Part II. LNCS, vol. 6469, pp. 122–131. Springer, Heidelberg (2011)
9. Rigau, J., Feixas, M., Sbert, M.: An information theoretic framework for image segmentation. In: International Conference on Image Processing, pp. 1193–1196. IEEE (2004)
10. Wu, J., Lin, W., Shi, G., Liu, A.: Reduced-reference image quality assessment with visual information fidelity. IEEE Trans. Multimedia **15**(7), 1700–1705 (2013)
11. Zhai, G., Wu, X., Yang, X., et al.: A psychovisual quality metric in free-energy principle. IEEE Trans. Image Processing **21**(1), 41–52 (2012)
12. Wang, Z., Simoncelli, E.P.: Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In: Electronic Imaging 2005. International Society for Optics and Photonics, pp. 149–159. IEEE (2005)
13. Soundararajan, R., Bovik, A.C.: RRED indices: Reduced reference entropic differencing for image quality assessment. IEEE Trans. Image Processing **21**(2), 517–526 (2012)
14. Wu, J., Lin, W., Shi, G., et al.: Reduced-reference image quality assessment with local binary structural pattern. In: 2014 IEEE International Symposium on Circuits and Systems, pp. 898–901. IEEE (2014)
15. Uzair, M., Fayek, D.: Reduced reference image quality assessment using principal component analysis. In: IEEE International Symposium on Broadband Multimedia Systems and Broadcasting 2011, pp. 1–6. IEEE (2011)
16. Wang, Z., Simoncelli, E.P.: Reduced-reference image quality assessment using a wavelet domain natural image statistic model. In: Electronic Imaging 2005. International Society for Optics and Photonics, pp. 149–159 (2005)
17. Sheikh, H.R., Wang, Z., Cormack, L., et al.: LIVE Image Quality Assessment Database. http://live.ece.utexas.edu/research/quality/

Fast Image Quality Assessment via Hash Code

Lihuo He(✉), Qi Liu, Di Wang, and Wen Lu

School of Electronic Engineering, Xidian University, Xi'an 710071, China
{lihuo.he,qliu625,wangdi.wandy}@gmail.com, luwen@mail.xidian.edu.cn

Abstract. No-reference image quality assessment (NR-IQA) is significant for image processing and yet very challenging, especially for real-time application and big image data processing. Traditional NR-IQA metrics usually train complex models such as support vector machine, neural network, and probability graph, which result in long training and testing time and lack robustness. Hence, this paper proposed a novel no-reference image quality via hash code (NRHC). First, the image is divided into some overlapped patches and the features of blind/ referenceless image spatial quality evaluator (BRISQUE) are extracted for each patch. Then the features are encoded to produce binary hash codes via an improved iterative quantization (IITQ) method. Finally, comparing the hash codes of the test image with those of the original undistorted images, the final image quality can be obtained. Thorough experiments on standard databases, e.g. LIVE II, show that the proposed NRHC obtains promising performance for NR-IQA. And it has high computational efficiency and robustness for different databases and different distortions.

Keywords: No-reference · Image quality assessment · Hash code

1 Introduction

With the tremendous development of intelligent network, ultra-high resolution display, and wearable devices, high quality and credible visual information (image, video, etc.) is significant for the end user to obtain a satisfactory quality of experience (QoE). Where, assessing the quality of visual information, especially no-reference or blind image quality assessment (NR-IQA or BIQA), plays an important role in numerous visual information processing system and applications [1]. Moreover, effective (high quality prediction accuracy) and efficient (low computational complexity) NR-IQA is essential and has attracted a large number of attentions.

NR-IQA metric is designed to automatically and accurately predict image quality without reference images. Hence, it is a difficult and challenging work and has attracted many researchers' attentions. Traditional methods focus on designing distortion-specific methods [2]-[4], which means that these methods evaluate images with only one kind of distortions effectively, such as JPEG compression, JPEG2000 compression, white noise, and Gaussian blurring. Therefore,

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 276–285, 2015.
DOI: 10.1007/978-3-662-48570-5_27

it is imperative to build the general purpose NR-IQA metric to handle different types of distortions and even multi-distortions.

Recently, great effort has been made to design general purpose NR-IQA metrics. A series of methods are presented in the literature [5]-[18]. Almost all of the reported NR-IQA methods include quality-aware feature extraction and effective evaluation model designing, which are the most important processing for building a NR-IQA method. Generally, natural scene statistical (NSS) properties [19] are most popular utilized features, which are extracted by generalized Gaussian distribution in wavelet domain usually. Also other features are extracted through Gabor in spatial domain [11] or statistical characteristics in discrete cosine transformation (DCT) domain [10]. Another key point is designing the prediction model. The latest methods can be divided into two categories, two-steps strategy and transductive approach. The former first determines the distortion type of a test image and then employs an associated distortion-specific no-reference image quality assessment metric to predict the quality of the given image, e.g. BIQI [5] and DIIVINE [6]. The BIQI trains a support vector machine (SVM) model to divide five different distortions and trains five different support vector regressions (SVR) model for a particular distortion to predict image quality. The DIIVINE, which is the extended work of BIQI, also is built in the two-steps framework. While the transductive approach aims to build a model to directly map image features to image quality without distinguishing different types of distortions, such as LBIQ [7], BLINDS [8], BLINDS-II [10], CORNIA [11], and SRNSS [12]. In those metrics, a large number of machine learning methods are utilized to train the quality prediction model, such as multiple kernels learning (MKL), neural network (NN), and the probabilistic model. Therefore, the reported metrics face a significant problem that they need long training and test time. This is because that complex machine learning model is adopted, the parameters are mostly defined by experience, and a large number of samples are utilized to train the prediction model. And these methods also would reduce the robustness of the quality evaluation system.

In order to solve the above problems, this paper proposed a novel no-reference image quality assessment metric via hash codes, which is simple yet very fast. The proposed method first divides the image into overlapped patches and extracts the blind/referenceless image spatial quality evaluator (BRISQUE) [9] features for each patch. Then the features are encoded into hash codes via an improved iterative quantization (IITQ) [20] method. The Hamming distance between the hash code of the test image and the original undistorted image is calculated to predict image quality. In the proposed method, the quality prediction includes hash coding and the Hamming distance calculation. They have the properties of fast speed and high efficiency. Hence, the proposed can satisfy the real-time applications and big image data processing.

The rest of the paper is organized as follows. Section 2 illustrates the proposed no-reference image quality assessment. Detailed experimental results are summarized and discussed in Section 3, and section 4 concludes the paper.

2 NR-IQA via Hash Code

In order to assess the image quality effectively and efficiently, a novel no-reference image quality assessment method is presented in the paper. The proposed method includes three major steps: feature extraction, hashing coding, and quality evaluation. For convenience, the proposed metric is named **NRHC**, which is short for fast **N**o-**R**eference image quality assessment via **H**ash **C**ode. And the framework of proposed NRHC is shown in figure 1.

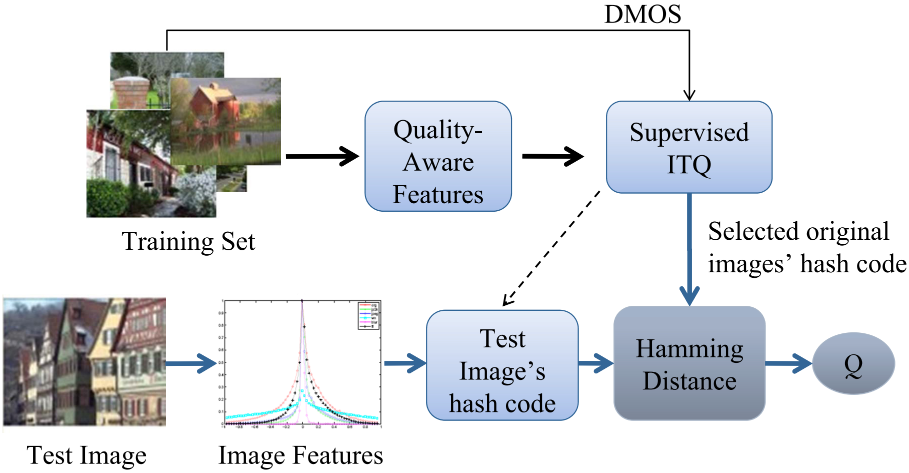


Fig. 1. The framework of the proposed NRHC.

2.1 Features Extraction

Generally, the image has some statistical properties [19], especially for natural scene images. The natural scene statistic is sensitivity to the presence of distortions, such as JPEG2000 compression. Hence, quantifying deviations from the normal natural scene statistics can assess the image quality, which is sufficient to quantify naturalness of the image. However, the NSS feature is extracted from the coefficient of some transform domain, such as DCT, wavelet and contourlet. And it has a defect that time of transformation is huge. Hence, this paper introduces the NSS properties into the BRISQUE [9] to describe the naturalness of images.

Let $\mathbf{X}$ denotes the training set of images including n images. First, every image in the training set is divided into overlapped patches with the size $\omega \times \omega$ and an over-lapping size of ω_0 pixels between neighboring patches. Then the local BRISQUE features are extracted for every patch. The local patch is processed by local mean subtraction and variance normalization on log-contrast values to produces decorrelated coefficients which follow Gaussian like distributions. Given a local patch $x_{l,k}$ (l-th image and k-th patch), the process is as follows

$$\hat{x}_{l,k}(i,j) = \frac{x_{l,k}(i,j) - \mu_{l,k}(i,j)}{\sigma_{l,k}(i,j) + \sigma_0}. \tag{1}$$

Where, i and j are the spatial indices in local patch. σ_0 is a constant that is used to prevent instabilities. μ and σ are the mean and variance with a Gaussian filter. The normalized luminance coefficients obey the asymmetric generalized Gaussian distribution (AGGD). And all the parameters are estimated for the AGGD with zero mode

$$p_X(x;\alpha,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{\alpha}{(\beta_l+\beta_r)\gamma(1/\alpha)} \exp(-(\frac{-x}{\beta_l})^{\alpha}) & x < 0 \\ \frac{\alpha}{(\beta_l+\beta_r)\gamma(1/\alpha)} \exp(-(\frac{-x}{\beta_r})^{\alpha}) & x \geq 0 \end{cases} \tag{2}$$

Let $\mathbf{f}_l = (f_{l,1}, \cdots, f_{l,k_l})$ represents the feature set of the l-th image. k_l is the number of patches for the l-th image. $f_{l,k}$ is the BRISQUE feature vector of the k-th patch in the l-th image. In this paper, the dimension of the feature is thirty-six for every patch.

All the features extracted from the training images are then clustered into c classes using the k-means clustering algorithm with the squared Euclidean distance metric. The cluster centers are used as the "quality-aware" visual words. Every patch is assigned to the nearest cluster center by vector quantization, and an empirical distribution over the visual words is calculated for each training image, denoted as $\mathbf{s}_l$ for every image.

2.2 Hashing Coding

The BRISQUE statistical features have a weak positive correlation with the differential mean opinion scores (DMOS) which are produced human evaluation. To get more accuracy prediction quality, it needs to design the evaluation model and learning algorithm. Most existing models are built through complex machine learning methods. Aiming to design as efficient and fast quality assessment model, the iterative quantization [20] hash code algorithm is adopted in this paper because of the fast properties. However, ITQ is an unsupervised method. To obtain accuracy image quality, ITQ is improved (named IITQ) with the supervised information (e.g. DMOS) to learning similarity binary codes for images with similarity DMOS.

Training
Given the training image set $S = [s_1, s_2, \cdots, s_n] \in \mathbb{R}^{c\times n}$ and corresponding DMOS $\{q_1, q_2, \cdots, q_n\}$, our goal is to learn a binary code matrix $B \in \{-1, 1\}^{n\times d}$, where d denotes the code length. For each bit r, the binary encoding function is built by h_r. Combining all bits, we can get hash functions $H = \{h_1, h_2, \cdots, h_d\}$. Following [19], we will apply linear dimensionality reduction to data, and then perform improved binary quantization in the resulting space. First, the PCA

projection matrix $W \in \mathbb{R}^{c \times d}$ is utilized to maximum the variance as follows

$$\begin{aligned} &\max F(W) = \tfrac{1}{n} tr\left(W^T S S^T W\right) \\ &s.t.\ W^T W = I. \end{aligned} \tag{3}$$

Where, maximizing the objective function F is a typical Eigen-problem which can be easily and efficiently solved by computing the eigenvectors of SS^T corresponding to the largest d eigenvalues. After getting PCA projection vectors W, we can get the low-dimensional embeddings $V = S^T W$. Then we orthogonally transform the projected data by an orthogonal rotation matrix $C \in \mathbb{R}^{d \times d}$ to minimize the quantization loss

$$\begin{aligned} &\min Q\left(B, C\right) = \|B - VC\|_F^2 \\ &s.t.\ C^T C = CC^T = I. \end{aligned} \tag{4}$$

Meanwhile, we intend to make the Hamming distance between a pair of binary codes preserving the DMOS difference of two images, that is

$$\min \delta\left(C\right) = tr(\mathrm{sgn}(VC)^T \tilde{S} \mathrm{sgn}(VC)), \tag{5}$$

where $\tilde{S} \in \mathbb{R}^{n \times n}$ is the difference matrix of DMOS. As the Eq. (5) is difficult to solve, we approximate binary codes B and rotation matrix C simultaneously

$$\min \tilde{\delta}\left(B, C\right) = tr(\mathrm{sgn}(VC)^T \tilde{S} VC) = tr(B^T \tilde{S} VC). \tag{6}$$

Then the overall objective function for minimizing the quantization loss and preserving the difference of DMOS simultaneously would be

$$\begin{aligned} &\min L(B, C) = \|B - VC\|_F^2 + 2\eta tr(B^T \tilde{S} VC) \\ &s.t.\ C^T C = CC^T = I. \end{aligned} \tag{7}$$

Where, η is a scaling parameter to balance the two contributions. Although the problem is NP-hard, its sub-problems w.r.t. each of B and C are convex. Therefore, we can minimize it by the alternating procedure.

Testing
Given a test image, we can obtain the features $s_t \in \mathbb{R}^{c \times 1}$. Then, its hash code can be obtained by

$$b_t = \mathrm{sgn}(s_t^T W C). \tag{8}$$

2.3 Quality Evaluation

For a new image y, we extract its features and quantify those features to get statistic properties of empirical distribution over the “quality-aware” visual words. Then the features are coded through the previous testing processing and we can obtain the hash code of the new image, indicated as b_t.

The original images from the training set are selected, indicated as b_i $(r = 1, \cdots, r)$, where r is the number of original images in the training set. Finally, the image quality can be obtained as follows

$$Q = \frac{100}{d}\frac{1}{r}\sum_{i=1}^{r}\text{Hamming}(b_i,b_t). \tag{9}$$

Where, Hamming(b_i, b_t) denotes the Hamming distance between b_i and b_t. And d is the number of hash code bits.

3 Experimental Results and Analysis

To validate the effectiveness and robustness of the proposed NR-IQA method, some experiments are conducted, including the consistency experiment, database independence, and time cost experiment.

Databases: LIVE II [23], TID [24], CSIQ [25], IVC [26], and MICT [27] are used as the standard databases. The LIVE (the Laboratory of Image and Video Engineering at the University of TEXAS at Austin) II database is the most popular adopted database and it is used as the benchmark database. It contains 29 high-resolution 24-bits/pixel RGB color original images and a series of distorted images (#982): JPEG2000 compression (JP2K, #227), JPEG compression (JPEG, #233), white noised in the RGB components (WN, #174), Gaussian blurring (Gblur, #174) and transmission error in the JPEG2000 bit stream using a fast-fading Rayleigh channel (FF, #174). All the images are presented with differential mean opinion scores.

Criterion: Video quality expert group (VQEG) [22] provides the comparison criterion in Phase-I and -II. A nonlinear mapping is first built between the predicted quality and DMOS using logistic non-linear regression analysis. And the criteria of LCC and SROCC are used to compare the performance of metrics. The Pearson linear correlation coefficient (**LCC**) provides an evaluation of prediction accuracy. The Spearman rank-order correlation coefficient (**SROCC**) is considered as a measure of prediction monotonicity. A larger value indicates better performance.

Settings: The bit size of binary hash code will directly affect the performance of the proposed NRHC. Generally, the size is larger than 100 for the proposed metric. On the other hand, the size of hash code is limited to the dimension of the image features. When all of these conditions are considered together, the size of hash code is set to 120 in this paper. And this setting can lead to a good performance.

3.1 Consistency

The subjective score (DMOS or MOS) is the most reliable evaluation because human beings are the ultimate recipients of the image. Therefore, the consistency

Table 1. Comparison of the Performance (LCC) on the LIVE II Database.

Metric	Type	JP2K	JPEG	WN	Gblur	FF	All
PSNR	FR	0.8962	0.8596	0.9858	0.7834	0.8895	0.8240
SSIM	FR	0.9367	0.9283	0.9695	0.8740	0.9428	0.8634
BIQI	NR	0.8086	0.9011	0.9538	0.8293	0.7328	0.8205
LBIQ	NR	–	–	–	–	–	–
DIIVINE	NR	0.9220	0.9210	0.9880	0.9230	0.8880	0.9170
BLIINDS	NR	0.8070	0.5970	0.9140	0.8700	0.7430	0.6800
LQF	NR	0.8424	0.8310	0.8523	0.8459	0.7976	0.8021
QAC	NR	0.8648	0.9435	0.9180	0.9105	0.8248	0.8625
NRHC	**NR**	**0.9006**	**0.8674**	**0.8845**	**0.8887**	**0.8955**	**0.8714**

Table 2. Comparison of the Performance (SRCC) on the LIVE II Database.

Metric	Type	JP2K	JPEG	WN	Gblur	FF	All
PSNR	FR	0.8898	0.8409	0.9853	0.7816	0.8903	0.8197
SSIM	FR	0.9317	0.9028	0.9629	0.8942	0.9411	0.8510
BIQI	NR	0.7995	0.8914	0.9510	0.8463	0.7067	0.8195
LBIQ	NR	0.9000	0.9200	0.9700	0.8800	0.7800	0.8900
DIIVINE	NR	0.9130	0.9100	0.9840	0.9210	0.8630	0.9160
BLIINDS	NR	0.8050	0.5520	0.8900	0.8340	0.6780	0.6630
LQF	NR	0.8389	0.8323	0.8472	0.8456	0.8018	0.8156
NIQE	NR	0.9187	0.9422	0.9718	0.9329	0.8639	0.9086
QAC	NR	0.8621	0.9362	0.9509	0.9134	0.8231	0.8683
NRHC	**NR**	**0.8622**	**0.8428**	**0.8518**	**0.8806**	**0.8651**	**0.8776**

between the objective evaluations and the subjective scores is the most important performance. To verify that the algorithms are robust to the image content, a cross-validation experiment is conducted on the database. Part of original images and their corresponding distorted images (80%) are randomly selected for model training, with for the remainder being used as test images. The performance of LCC and SROCC is the average of the experimental results with 100 times of random cross-validation. The results are shown in Table 1 and 2. Where, PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural SIMilarity) [21] are typical full-reference image quality assessment metrics, and BIQI (the Blind Image Quality Indices) [5], LBIQ (Learning based Blind Image Quality measure) [7], DIIVINE (the Distortion Identification-based Image Verity and INtegrity Evaluation) [6], BLIINDS (BLind Image Integrity Notator using DCT Statistics) [8] [10], LQF (Latent Quality Factors) [15] and QAC (Quality-Aware Clustering method) [17] are no-reference image quality assessment metrics.

Table 1 and 2 show the comparison of performances on the sub (JP2K, JPEG, WN, Gblur, FF) and the entire LIVE II database. It can be found that different NR-IQAs present the best performance on some distorted sub database. The proposed NRHC obtains better performance than the state-of-the-art methods on most conditions. Furthermore, the proposed NRHC has similar performance on different distortions. Hence, the proposed NRHC has greater robustness than other metrics.

3.2 Database Independence

Most of the NR-IQA metrics need to determine model parameters through learning algorithm on the training set. Hence, it is necessary to verify the robustness and generalization. It means that whether the learned model is sensitive to different databases or database independence.

The metrics are trained on the LIVE II database and tested on other databases including TID (1700 distorted images with 17 different distortions), CSIQ (900 distorted images with 6 different distortions), IVC (195 distorted images with 4 different distortions), and MICT (168 distorted images with 2 different distortions). The experimental results of PSNR, SSIM, BIQI, DIIVINE, BLIINDS, LQF, QAC and the proposed NRHC on these publicly available databases are shown in Table 3. It can be found that the proposed metric has better stability than other metrics.

Table 3. LCC on other Public Databases.

Metric	Type	TID	CSIQ	IVC	MICT
PSNR	FR	0.5643	0.8772	0.7192	0.6355
SSIM	FR	0.6387	0.8060	0.7924	0.7979
BIQI	NR	0.4192	0.6601	0.5346	0.6853
DIIVINE	NR	0.7749	0.8284	0.3300	0.6416
BLIINDS	NR	0.5086	0.7529	0.7013	0.7924
LQF	NR	0.4231	0.6396	0.6191	0.7042
QAC	NR	0.8538	0.8416	0.7676	0.5189
NRHC	**NR**	**0.5387**	**0.6826**	**0.6237**	**0.7187**

Table 4. Computational time on LIVE II Databases.

Metric	Type	Training	Testing
PSNR	FR	–	1.86s/100p
SSIM	FR	–	7.20s/100p
BIQI	NR	–	74.25s/100p
BLIINDS	NR	–	85.27s/100p
NRHC	**NR**	**1.75s/100p**	**0.21s/100p**

3.3 Computational Time

Time cost will greatly affect the effectiveness and efficiency and plays a significant role in the real-time application and big image data processing system. In this subsection, we test and compare the computing time of some existing methods with the proposed NRHC. The computational time experiment is conducted on the LIVE II database and the same runtime environment. The results are shown in Table 4. The computational time is recorded and presented under processing for every 100 images. From the table, we can find that the proposed NRHC has the best performance on computational time cost.

4 Conclusions

This paper proposed a novel no-reference image quality assessment method via hash codes. The proposed NRHC is effective and efficient which are demonstrated by the analysis and experiments. The proposed NRHC first extract the spatial natural scene statistics features, embed the features into hash codes via an improved iterative quantization method, and calculate the Hamming distance between the hash code of the test image and the original undistorted image to predict image quality. The hash coding and the Hamming distance calculation have the properties of fast speed and high efficiency. Hence, the proposed NRHC can satisfy the real-time applications and big image data processing. However, the proposed NRHC has not considered the characteristics of the human visual system (HVS), such as visual saliency. Therefore, the method combining the fast algorithm and the properties of HVS need to be studied. Additionally, in order to assess stereo images, video and high definition, the proposed method also needs to be extended to fit new applications.

Acknowledgments. This research was supported partially by the National Natural Science Foundation of China (No. 61125204, 61372130, 61432014, 61501349), the Fundamental Research Funds for the Central Universities (No. BDY081426, JB140214, XJS14042), the Program for New Scientific and Technological Star of Shaanxi Province (No. 2014KJXX-47), the Project Funded by China Postdoctoral Science Foundation (No. 2014M562378).

References

1. Wang, Z., Bovik, A.C.: Modern Image Quality Assessment. Morgan and Claypool, New York (2006)
2. Sheikh, H.R., Bovik, A.C., Cormack, L.: No-reference quality assessment using natural scene statistics: JPEG2000. IEEE Trans. Image Processing **14**(11), 1918–1927 (2005)
3. Wang, Z., Bovik, A.C., Evans, B.L.: Blind measurement of blocking artifacts in images. In: Proc. IEEE Int. Conf. Image Processing, vol. 3, pp. 981–984 (2000)
4. Li, L., Wang, Z.-S.: Compression quality prediction model for JPEG2000. IEEE Trans. Image Processing **19**(2), 384–398 (2010)
5. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. IEEE Signal Processing Letters **17**(6), 513–516 (2010)
6. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: from natural scene statistics to perceptual quality. IEEE Trans. Image Processing **20**(12), 3350–3364 (2011)
7. Tang, H., Joshi, N., Kapoor, A.: Learning a blind measure of perceptual image quality. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 305–312 (2011)
8. Saad, M.A., Bovik, A.C., Charrier, C.: A DCT statistics based blind image quality index. IEEE Signal Processing Letters **17**(6), 583–586 (2011)

9. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: Proc. Asilomar Conf. Signals, Systems and Computers, pp. 723–727 (2011)
10. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: a natural scene statistics approach in the DCT domain. IEEE Trans. Image Processing **21**(8), 3339–3352 (2012)
11. Ye, P., Doermann, D.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Processing **21**(7), 4695–4708 (2012)
12. He, L., Tao, D., Li, X., Gao, X.: Sparse representation for blind image quality assessment. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1146–1153 (2012)
13. Mittal, A., Muralidhar, G.S., Ghosh, J., Bovik, A.C.: Blind image quality assessment without human training using latent quality factors. IEEE Signal Processing Letters **19**(2), 75–78 (2012)
14. Ye, P., Doermann, D.: No-reference image quality assessment using visual codebooks. IEEE Trans. Image Processing **21**(7), 3129–3138 (2012)
15. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013)
16. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural network for no-reference image quality assessment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2014)
17. Xue, W., Mou, X., Zhang, L., Bovik, A.C., Feng, X.: Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. IEEE Trans. Image Processing **23**(11), 4850–4862 (2014)
18. Gu, K., Zhai, G., Yang, X., Zhang, W.: Using free energy principle for blind image quality assessment. IEEE Trans. Multimedia **17**(1), 50–63 (2015)
19. Buccigrossi, R.W., Simoncelli, E.P.: Image compression via joint statistical characterization in the wavelet domain. IEEE Trans. Image Processing **8**(12), 1688–1701 (1999)
20. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans. Pattern Analysis and Machine Intelligence **35**(12), 2916–2929 (2013)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing **13**(4), 600–612 (2004)
22. VQEG, Validation of reduced-reference and no-reference objective models for standard definition television, Phase I (2009). http://www.vqeg.org/
23. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: Live image quality assessment database release 2 (LIVE II) (2003). http://live.ece.utexas.edu/research/quality
24. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: TID2008 - a database for evaluation of full-reference visual quality assessment metrics. Advances of Modern Radioelectronics **10**, 30–45 (2009)
25. Larson, E.C., Chandler, D.M.: Categorical image quality (CSIQ) database (2009). http://vision.okstate.edu/csiq
26. Callet, P.L., Autrusseau, F.: Subjective quality assessment - IVC database (2006). http://www.irccyn.ec-nantes.fr/ivcdb/
27. Horita, Y., Shibata, K., Kawayoke, Y., Sazzad, Z.M.P.: MICT Image Quality Evaluation Database (2000). http://mict.eng.u-toyama.ac.jp/mictdb.html

Shooting Recognition and Simulation in VR Shooting Theaters

Shuo Feng[1], Wei Gai[1], Yanning Xu[1,2] (✉), Tingting Cui[1], Pu Qin[1], Huiyu Li[1],
Dongdong Guan[1], Chenglei Yang[1,2](✉), and Xiangxu Meng[1,2]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
{xyn,chl_yang}@sdu.edu.cn
[2] Engineering Research Center of Digital Media Technology,
Ministry of Education of PRC, Jinan, China

Abstract. VR shooting theaters are very popular now, and shooting recognition system is one of the important components. In this paper, we will introduce a new simulation gun and a novel shooting recognition system used in VR shooting theaters in which multi-players can moving freely, and describe how to improve the realistic experience of shooting actions in virtual worlds. We will also show one of its application in two-view VR shooting theaters in which players can see the different pictures rendered from different viewpoints.

Keywords: Shooting game · VR theater · Simulation gun

1 Introduction

The traditional first-person shooter video game exists mainly in the consoles, and users play games with keyboard, mouse, or joystick in an unnatural way. Currently, more and more virtual reality (VR) theaters come forth in cultural theme parks, exhibition museums and so on, with the development of VR, human-computer interaction, computer animation, computer game and projector technology. The existing VR theaters [1] usually involve a large screen, 3D glasses, surround-stereo sound systems and dynamic seats. They allow the interaction of a large number of audiences. Generally, all the players share one same scene image. As one kind of the VR theaters, the shooting theater systems provide shooting game with cooperative mode, usually limiting players in the seats, so the players cannot move freely to shoot the hiding targets [2]. To enhance the immersive experience, simulation guns based on infrared and laser technology are made-up of complex and expensive apparatus [3]. The guns are connected to the seats by cables in most scenarios. However, the use of simulation guns becomes the biggest drawback of this form of the game, mainly embodying as follows: First, the simulation guns correction process is very complex, affected deeply by subjective ideas of the correcting people. Second, simulation guns on each seat needs to be corrected, which are large correction tasks. Third, the use of wired simulation guns, limits user interaction and impacts on the user's immersive experience.

We implemented a two-view VR shooting theater system in which multi-players can move freely and interact with the movies, and briefly introduced a shooting

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 286–292, 2015.
DOI: 10.1007/978-3-662-48570-5_28

recognition system [4]. In this paper, we focus on introducing an improved shooting recognition system. The simulation guns only have laser transmitters and have no cameras in them. The system only uses an infrared camera to capture screen images, and we complete interactive shooting task by analyzing the position of front sight and users' shooting information. Without the limitation of power cable, each user can move freely. We also discuss how to improve the realistic experience of shooting actions in virtual worlds in this paper, and show one of its application in two-view VR shooting theaters in which players can see the different pictures rendered from different viewpoints.

2 System Architecture of the Shooting Recognition System

Figure 1 shows the system architecture of the shooting recognition system. The camera is placed where lens is on the centerline of the curtain, and install it on the roof or other high at 2.5 to 3 meters away from the curtain. Adjust the position by monitoring software that comes with the camera. Then the projector is also placed where lens is on the centerline of the curtain. We install it on the roof or other high (It can be placed at the bottom of the camera). Adjust the position of the projector imaging. Make sure the camera and projector are set up a good network connection to the computer. We complete the front sight positioning and determine the case of the simulation guns by infrared emitters on the guns. The simulation gun has two laser emitters. One is on all the time and utilized to locate front sight, another is on while player is shooting, and is utilized to determine whether player is shooting.

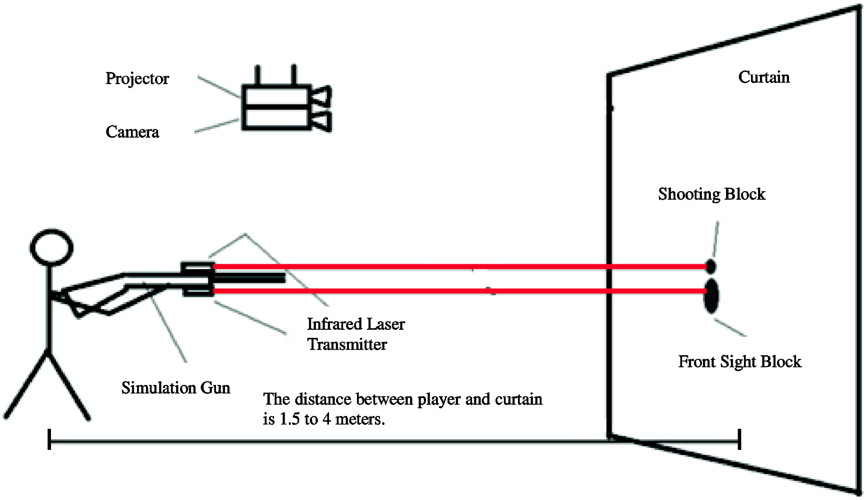

Projector
Camera
Curtain
Shooting Block
Infrared Laser Transmitter
Front Sight Block
Simulation Gun
The distance between player and curtain is 1.5 to 4 meters.

Fig. 1. Multiplayer free shooting recognition system and device

3 The Method of Shooting Recognition

We briefly introduced a shooting recognition system in reference [4]. The simulation gun was equipped with three infrared lasers on the head. Two of them on the side are turned on to locate the front sight, and the last one in the middle is controlled by the trigger as the signal of shoot. When the trigger was pulled, the infrared camera will catch the light of middle laser, and the shooting event will be triggered. The infrared lasers can compose different patterns with the parameters such as different angles, distance etc., and it's convenient to be recognized using pattern recognition technology. As the example in Figure 2, gun A shows a line shape and gun B shows a triangle shape, so the infrared camera can distinguish them through capturing and detecting the pattern of laser points. But it is difficult to recognize the guns when several patterns are overlap.

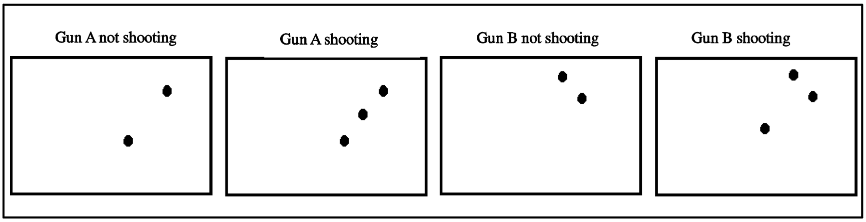


Fig. 2. Different patterns of simulation guns

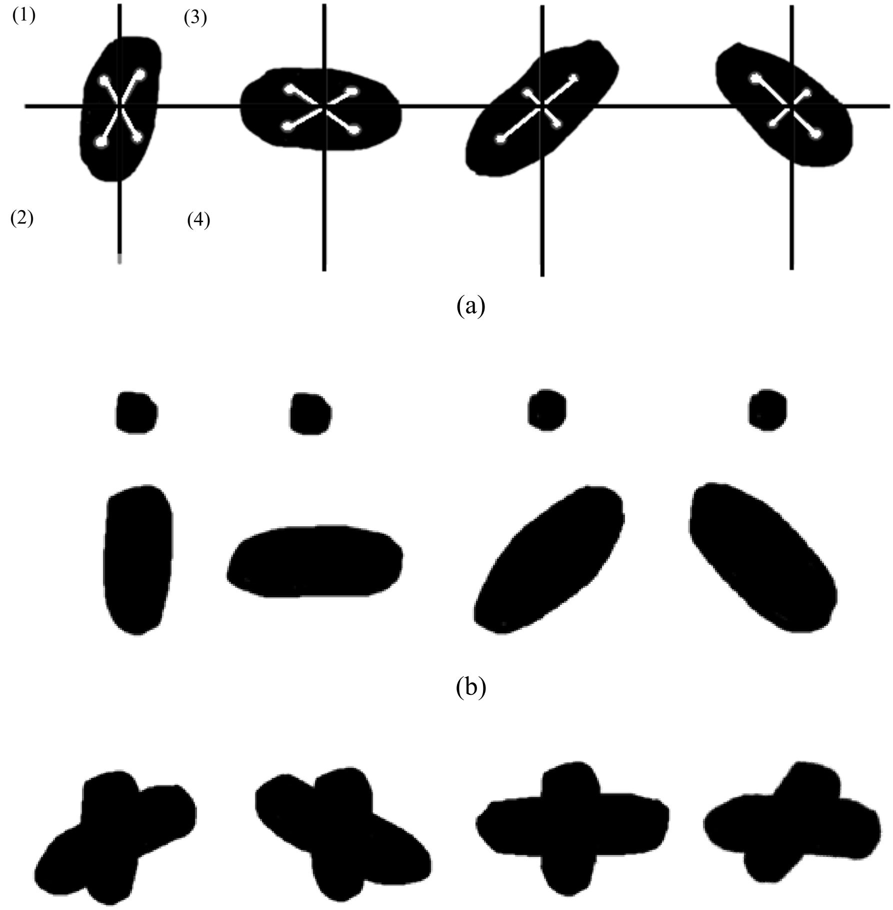
In this paper, a new simulation gun recognition method is presented, which enables users to play a shooting game with wireless simulation guns in a more natural way. The simulation gun is equipped with two infrared lasers. One for target locating is always on during the game (Figure 3.a), and the other one for trigger status will be turned on only if the user pulled the trigger (Figure 3.b). An infrared camera is adopted to capture the infrared laser spots from the curtain. Because a filter is covered in front of the camera to filter out contents from the projector, we can easily capture laser spots generated by the infrared laser transmitter and record them with black color blocks. Our method is supposed to recognize the front sight position and the triggers status.

Figure 3 shows us different patterns might be produced by our infrared simulation guns. Figure 3(a) describes the four predefined status that users do not pull the trigger; figure 3(b) describes status when users pull the trigger, and we can notice that an extra dot block for firing status is captured; and patterns in figure 3(c) is generated when an overlap is happened during the shooting process.

Before shooting patterns recognition, we define the feature value of shooting patterns as

$$v = (v_1, v_2)$$

and $v_1 = ((x_1 - x_4), (y_1 - y_4)), v_2 = ((x_3 - x_2), (y_3 - y_2))$. We calculate the center point of the front sight block, and build a frame with the center point



(a)

(b)

(c)

Fig. 3. Patterns of infrared simulation guns

as the original point. Then we find a center for each block area in on phase. $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$ are coordinate of center points in phases 1, 2, 3, 4 respectively(Figure 3(a)).

The basic idea is to calculate the feature value of the captured image and compare it with the predefined patterns to decide the status of the simulation gun. The main steps can be described as:

Step 1: Calculate the area of each shooting pattern.

Step 2: Decide the status of each shooting pattern. Set up three predefined ascending value C_1, C_2 and C_3. If the area of block is lower than C_1, we consider it as noise. If between C_1 and C_2, it is a shooting block and goes to step 3. If between C_2 and C_3, it is a front sight block and goes to step 4. If bigger than C_3, it is an overlap block and goes to step 5.

Step 3: Enumerate each front sight block, and calculate distances between the front sight block and the shooting block. Find the nearest pairs that meet the distance requirements and mark the front sight block as firing status.

Step 4: Calculate the feature value, compare it with the shooting patterns, and find the best match pattern with minimum sum of square difference method.

Step 5: Decide the pattern of the overlap front sight blocks. Calculate the missing front sight block of current frame from the previous frame. Find a nearest overlap block for each missing block. If the distance between the missing block and its nearest overlap block is less than a predefined value, the missing block exists and we assign the overlap block coordinate as its new coordinate. Otherwise, the missing block will be skipped.

4 Shooting Effects Simulation

Three types of particle systems are used to simulate the visual effects of firing flare, firing smoke, and hitting smash correspondingly. Transparent billboard pictures are used to represent particles. Particles to simulate the firing flare and the firing smoke are both emitted along the virtual gun direction. The former particle is emitted with high speed and short life cycle. The latter one has long life cycle and low spread speed. Particles for hitting smash simulation are relatively complex to design. The difficulty is that different visual effects are supposed to be displayed when the bullet hits different objects. We have categorized objects into three kinds and designed smoke, splat, and explosion effects to simulate the hitting effects for hard objects, fragile objects, and explosive objects perceptively.

Gunkick and viewkick are introduced to enhance the player experience for firing. A quick backward gunkick movement or a predefined reload animation for virtual gun is displayed and produces the shooting delay. We change the view position slightly after each shot for two reasons. The swift and slight shake of view position can produce the viewkick effects. And players are expected to adjust targets after each shooting action due to the fact that their positions are changed slightly when shooting.

5 Application in Two-View VR Shooting Theater Systems

In conventional projection-based display systems, stereoscopic images are projected onto a single large screen to allow groups of people to view the virtual environment. These systems provide only one stereoscopic image pair in a shared virtual environment, and also all viewers observe stereoscopic images from a single viewing position, this case lacks realistic experience. The multi-view projection display system which presents multiple viewpoints of the same screen concurrently has become the focus in recent years[5][6]. Using this kind of technology, we developed a shooting theater system, equipped with two-view projecting system, 3D shutter glasses, individual surround-stereo earphone and user-customized simulation guns [4]. To provide a more friendly interaction for the players, we use Kinect networks to capture the movement of players.

The system mainly has four modules: output module, interactive input module, real-time parallel rendering module, and integrated processing module. Kinects and simulation guns are utilized as interactive tools to trace information such as shooting position, shooting action, player's dynamic position and ID of simulation gun. Images are rendered by the rendering server based on the information, and then projected on the large screen. Players can see their individual view of 3D scene through the customized shutter glasses and hear sound with stereo sound earphone. At the same time, the integrated processing server also drives the devices to produce special effects of smog, rain, bubbles in line with the story to form a unique experience.

As there could be multiple players in the game, cooperative strategy needs to be made. The system supports two kinds of game mode: collaborative mode in which players will cooperate with each other and adversarial mode in which players in different team will shoot each other. Both modes follow the same rules: when there are two players, they can move freely to drive the virtual avatar and each of them will see his own scene image with different shutter glasses. When there are more than two players, they will be divided into two teams. Players can only see the scene of their own team. Each team should assign a leader who commands the team, control the story plot and change the scene. Each member in one team wears the same kind of shutter glasses.

The multi-view projection display system which presents multiple viewpoints of the same screen concurrently has become the focus in recent years[5][6]. Using this kind of technology, we can also develop a six-view shooting theater system. The method of shooting recognition and simulation in this paper also can be adapted to the six-view shooting theater system.

Different from the traditional single view system, it is necessary to render the scene from different viewpoints simultaneously for multi-view systems. In our system, we maintain different worldview transformation matrix for different players. Although, the positions and shooting information of two players are recognized with the same Kinect network and camera, positions and directions of different group of players are transformed to different places in the scene after the worldview transformation.

6 Conclusion

This paper introduces a kind of multiplayer free shooting recognition system and device for 7D shooting theater. This system can rapidly identify the ID of each player, locate front sight and obtain shooting information. No need to pre-correction, easy to use, more freedom because of wireless. In addition, it is easy to transport and set up, low cost, suitable for family.

For future work, we would like to evaluate the two-view free shooting recognition system by a user study. We will design a shooting task and ask volunteers to fulfill the task in different systems including our two-view system. Objective indicators such as task completion time and subjective indicators such as user satisfactory will be used to evaluate the system.

Acknowledgements. This work was supported by the China National Natural Science Foundation (61272243, 61202146, 61472225), the China National Science and Technology Support Program (2013BAH39F00), Shandong Province Natural Science Foundation (BS2009DX003).

References

1. Ahn, S.C., Kim, I.-J., Kim, H.-G., Kwon, Y.-M., Ko, H.: Audience interaction for virtual reality theater and its implementation. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology-VRST 2001, pp. 41–45 (2001)
2. Virtual Reality 7D Movie Theater With Infrared Control Gun Shooting Games - quality 7D Movie Theater for sale. http://www.5dmovietheater.com/china-virtual_reality_7d_movie_theater_with_infrared_control_gun_shooting_games-2201558.html (accessed on April 23, 2015)
3. Professional PC Light Guns for Arcade Games. http://www.arcadeguns.com/index.php?main_page=index (accessed on April 23, 2015)
4. Yu, H.-D., Zeng, W., Li, H.-Y., Sun, W.-S., Gai, W., Cui, T.-T., Wang, C.-T., Guan, D.-D., Yang, Y.-J., Yang, C.-L.: A two-view VR shooting theater system. In: Proceedings of the 13th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry-VRCAI 2014, pp. 223–226 (2014)
5. Bolas, M., McDowall, I., Corr, D.: New research and explorations into multiuser immersive display systems. IEEE Comput. Graph. Appl. **24**(1), 18–21 (2004)
6. Kulik, A., Kunert, A., Beck, S., Reichel, R., Blach, R., Zink, A., Froehlich, B.: C1x6:a stereoscopic six-user display for co-located collaboration in shared virtual environments. In: Proceedings of the 2011 SIGGRAPH Asia Conference on-SA 2011, vol. 30, no. 6, p. 1 (2011)

An Improved Image Quality Assessment in Gradient Domain

Yuling Ren, Wen Lu(✉), Lihuo He, and Tianjiao Xu

School of Electronic Engineering, Xidian University, Xi'an 710071, China
renyuling@stu.xidian.edu.cn, {luwen,lhhe}@mail.xidian.edu.cn

Abstract. The available image quality assessment (IQA) methods based on gradient calculation are mostly implemented without considering visual perception threshold (VPT) and color information. However, incorporating VPT with IQA model can reduce redundant information and human visual system (HVS) is extremely sensitive to color variation. An improved image quality assessment in gradient domain is proposed which utilizes minimum amount of gradient coefficients to capture the color and structure distortion of degraded image by applying a VPT to remove the unperceived gradient coefficients. The difference of perceived gradient coefficients between distorted and reference image is measured to acquire image quality score. Experimental results on two benchmarking databases (LIVEII and TID2008) indicate the rationality and validity of the proposed method.

Keywords: Image quality assessment · Human visual system · Gradient calculation · Visual perception threshold

1 Introduction

Objective image quality assessment is designed with the aim of interpreting the quality of distorted image automatically and responding consistently with the behavior of the HVS [1-2]. A huge number of IQA algorithms have been emerged with the evolution of image processing technology, which can be divided into two categories, namely HVS based paradigm and non-HVS based metrics. The traditional peak signal to noise ratio (PSNR) [3] just measure the pixel difference between degraded and reference image to obtain the image quality score, which doesn't accord with the way of human perceive information. The perfect IQA model is required to simulate the actual process of HVS perceive image. However, the HVS is extremely complex and the research on it is limited, which lead to the mainstream IQA methods are designed based on certain properties of HVS. The Multi-Scale structural similarity (MS-SSIM) [4] assumes that HVS is sensitive to structure information in an image when perceiving the image quality. Motivated by SSIM, the gradient SSIM (G-SSIM) [5] is built by Chen et al, which first compute the gradient of distorted image and reference image and then measure the luminance similarity, contrast similarity and structural similarity of gradient maps. Given the gradient magnitude maps, the gradient orientation

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 293–301, 2015.
DOI: 10.1007/978-3-662-48570-5_29

maps and contrasts of reference and distorted image, the similarity among them is computed in geometric structure distortion (GSD) [6] method to acquire the image quality score. RR-VIF [7] constructs the IQA model by measuring the change of visual information fidelity in the distorted image. GMSD [8] explores the use of global variation of gradient based local quality map for overall image quality prediction.

The available IQA methods based on gradient calculation are mostly implemented ignoring VPT and color information. However, incorporating VPT with IQA model can reduce redundant information and HVS is extremely sensitive to color variation [9-10]. The human eyes cannot perceive image gradient with its magnitude under VPT. However, there is no consideration of this aspect for these models [4-8]. An improved image quality assessment in gradient domain is proposed in this paper. In the proposed framework, we first calculate the gradient of an image in RGB color space and grayscale domain to capture its color and structure distortion. And then the VPT is determined according to the properties of HVS, which is used to calculate the perceived visual feature. Finally, the difference of perceived visual feature between distorted and reference image is measured to acquire image quality score.

The remainder of this paper is organized as follows. Section 2 presents the comprehensive implementation of proposed algorithm. Section 3 illustrates the experimental result and a though analysis. Finally, conclusion is made in section 4.

2 Image Quality Assessment in Gradient Domain

Fig. 1. presents the structure of the proposed metric. In the first step, we calculate the gradient of distorted and reference image in RGB color space and grayscale domain. The greater of gradient magnitude imply the huger variation in the image and yet the tiny change in the image can’t be perceived by human eyes. So the next step is to compute the VPT of reference image gradient magnitude. Afterwards, the proportion of perceived gradient magnitude of reference and distorted image is calculated according to the VPT. Finally, the objective image assessment is acquired by comparing the difference of the proportion of perceived gradient magnitude between reference and distorted image.

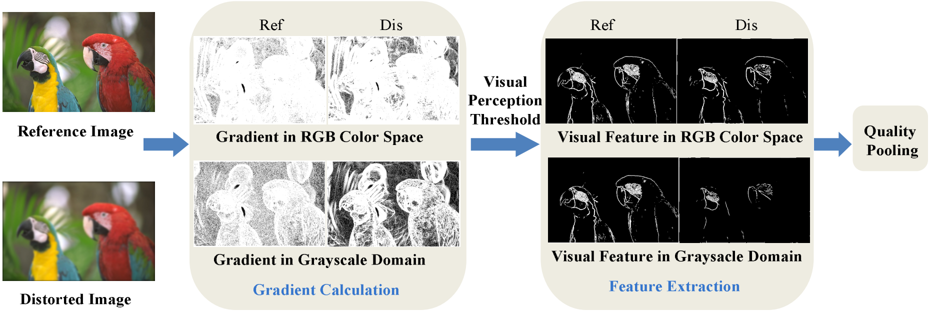

Reference Image
Distorted Image
Ref
Dis
Gradient in RGB Color Space
Gradient in Grayscale Domain
Gradient Calculation
Visual Perception Threshold
Ref
Dis
Visual Feature in RGB Color Space
Visual Feature in Graysacle Domain
Feature Extraction
Quality Pooling

Fig. 1. The proposed image quality assessment algorithm framework

2.1 Gradient Calculation

It is well known that image gradient is sensitive to distortions, however, the mostly existing IQA methods just compute image gradient in grayscale domain, which ignores the fact that image gradient in RGB color space has a great influence on quality prediction. With g_x denotes the horizontal direction of the filter and g_y denotes the vertical direction, the calculation of gradient magnitude complies with the the following rules. Let I denote an image.

$$G_{gray} = \sqrt{(I \otimes g_x)^2 + (I \otimes g_y)^2} \text{ ,} \tag{1}$$

$$G_{rgb} = \sqrt{\begin{array}{l}(I_r \otimes g_x)^2 + (I_r \otimes g_y)^2 + (I_g \otimes g_x)^2 \\ + (I_g \otimes g_y)^2 + (I_b \otimes g_x)^2 + (I_b \otimes g_y)^2\end{array}} \text{ ,} \tag{2}$$

Where " $\otimes$ " is the linear convolution operator and G_{gray} denotes the gradient in the grayscale domain, G_{rgb} denotes the gradient in RGB color space. I_r, I_g and I_b denote the R, G and B channel of image respectively. The gradient in the grayscale domain is compute at four scales to capture multiscale behavior, by low pass filtering. g_v is the Gaussian partial derivative filter applied along the horizontal (x) or vertical (y) direction:

$$g_v((x,y) \mid \alpha) = \frac{\partial}{\partial v} g((x,y) \mid \alpha) \tag{3}$$

$$= -\frac{1}{2\pi\alpha^2} \frac{v}{\alpha^2} \exp(-\frac{x^2+y^2}{2\alpha^2}), v \in \{x,y\} \text{ ,}$$

Where α is the scale parameter. Fig. 2 shows the gradient map in RGB color space and grayscale domain of natural image and corresponding distorted image. What we can see from Fig. 2 is that the degradation of image will induce obvious change of image gradient in RGB color space and grayscale domain.

2.2 Visual Perception Threshold

The available IQA models based on gradient just compute the similarity of image gradient structure without considering the human visual perception threshold. However, the tiny change in the image can't be perceived by human and therefore the VPT is required to remove the diminutive gradient magnitude which doesn't arouse respond in HVS [11-12]. The VPT is defined by.

$$T = \omega \sqrt{\frac{C}{C-1} \sum_{k=1}^{C} (g(k) - \bar{g})^2} \text{ ,} \tag{4}$$

Where C is the amount of gradient coefficients, $g(k)$ is the kth gradient coefficients and $\overline{g}$ is the mean of all gradient coefficients, ω a tuning parameter. Based on the eq. (4), we can obtain the visual perception threshold in RGB color space and grayscale domain denoted by T_{rgb} and T_{gray}.

It is valuable to preserve visually sensitive gradient coefficients by VPT and the amount of visual sensitive gradient coefficients reflects the visual quality of the images, which reduce the amount of feature and decrease the complexity of algorithm.

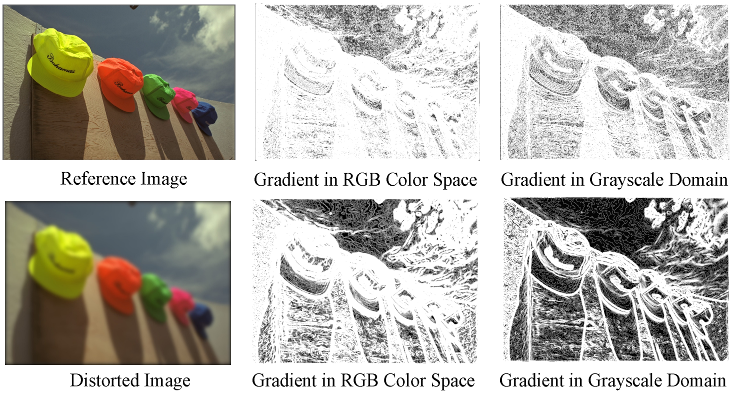

Reference Image
Gradient in RGB Color Space
Gradient in Grayscale Domain
Distorted Image
Gradient in RGB Color Space
Gradient in Grayscale Domain

Fig. 2. Gradient map in RGB color space and grayscale domain

2.3 Visual Perception Feature Extraction

By using VPT obtained by eq. (4), we can count the number of visually sensitive gradient coefficients in RGB color space and grayscale domain. Therefore, for a given image, we can obtain the proportion of perceived gradient coefficients N based on eq. (6).

$$C_T = \{C > T\}, \tag{5}$$

$$N = \frac{C_T}{C}, \tag{6}$$

Where C is total number of gradient coefficients and C_T is the visual perceived gradient coefficients which are greater than VPT. With eq. (5), (6) the proportion of visual perceived gradient coefficients in RGB color space and grayscale domain are got and denoted by N_{rgb}, N_{gray} , which are defined as the visual perception feature.

2.4 Quality Pooling

In the proposed framework final quality index is defined by weighted strategy of Q$_{rgb}$ and Q$_{gray}$.

$$Q_{rgb} = \frac{1}{1+\log_2(1+\frac{D_{rgb}}{\lambda})}, \tag{7}$$

$$Q_{gray} = \frac{1}{1+\log_2(1+\frac{D_{gray}}{\lambda})}, \tag{8}$$

$$Q = \alpha Q_{rgb} + \beta Q_{gray}, \tag{9}$$

Where λ , α , β are the tuning parameters, D_{rgb} and D_{gray} are the difference of visual feature in the RGB color space and grayscale domain, which are obtained by the following equations.

$$D_{rgb} = |N_{rgb\_r} - N_{rgb\_d}|, \tag{10}$$

$$D_{gray} = \sum_{i=1}^{S} |N_{gray\_r}(i) - N_{gray\_d}(i)|, \tag{11}$$

Where N_{rgb_r}, $N_{rgb\_r}(i)$ and N_{rgb_d} , $N_{rgb\_d}(i)$ are the visual perception feature of reference and distorted image in RGB color space and grayscale domain. S is the number of image scale in grayscale domain obtained by low pass filtering and i is the scale index.

3 Experimental Results

Experiments are done on the LIVE database II [13] and the TID2008 database [14] to verify the rationality and validity of proposed. LIVE database II contains 29 high-resolution 24 bits/pixel RGB color images and 175 corresponding JPEG and 169 JPEG2K compressed images, as well as 145 white noisy (WN), 145 Gaussian blur (GB), and 145 fast-fading (FF) Rayleigh channel noisy images at a range of quality levels. We select five types of distortion in the TID2008 database to complete the experiment, i.e., Gaussian blur (GB), Image denoising (DEN), JPEG compression (JPEG), JPEG2K compression (JPEG2K) and JPEG transmission errors (JGTE). The assessment indexes considered in the experiment is spearmans rank ordered correlation coefficient (SROCC). The value of SROCC closer to 1 implies superior consistency with human perception.

3.1 Consistency Experiment

In this section, we compare the performance of the proposed framework with standard IQA methods, i.e., PSNR [3], MS-SSIM [4], G-SSIM [5], GSD [6] and RR-VIF [7]. The values for SROCC of all the IQA metrics mentioned above are given in tables 1, 2. Fig. 3 presents the nonlinear fitting of the objective quality score obtained by proposed versus mean opinion score (MOS) on LIVE database II. In implement the IQA task, methods [3-6] require full information of reference image while the proposed just utilize a fraction of information, which reduce the redundant information and complexity of algorithm. RR-VIF [7] build IQA model based on the additive noise model, while the mostly of distortions on LIVE II database are induced by additive noise and hence the performance of RR-VIF [7] on LIVE II database is superior to the proposed. However, the distortions on TID2008 are generated by additive noise and multiplicative noise and therefore the performance of RR-VIF [7] declined. Mostly of the SROCC values on TID2008 for the proposed are higher than that for algorithms [3-7]. In general, consistency experiment shows that the proposed owns a preferable result.

Table 1. SROCC of different metrics on LIVE II database

Metric	JPEG2K	JPEG	WN	GB	FF
PSNR	0.895	0.881	0.985	0.782	0.891
MS-SSIM	0.963	0.981	0.973	0.954	0.947
G-SSIM	0.935	0.944	0.926	0.968	0.948
GSD	0.911	0.931	0.879	0.964	0.953
RR-VIF	0.950	0.885	0.946	0.961	0.941
Proposed	0.927	0.827	0.919	0.956	0.937

Table 2. SROCC of different metrics on TID2008 database

Metric	GB	DEN	JPEG	JPEG2K	JGTE
PSNR	0.870	0.942	0.872	0.813	0.752
MS-SSIM	0.691	0.859	0.956	0.958	0.932
G-SSIM	0.924	0.880	0.859	0.944	0.855
GSD	0.911	0.878	0.839	0.923	0.880
RR-VIF	0.942	0.948	0.599	0.928	0.891
Proposed	0.937	0.946	0.796	0.951	0.787

3.2 Rationality Experiment

To verify the rationality of the proposed metric, we choose four sets of images with different distortions, which are Gaussian blur, spares sampling and reconstruction, chromatic aberrations and Image denoising. Fig. 4 illustrates the prediction trend of the four sets images with different quality. It can be observed that the proposed method prediction trend rises with the increasing of MOS on different types of distortions. It proves the rationality of the proposed framework.

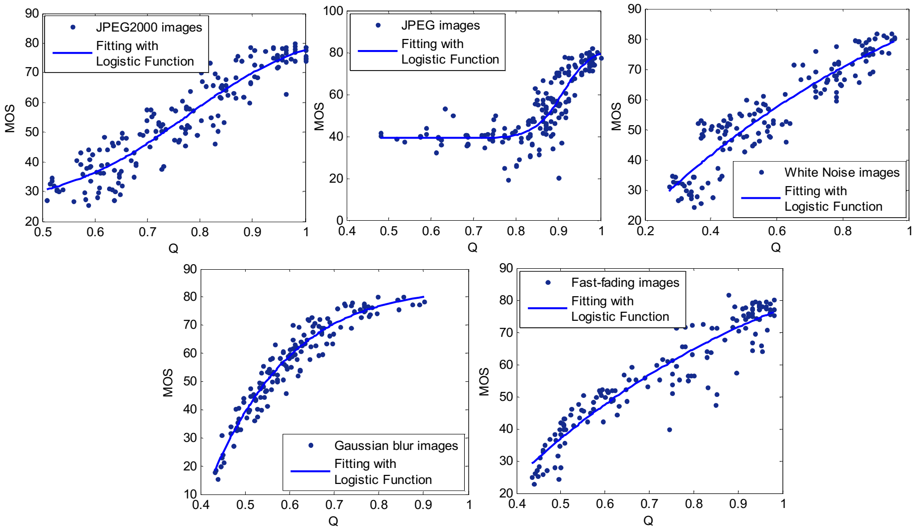


Fig. 3. Nonlinear scatter plots of MOS versus the proposed metric.

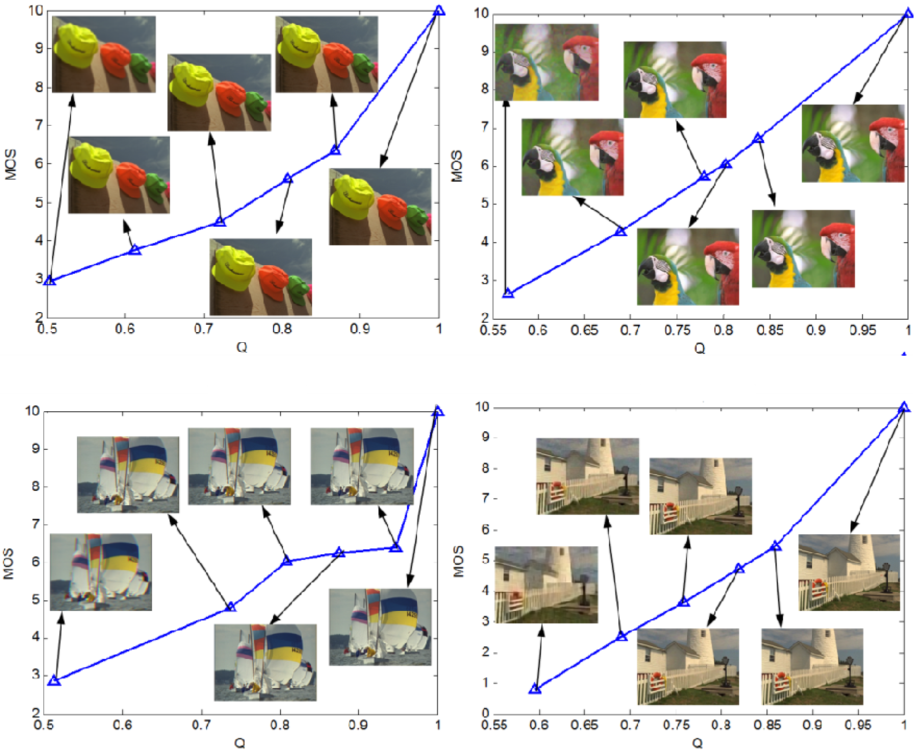


Fig. 4. Results of rationality experiment

3.3 The Performance of Gradient in RGB Color Space

The available IQA metrics based on image gradient just calculate the gradient in grayscale domain, which fails to consider the case that HVS is sensitive to color change. Therefore, we compute the gradient both in the grayscale domain and RGB color space to capture the structure and color feature. The first strategy map gradient in grayscale domain to the quality score and denote as G_{gray}. Both the gradient in grayscale domain and RGB color space is used to obtain the quality score denote as $G_{gray+rgb}$, which is defined as the second strategy. Tables 3, 4 show the performance of the two different strategies. The values for SROCC of $G_{gray+rgb}$ are higher than that of G_{gray}, which verifies the rationality and validity of gradient in RGB color space in the proposed IQA model.

Table 3. SROCC of different quality pooling on LIVE database II

	JPEG2K	JPEG	WN	Gblur	FF
G_{gray}	0.910	0.812	0.850	0.946	0.924
$G_{gray+rgb}$	0.927	0.827	0.919	0.956	0.937

Table 4. SROCC of different quality pooling on TID2008 database

	GB	DEN	JPEG	JP2K	JGTE
G_{gray}	0.871	0.915	0.745	0.933	0.760
$G_{gray+rgb}$	0.937	0.946	0.796	0.951	0.787

4 Conclusion

A novel image quality assessment metric in gradient domain is proposed. In the proposed framework, the gradient is first calculated in RGB color space and grayscale domain to obtain the change in the color and structure of a distorted image. VPT which determined from the reference image is utilized to produce a noticeable variation in sensory experience. Finally, the objective image quality assessment is acquired by measuring the difference of the proportion of visual sensitive gradient coefficients between reference and distorted image. Although the proposed achieves a desirable performance, it is essential to develop blind image quality metrics that estimate the quality of images without any prior information of nature image.

Acknowledgements. This research was supported partially by the National Natural Science Foundation of China (No. 61125204, 61372130, 61432014), the Fundamental Research Funds for the Central Universities (No. BDY081426, JB140214, XJS14042), the Program for New Scientific and Technological Star of Shaanxi Province (No. 2014KJXX-47), the Project Funded by China Postdoctoral Science Foundation (No. 2014M562378).

References

1. Wang, Z., Bovik, A.C.: Modern image quality assessment. Synthesis Lectures on Image, Video, and Multimedia Processing **2**(1), 1–156 (2006)
2. Park, H., Har, D.H.: Subjective image quality assessment based on objective image quality measurement factors. IEEE Transactions on Consumer Electronics **57**(3), 1176–1184 (2011)
3. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 20th International Conference on Pattern Recognition, pp. 2366–2369. IEEE, Istanbul (2010)
4. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers 2004, pp. 1398–1402 (2003)
5. Chen, G.H., Yang, C.L., Xie, S.L.: Gradient-based structural similarity for image quality assessment. In: 2006 IEEE International Conference on Image Processing, pp. 2929–2932. IEEE (2006)
6. Cheng, G., Huang, J.C., Zhu, C., et al.: Perceptual image quality assessment using a geometric structural distortion model. In: International Conference on Image Processing, pp. 325–328. IEEE (2010)
7. Wu, J., Lin, W., Shi, G., Liu, A.: Reduced-reference image quality assessment with visual information fidelity. IEEE Transactions on Multimedia **15**(7), 1700–1705 (2013)
8. Xue, W., Zhang, L., Mou, X., et al.: Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Transactions on Image Processing **23**(2), 684–695 (2014)
9. McCollough, C.: Color adaptation of edge-detectors in the human visual system. Science **149**(3688), 1115–1116 (1965)
10. Gerhard, H.E., Wichmann, F.A., Bethge, M.: How sensitive is the human visual system to the local statistics of natural images. PLoS Computational Biology **9**(1), e1002873 (2013)
11. Blakemore, C., Muncey, J.P.J., Ridley, R.M.: Stimulus specificity in the human visual system. Vision Research **13**(10), 1915–1931 (1973)
12. Chitprasert, B., Rao, K.R.: Human visual weighted progressive image transmission. IEEE Transactions on Communications **38**(7), 1040–1044 (1990)
13. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: Live image quality assessment database, release 2. http://live.ece.utexas.edu/research/quality
14. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: Tid 2008 - a database for evaluation of full-reference visual quality assessment metrics. Advances of Modern Radioelectron **10**(4), 30–45 (2009)

A New Dataset and Evaluation for Infrared Action Recognition

Chenqiang Gao[1](✉), Yinhe Du[1], Jiang Liu[1], Luyu Yang[1], and Deyu Meng[2]

[1] Chongqing Key Laboratory of Signal and Information Processing,
Chongqing University of Posts and Telecommunications, Chongqing, China
gaocq@cqupt.edu.cn

[2] Institute for Information and System Sciences
and Ministry of Education Key Lab of Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an, China

Abstract. Action recognition (AR) is one of the most important tasks in computer vision and there are a large number of related research works along this line. While most of these works are investigated on AR datasets collected from the visible spectrum, the AR problem on infrared scenarios still has not attracted much attention, and there is even few public infrared datasets available for supporting this research. This study aims to emphasize the importance of the infrared AR problem in real applications and arouse researchers' attention on this task. Specifically, we construct a new infrared action dataset and evaluate the state-of-the-art AR pipeline, including widely-used low-level local descriptors, coding methods and fusion strategies, on it. Through these evaluations, we find some interesting results. E.g., dense trajectory feature can achieve the best performance while the appearance features, e.g., HOG, has relatively poorer performance; the coding method of vector of locally aggregated descriptors is evidently better than that of the widely-used fisher vector; the late fusion facilitates a better performance than early fusion. Furthermore, the best performance achieved on our dataset is 70%, leaving a relative large space for promoting new methods on this infrared AR task.

Keywords: Infrared action dataset · Action recognition · Local descriptors · Feature fusion

1 Introduction

Action recognition (AR) is one of the most important tasks in computer vision. Its potential applications include video surveillance, video indexing, human-computer interaction (HCI), etc. [1]. Over the past decades, human action recognition has attracted extensive attention and a number of methods have been proposed to address this task [24]. Basically, most of the efforts have been put into visible imaging videos and many existing methods follow the pipeline: raw feature extraction, feature coding and classifier learning. Generally speaking, the description ability of the adopted features is very important to the performance

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 302–312, 2015.
DOI: 10.1007/978-3-662-48570-5_30

of the method. So far, many good feature descriptors have been widely used for action recognition, such as STIP [18], HOG3D [14], 3DSIFT [23], etc.

The development of feature descriptors needs to be refined and substantiated on proper AR datasets. Recently, many AR datasets have been constructed to research purposes, such as KTH [22], UCF sports [26], HMDB51 [16],WEB-interaction [8], etc. The recently proposed AR datasets [15] more and more simulate real scenarios. While benefited from these datasets, recently designed methods for AR can better adapt real applications, these methods still often encounter great challenges, such as illumination change, shadow, background clutter, occlusion of the object, etc. Actually, these challenges also make other computer vision tasks, like object detection, very hard to be effective only based on the provided visible information.

Compared to visible spectrum imaging, the infrared thermal imaging have many complementary advantages over the aforementioned challenges [10], e.g., the infrared imaging is able to work well under poor light condition, like imaging at night. These advantages have been utilized in pedestrian detection [30], face recognition [13] and other computer vision tasks, but still have not been attracted much attention in AR community [9]. Especially,to the best of our knowledge, there is still no public dataset available for AR research purpose so far.

To the aforementioned issues, we set up a new infrared dataset, called infrared action dataset (IAD), for the infrared AR task. Following the approach of construction of existing AR datasets in visible spectrum [3], the new dataset collects 12 kinds of common human actions. The samples vary from simple to complex scenes. We further evaluate the state-of-the-art AR pipeline on our dataset. Specifically, our evaluation emphasis is put on widely-used low-level local descriptors, the coding strategies and the fusion strategies. This work is expected to establish a benchmark and baseline for infrared AR research, like KTH dataset for AR of visible spectrum.

The rest of this paper is organized as follows: Section 2 introduces details of the newly constructed dataset. Section 3 introduces the employed local descriptors and the utilized evaluation methods. Section 4 presents experimental setup and evaluation results on the dataset. The conclusion is drawn in Section 5.

2 Infrared Action Dataset(IAD)

Following the approach to construct a AR dataset from the visible spectrum [3], we collect 12 common human actions from infrared videos. As shown in Fig. 1, the action types include one hand wave(wave1), multiple hands wave(wave2), handclapping, walk, jog, jump, skip, handshake, hug, push, punch and fight. Each action type has 30 video clips. All actions are performed by 25 different volunteers. The videos are captured by a handled infrared camera IR300A. Each clip lasts 4 seconds in average. The frame rate is 25 frames per second and the resolution is 293×256. Each video contains one person or several persons performing one action or more actions. Some of them are interactions between multiple persons. Table 1 lists the detailed information of our IAD and some known existing visible AR datasets.

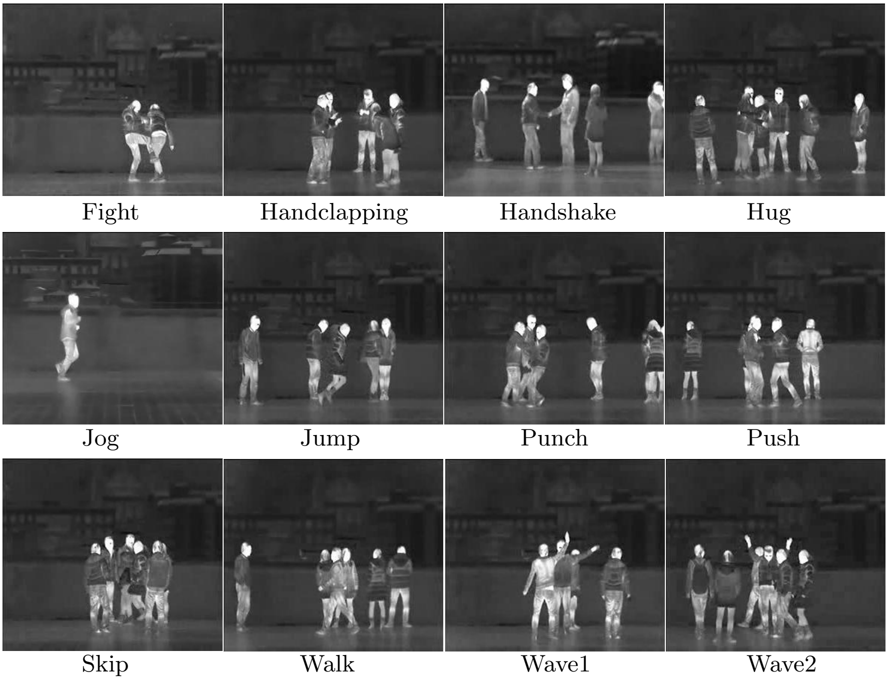

Fight
Handclapping
Handshake
Hug
Jog
Jump
Punch
Push
Skip
Walk
Wave1
Wave2

Fig. 1. 12 actions of the newly constructed infrared AR dataset.

Table 1. Comparison of existing AR datasets and the new IAD dataset.

	KTH	IXMAS	UCF sports	HMDB51	IAD
Video clips	600	2236	156	(min)5151	360
Action Class	6	13	13	51	12
Resolution	160×120	390×291	480×360	320×240	293×256
Frame Rate	25	25	25	30	25
Average Length(s)	4	3	3	3	4
Data Type	visible	visible	visible	visible	infrared
Reference	[29]	[28]	[21]	[29]	-

In order to make our dataset more representative for real-world varying scenarios, we consider four intra-class variations: **(a)** The background varies from simple scene (clean background) to complex one (including real-life background with moving humans). For some clips with simple background, there are only the person performing actions with clean background, as shown in Fig. 2(a). On the contrary, for some other clips with complex background, there are interrupting pedestrian activities concurring with the action, as shown in Fig. 2(b)-(f). **(b)** We specify 2-3 video clips with over 50% occlusion in each class, as shown in Fig. 2(c). **(c)** The pose variation is considered even for the same person, as shown in Fig. 2(d)-(f). **(d)** The viewpoint variation is also considered. Around

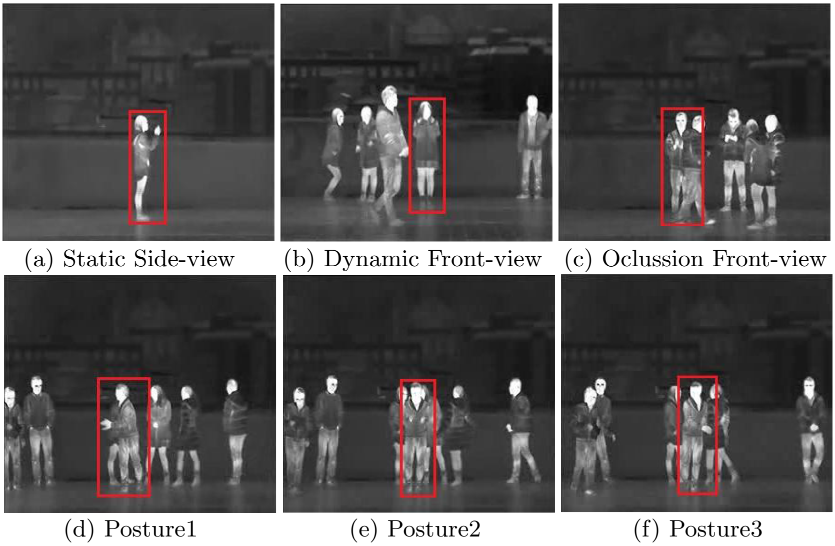


Fig. 2. Examples of intra-class variations of the handclapping action. (a) the ideal case with side-view angle and single person. (b) and (c) two cases with dynamic and occlusion variations in the background. (d), (e) and (f) cases with different postures.

20 video clips are taken under the front-view, and the remaining are taken under side-view, as shown in Fig. 2(a) and (b)-(c).

3 Evaluation Pipeline

3.1 Local Descriptors

Seven widely-employed low-level local descriptors are extracted from infrared video, including STIP [18], HOG3D [14], 3DSIFT [23], trajectory feature TRAJ [27], appearance feature HOG [4], and motion features HOF [19] and MBH [5]. Combination of TRAJ, HOG, HOF and MBH forms the dense trajectory feature [27], denoted as Dense-traj. In order to further introduce our evaluation settings, we briefly review these descriptors below.

STIP: The spatio-temporal interest points (STIP) is proposed by Laptev et al. [18] based on the idea from the Harris and Forstner interest point operators [11], which is widely used as a video representation to handle videos with complex and dynamic background recently [31]. As actions often have characteristic extending both in spatial and temporal domain, Laptev el al. extended the notion of interest points into the spatio-temporal domain and adapted both spatial and temporal scales of the detected features. In our experiment, we use the off-the-shelf binary package available online to extract this feature.

HOG3D: This feature is the local descriptor proposed by Klaser et al. [14], which is based on histograms of oriented 3D spatio-temporal gradients. It computes 3D gradient from an integral video representation. Then regular polyhedrons are used to quantize orientation of 3D gradient. The author divided a 3D patch from videos into $n_x \times n_y \times n_t$. The corresponding descriptor concatenates gradient histograms of all cells and is then normalized. In our paper, we firstly detect interest points with Harris corner detector, and then represent them with the HOG3D descriptor. We use executable programs from the author website and apply their recommend parameter setting as: $n_x = n_y = 4, n_t = 3$, where n_x, n_y and n_t are numbers of spatial and temporal cells, respectively.

3DSIFT: The 3 Dimensional Scale-Invariance Feature Transform (3DSIFT) descriptor was proposed by Scovanner el at. [23] which encodes gradient characteristics in 3 dimensional space. First the gradient magnitude and orientations in 3D are computed, and then A sub-histogram is created by sampling subregions surrounding the interest points. For each sub-region the orientations are accumulated into sub-histogram. The final descriptor is a vectorization of the sub-histogram. Here we detect interest points using the Harris Corner detector. We use the publicly available code from the Scovanner's website with the suggested parameter settings.

TRAJ: Want et al. [27] put forward a trajectory descriptor which encodes local motion of the densely-sampled interest points. The trajectory shapre is described by the sequence of the relative motion between every two consecutive frames, and the feature points are sampled on the trajectory with a fixed number of frames. This feature can well capture the motion characteristics of the video, which is significant in action recognition.

HOG/HOF: The Histogram of Gradients (HOG) and the Histogram of Optical flow (HOF) descriptors are introduced by Laptev et al. [19]. The authors compute histograms of spatial gradient and optical flow accumulated in space-time neighborhoods of detected points, which can be detected using any interest point detectors [7,18]. In our experiment, these points are selected along the motion trajectory [27] and features are computed within a $N \times N$ volume around these points. Each volume is subdivided into a space-time grid of size $n_\sigma \times n_\sigma \times n_T$. The default parameters for our experiments are $N = 32, n_\sigma = 2, n_T = 3$.

MBH: The Motion Boundary Histogram (MBH) descriptor is proposed in the work of Dalal et al. [5], where derivatives are computed separately for the horizontal and vertical components of the optical flow. The descriptor encodes the relative motion between pixels. Since MBH represents the gradient of the optical flow, constant motion information is suppressed and only the information concerning changes in the flow field (i.e., motion boundaries) is kept. This descriptor yields good performance when combined with other local descriptors. In our evaluation, we used the same MBH parameters as used in the work of Wang et al. [27].

3.2 Feature Encoding Methods

In this paper, two encoding methods, namely the Fisher Vector [20] and the Vector of Locally Aggregated Descriptors (VLAD) [12], are used. The former utilizes Maximum Likelihood (ML) estimation to train a Gaussian mixture model (GMM), which is later employed to form the description of low level features. The latter, however, utilizes the k-means technique to assign each feature to the closest cluster of a vocabulary with size k.

3.3 Fusion Strategy

At present, early-fusion and late-fusion are the basic feature fusion strategies. Early fusion [25] combines multiple features before classification. In our work, the concatenation of multiple features is employed since it is a commonly-used way in early fusion. Late fusion [17] requires more computation, since it combines the outputs of each type of feature. In our work, the average of the output scores is adopted for late-fusion.

4 Experiments

In this section, we first describe the implementation details, and then present the evaluation results of low-level local descriptors, including the encoding strategies on our IAD dataset. Besides, the fusion strategies are also evaluated.

4.1 Implementation Details

In our experiments, we follow the widely-used pipeline of raw feature extraction, feature encoding and classifier training in general AR systems. Basically, the raw feature extraction adopted the off-the-shelf coding and the default parameter configure as aforementioned. For the Fisher Vector, the number K of the Gaussian distributions model is an important parameter. We tested many values and empirically found that $K = 90$ can get relative better performance. For the VLAD, the size K of the codebook is also empirically determined as 500. We adopted the SVM [6] as the classifier and the libSVM [2] software in our experiments. We tested two kernels of Linear kernel and RBF kernel for two coding methods. The corresponding optimal parameters C and γ are obtained using 5 fold cross validations with a grid searching algorithm. Using the Fisher Vector and VLAD encoding methods, the searching results are as follows: For the linear kernel the optimal C is 30 and 80, respectively, and for the RBF kernel, the optimal C is 50 and 8, and γ is 2.7×10^{-3} and 5.0×10^{-1}, respectively.

4.2 Evaluations on Low-Level Local Descriptors

We evaluate 8 local feature descriptors as aforementioned with respect to different coding methods and different classifier kernels. For each evaluation, we randomly select 20 samples as the training set out of a sample set containing 30 samples and the rest 10 samples are used as the test set. We conduct the experiments of the same settings five times and the average is used as the final result. All evaluation results are shown in Table 2.

Table 2. The average precision (%) of different local descriptors with different kernels and coding methods.

Descriptor	Fisher Vector		VLAD	
	Linear	RBF	Linear	RBF
TRAJ	55.74	51.66	62.5	60.49
Dense-Traj	**68.15**	65.83	**74.83**	72
HOG	48.61	43.66	52.5	50.83
HOF	66.94	65.5	69.16	69.16
MBH	64.53	64.16	70	68.83
STIP	62.66	45.66	61.83	58.16
HOG3D	57.16	37.83	56.66	56
3DSIFT	53.66	20.33	58.33	54.16

From Table 2, we can observe that the best performance is from the Dense-traj [27]. This is basically in accordance with the situation on some other available datasets of visible spectrum [3]. Overall, the performance of the HOG is relatively bad among these descriptors. The reason may be caused by the lack of local texture information in infrared images (please see Fig. 2). Since the HOG descriptor is good at appearance description, its strength may not be revealed in the situations where local texture is relatively weak.

It can be also observed that the performance of linear kernel is much better than the RBF one, especially for those descriptors with higher feature dimensions (e.g., HOG3D, 3DSIFT) under the Fisher vector coding strategy. One possible reason is that the RBF kernel causes over-fitting in our task. It is also interesting to see that the performance of VLAD is better than Fisher Vector. This may be the fact that the codewords generated by HOG3D, MBH, etc. are not suitable to be described as the GMM model.

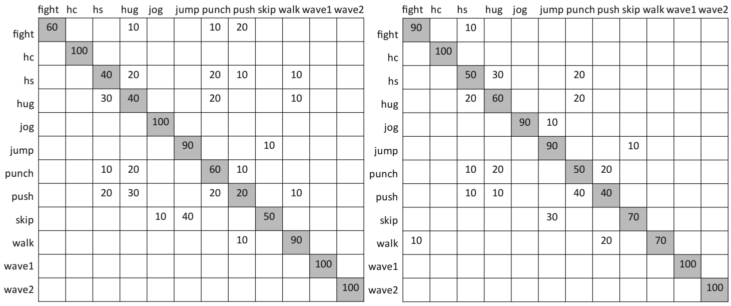
4.3 Early Fusion and Late Fusion

The early fusion and late fusion strategies are evaluated on the IAD dataset. We mainly consider STIP, TRAJ, HOG, HOF and MBH descriptors since they are of different and complementary types. The combinations of different numbers of features are tested, respectively, and the results are shown in Table 3. We can observe that the late fusion benefits more to the overall performance than the early fusion. Besides, the number of features for fusion does not determinate the final performance. In our case, the best performance for both fusion strategies is obtained when using STIP, HOF and MBH.

In order to further explore the classification performance for each action, we illustrate two confusion matrices shown in Fig. 3. The left one is the result of early fusion of STIP and HOF, and the right one is the result of late fusion of STIP and MBH. It can be seen that four actions of handshaking, hugging, punching and pushing have relative lower precisions. These actions are easily confused with other actions, e.g., handshaking and hugging, punching and pushing, pushing and hugging, etc. Fig. 4 shows two pairs of frames from four action videos. From the left pair

Table 3. The evaluation results (AP) of Early vs. Late Fusion with the same coding method of fisher vector.

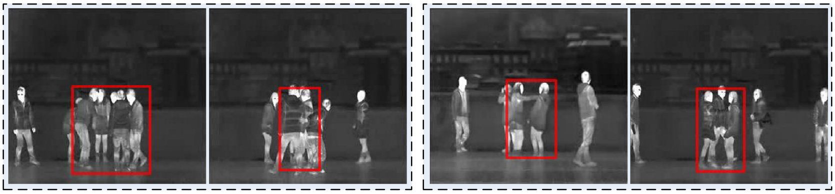
Fusion Type	Descriptor	Early Fusion	Late Fusion
Two features fusion	STIP+TRAJ	63.33	62.5
	STIP+HOG	58.33	61.67
	STIP+HOF	70.83	74.17
	STIP+MBH	69.16	75.83
Three features fusion	STIP+TRAJ+HOG	63.33	65
	STIP+HOG+HOF	70	72.5
	STIP+HOF+MBH	78.33	77.50
Four feature fusion	STIP+TRAJ+HOG+HOF	70.83	71.67
	STIP+HOG+HOF+MBH	72.5	72.5
Five feature fusion	STIP+TRAJ+HOG+HOF+MBH	73.33	72.5



	fight	hc	hs	hug	jog	jump	punch	push	skip	walk	wave1	wave2
fight	60			10			10	20				
hc		100										
hs			40	20			20	10		10		
hug			30	40			20			10		
jog					100							
jump						90			10			
punch			10	20			60	10				
push			20	30			20	20		10		
skip					10	40			50			
walk								10		90		
wave1											100	
wave2												100

	fight	hc	hs	hug	jog	jump	punch	push	skip	walk	wave1	wave2
fight	90		10									
hc		100										
hs			50	30			20					
hug			20	60			20					
jog					90	10						
jump						90			10			
punch			10	20			50	20				
push			10	10			40	40				
skip						30			70			
walk	10							20		70		
wave1											100	
wave2												100

Fig. 3. The comparative results of two fusion strategies, where the left is from early fusion strategy of STIP and HOF, while the right is from the late fusion strategy of STIP and MBH. Note that "hc" stands for "handclapping", "hs" stands for "handshake".



handshaking hugging punching pushing

Fig. 4. Two pairs of easily confused actions. The left shows two actions of handshaking and hugging with heavy occlusion, while the right shows two similar actions of punching and pushing.

of frames, we can see that the handshaking and hugging actions are both occluded by crowded persons around. These background clutter would bring big confusion. From the right pair, we can see that punching and pushing are so similar that it may even be deceitful for human eyes to recognize.

5 Conclusion

In this paper, we introduce a new infrared action dataset and evaluate the state-of-the-art AR pipeline on it. The evaluation results reveal that the dense trajectory feature can achieve the best performance on our dataset and the appearance features have relative poorer performance. Besides, the coding method of vector of locally aggregated descriptors is better than the widely-used fisher vector, and the late fusion benefits more to performance than early fusion. In addition, the best average precision on our infrared action dataset is around 70%, which leaves sufficient space for promoting new infrared-oriented AR methods.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61102131, 61571071), the Natural Science Foundation of Chongqing Science and Technology Commission (No. cstc2014jcyjA40048), Cooperation of Industry, Education and Academy of Chongqing University of Posts and Telecommunications (No. WF201404).

References

1. Aggarwal, J., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys (CSUR) **43**(3), 16 (2011)
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm
3. Chaquet, J.M., Carmona, E.J., Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding **117**(6), 633–659 (2013)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
5. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
6. Dikmen, M., Ning, H., Lin, D.J., Cao, L., Le, V., Tsai, S.F., Lin, K.H., Li, Z., Yang, J., Huang, T.S., et al.: Surveillance event detection. In: TRECVID (2008)
7. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. IEEE (2005)
8. Gao, C., Yang, L., Du, Y., Feng, Z., Liu, J.: From constrained to unconstrained datasets: an evaluation of local action descriptors and fusion strategies for interaction recognition. In: World Wide Web, pp. 1–12 (2015)
9. Han, J., Bhanu, B.: Human activity recognition in thermal infrared imagery. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, CVPR Workshops 2005, p. 17. IEEE (2005)
10. Han, J., Bhanu, B.: Fusion of color and infrared video for moving human detection. Pattern Recognition **40**(6), 1771–1784 (2007)

11. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, Manchester, UK, vol. 15, p. 50 (1988)
12. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3304–3311. IEEE (2010)
13. Klare, B.F., Jain, A.K.: Heterogeneous face recognition using kernel prototype similarities. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(6), 1410–1422 (2013)
14. Klaser, A., Marszalek, M.: A spatio-temporal descriptor based on 3d-gradients (2008)
15. Kuehne, H., Jhuang, H., Stiefelhagen, R., Serre, T.: Hmdb51: A large video database for human motion recognition. In: High Performance Computing in Science and Engineering 2012, pp. 571–582. Springer (2013)
16. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2556–2563. IEEE (2011)
17. Lan, Z., Bao, L., Yu, S.-I., Liu, W., Hauptmann, A.G.: Double fusion for multimedia event detection. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 173–185. Springer, Heidelberg (2012)
18. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2–3), 107–123 (2005)
19. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
20. 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, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
21. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1234–1241. IEEE (2012)
22. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3, pp. 32–36. IEEE (2004)
23. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th International Conference on Multimedia, pp. 357–360. ACM (2007)
24. Shao, L., Zhen, X., Tao, D., Li, X.: Spatio-temporal laplacian pyramid coding for action recognition. IEEE Transactions on Cybernetics **44**(6), 817–827 (2014)
25. Snoek, C.G., Worring, M., Smeulders, A.W.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 399–402. ACM (2005)
26. 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 (2012)
27. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE (2011)
28. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103**(1), 60–79 (2013)

29. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3551–3558. IEEE (2013)
30. Wang, J.T., Chen, D.B., Chen, H.Y., Yang, J.Y.: On pedestrian detection and tracking in infrared videos. Pattern Recognition Letters **33**(6), 775–785 (2012)
31. Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2834–2841. IEEE (2013)

Building Segmentation and Classification from Aerial LiDAR via Local Planar Features

Jun Chu[1,2](✉), Wei Han[1,2,3], Wei Sui[3], Lingfeng Wang[3], Qi Wen[4], and Chunhong Pan[3]

[1] Institute of Computer Vision, Nanchang Hangkong University, Nanchang, China
chuj@nchu.edu.cn
[2] Key Laborator of Jiangxi Province for Image Processing and Pattern Recongnition, Nanchang, China
[3] NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4] National Disaster Reduction Center of China, Ministry of Civil Affairs, Beijing, China

Abstract. In this paper, we propose a framework on building segmentation and classification from Aerial Lidar data via planar features. In this framework, the planar points corresponding to planar objects are obtained first by an unsupervised Markov random field clustering model. The ground normal is detected from planar points via the proposed constrained K-means algorithm. Within constrained K-means algorithm, the building points are generated by removing ground points from planar points. Furthermore, the candidate buildings are obtained by using region growing algorithm. Finally, these candidate buildings are classified into two types, that is, abnormal building and normal building based on the proposed vertical feature. Experimental results on a real world dataset demonstrate the effectiveness of our framework.

Keywords: Building segmentation · Planar objects · Aerial lidar data · Ground detection

1 Introduction

Earthquake and flood have taken place frequently over the world and brought disasters to natives. Many buildings are collapsed and damaged in the affected areas. In practice, it is difficult to measure and evaluate the damaged condition of buildings by manpower. Many algorithms based on computer vision have been proposed, among which image-based approaches are widely used. However, image-based approaches may not be applied to real-world problems effectively because the image acquisition is susceptible to lighting conditions. By contrast, 3D point cloud is robust to light. Hence, we propose a method to detect and classify buildings from 3D point cloud.

J. Chu—This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61263046, 61403376, 61175025 and 41301485).

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 313–322, 2015.
DOI: 10.1007/978-3-662-48570-5_31

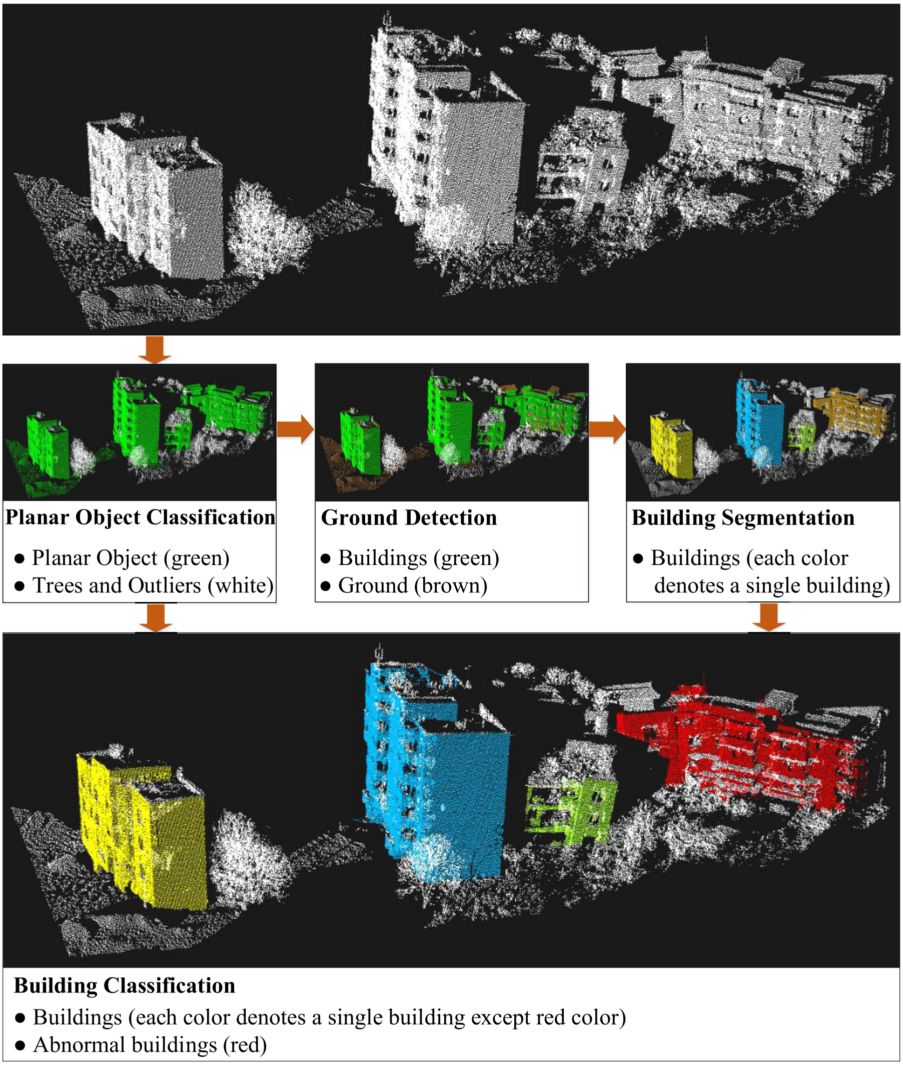

Planar Object Classification
● Planar Object (green)
● Trees and Outliers (white)
Ground Detection
● Buildings (green)
● Ground (brown)
Building Segmentation
● Buildings (each color denotes a single building)
Building Classification
● Buildings (each color denotes a single building except red color)
● Abnormal buildings (red)

Fig. 1. A framework of building detection and classification.

1.1 Related Work

Building detection and classification approaches can be mainly divided into two categories, i.e., supervised learning based methods [1–4], and unsupervised learning based methods [5–9].

For the supervised learning based approaches, the features for points classification are learned by fitting a mixture of Gaussian model by Charaniya et al. [1] and Lalonde et al. [2] [3]. Secord et al. [4] proposed a method based on support vector machines for object detection using aerial lidar and image data.

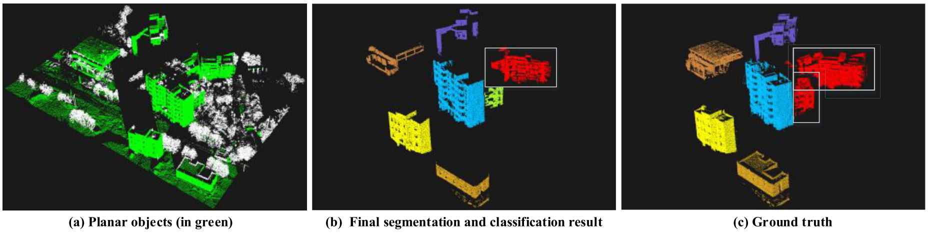

(a) Planar objects (in green)
(b) Final segmentation and classification result
(c) Ground truth

Fig. 2. One segmentation and classification result on **data1**. In sub-figures (b) and (c), each color represents a normal building except the red color. The red color represents abnormal buildings. Specifically, for red color buildings, each white box is a single abnormal building.

This method can provide satisfactory result. Unfortunately, the training sets are hard to be labeled due to the unstructured distribution of 3D point cloud.

The unsupervised learning approaches directly use the scatter or elevation for object including building detection and classification from 3D point cloud. Generally, for feature extraction, they focused on the neighborhood size of each point. For example, Weinmann et al. [7] minimized a energy function to obtain the optimal neighborhood size with the unknown density of 3D point cloud. The main limitation of this method is its computational complexity and fixing-density assumption. For the application of building detection, Carlberg et al. [5] and Lafarge et al. [6] used 3D point cloud to detect buildings in urban scenes. Specifically, Carlberg et al. [5] used a multi-category classifier to classify water, ground, roof, and trees. The height information has been used to remove ground and water under the assumption that the ground and water are in low height. Buildings and trees are detected through 3D shape analysis and region growing. Lafarge et al. [6] combined the features of local non-planarity, elevation, scatter and regular grouping to classify 3D point cloud data into buildings, vegetation, ground and clutters. The constructive solid geometry is then used to reconstruct buildings. The region growing segmentation and gradient orientation segmentation algorithms are used for classification of building, ground and vegetation in [8]. Matei et al. [9] proposed a building segmentation method by using error back propagation algorithm. These methods perform well on urban scenes, however, they may not be robust for rural scenes, where houses are built with different elevations.

1.2 Our Method

The framework of our method is illustrated in Fig. 1. The planar objects are first detected from 3D point cloud by a Markov random field (MRF) based model, which is motivated by [6] . Then, the ground normal is extracted by constrained K-means to further remove the ground points. After that, region growing method is used to obtain a single building and remove some of the clusters. Finally, the vertical feature is utilized for building classification.

The main contributions are highlighted as follows:

1. We propose an unsupervised building detection and classification framework based on planar features. Experimental results show that our method can provide satisfactory results.

2. A new dataset, which includes major difficulties in building detection and classification, is created to evaluated performance of our method. For quantitative evaluation, we have labeled the ground truth including all normal and abnormal buildings.

2 Local Planer Feature

In this section, we introduce the features used in our method. Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ be the set of input noise-free 3D points[1] in the scene, where N is the number of points. For each point $\mathbf{x}_i$, we first consider a subset $\mathcal{X}_i = \{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \cdots, \mathbf{x}_{i,K}\} \subset \mathcal{X}$ as its K-nearest neighbors (K-NN). Then, the covariance matrix $\mathbf{C}_i \in \mathbb{R}^{3\times 3}$ of these K-NN points $\mathcal{X}_i$ is calculated by

$$\mathbf{C}_i = \sum_{k=1}^{K} (\mathbf{x}_{i,k} - \mu_i)(\mathbf{x}_{i,k} - \mu_i)^{\mathsf{T}} , \tag{1}$$

where $\mu_i = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_{i,k}$ is the mean of the K-NN points. After that, we can obtain the eigenvalues $\{\lambda_{i,j}\}_{j=1}^{3}$ of the covariance matrix $\mathbf{C}_i$ and the corresponding eigenvectors $\{\mathbf{u}_{i,j}\}_{j=1}^{3}$ by performing singular value decomposition, namely,

$$\mathbf{C}_i = \sum_{j=1}^{3} \lambda_{i,j} \mathbf{u}_{i,j} \mathbf{u}_{i,j}^{\mathsf{T}} . \tag{2}$$

Without loss of generality, we assume that $\lambda_{i,1} \geq \lambda_{i,2} \geq \lambda_{i,3}$. The local planar features of the point $\mathbf{x}_i$ are composed of two parts, that is,

$$\{f_i, \mathbf{n}_i\} = \left\{ \frac{\lambda_{i,2}}{\lambda_{i,3} + \epsilon}, \mathbf{u}_{i,3} \right\} ,$$

where $\epsilon = 10^{-5}$ is a small positive value to prevent dividing by zero. On the one hand, f_i represents the degree of planarization of the K-NN points $\mathcal{X}_i$ and the larger f_i is, the closer the plane distribution is. On the other hand, $\mathbf{n}_i$ is the normal of this potential plane.

3 The Proposed Framework

3.1 MRF-Based Planar Object Classification

Let $\mathcal{L} = \{0, 1\}$ be two class labels to represent whether a 3D point locates on the planar object. Denote

$$\mathcal{Z} = \{z_1, z_2, \cdots, z_N\}$$

[1] The noise is removed by applying the method proposed in [10].

as a potential classification result of all points $\mathcal{X}$, where $z_i \in \mathcal{L}$ is the class label of the i-th point. Note that, $z_i = 1$ represents the i-th point that locates on a planar object. The MRF model [11] used in this work is defined as

$$E(\mathcal{Z}) = \sum_{i=1}^{N} D(z_i) + \gamma \sum_{i \sim j} S(z_i, z_j) \ , \tag{3}$$

where $D(\cdot)$ is the data term, $S(\cdot, \cdot)$ is the pairwise smooth term, $i \sim j$ represents the pairs of neighboring points, and γ is a weighting constant. In this paper, the data term is defined as

$$D(z_i) = \begin{cases} 1, & f_i \geq \theta \\ 0, & \text{otherwise} \end{cases} \ , \tag{4}$$

where θ is a positive threshold. The smooth term $S(z_i, z_j)$ is considered as the Potts model, given by

$$S(z_i, z_j) = \begin{cases} 1, & z_i = z_j \\ 0, & \text{otherwise} \end{cases} \ . \tag{5}$$

During implementation, the Graph-cut (GC) algorithm [12] is utilized to solve Eqn. (3).

After performing GC algorithm, all points $\mathcal{X}$ can be divided into two subsets, i.e., $\mathcal{X}^0$ and $\mathcal{X}^1$, satisfying that

$$\mathcal{X}^0 \cup \mathcal{X}^1 = \mathcal{X}, \quad \mathcal{X}^0 \cap \mathcal{X}^1 = \emptyset \ ,$$

where $\mathcal{X}^0 = \{\mathbf{x}_i | z_i = 0\}_{i=1}^{N}$ and $\mathcal{X}^1 = \{\mathbf{x}_i | z_i = 1\}_{i=1}^{N}$. The points in subset $\mathcal{X}^1$ represent planar objects, including buildings and ground (see Fig. 2). To segment and classify buildings, the ground should be removed beforehand. The details of removal of the ground is presented in the following subsection.

3.2 Ground Detection via Constrained K-Means

In real application, the normals of buildings and ground are mutually perpendicular. Based on this observation, the ground is detected by classifying all normals into two types, i.e., ground normal and building normal. To obtain these two types of normals, we propose the constrained K-means clustering model, which is defined as

$$\begin{aligned} &\min_{\{\mathbf{m}_j\}_{j=1}^{M}} \sum_{i=1}^{N} \sum_{j=1}^{M} \|\mathbf{n}_i - \mathbf{m}_j\|_2^2 \delta_{i,j} \ , \\ &\text{s.t.} \ \ \|\mathbf{m}_j\|_2^2 = 1, \quad \forall j \in \{1, 2, \cdots, M\} \ , \end{aligned} \tag{6}$$

where M is the number of clusters, $\{\mathbf{m}_j\}_{j=1}^{M}$ are M cluster centers, and $\delta_{i,j}$ is a Dirichlet function, satisfying that $\delta_{i,j} = 1$ if the i-th point belongs to the j-th

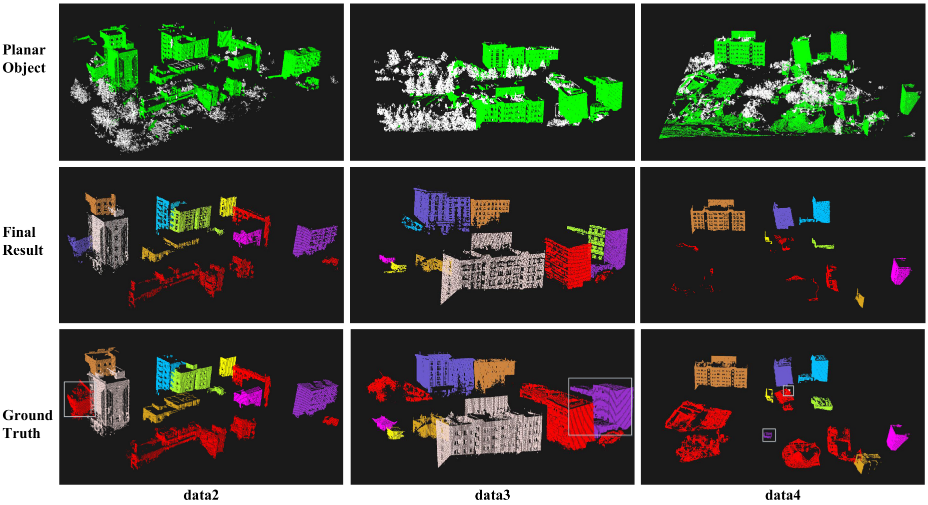


Fig. 3. Building segmentation and classification results on the other three data (each column). The first row represents the planar object by MRF. The second row shows our segmentation and classification results. The last row illustrates the ground truths labeled manually. The color indexes are the same with those of Fig. 2.

cluster, otherwise $\delta_{i,j} = 0$. In Eqn. (6), the constraint is to restrict the cluster centers to be normals. During implementation, the number of clusters M is set to be 3.

Optimization: The proposed constrained K-means in Eqn. (6) is optimized by iterating the following two steps (the max number of iterations is set to be 10):

Step_1: Computing normal clusters $\{\mathbf{m}_j\}_{j=1}^{M}$ by K-means.

Step_2: Restricting each normal cluster $\mathbf{m}_j$ with the normalizing operation, that is, $\mathbf{m}_j = \frac{\mathbf{m}_j}{\|\mathbf{m}_j\|_2}$.

Ground Detection: Let $\mathcal{I} = \{1, 2, \cdots, M\}$ be the indexes of all the normal clusters obtained by the proposed constrained K-means. In order to detect ground normal, we first define a perpendicular value for each normal cluster. For example, the perpendicular value p_i for the i-th normal cluster is calculated by

$$p_i = \sum_{j \in \{\mathcal{I}/i\}} \mathbf{m}_i^{\mathsf{T}} \mathbf{m}_j \ , \tag{7}$$

where $\mathcal{I}/i$ means all indexes except the index i. The index of the ground normal $i^\star$ is computed as the index with the minimal perpendicular value, given by

$$i^\star = \min_i \{p_1, p_2, \cdots, p_M\} \ . \tag{8}$$

Hence, the $i^\star$-th normal cluster $\mathbf{m}_{i^\star}$ is the ground normal.

For the i-th point, we first calculate the inner product q_i between its normal $\mathbf{n}_i$ and the ground normal $\mathbf{m}_{i^\star}$, that is, $q_i = \mathbf{n}_i^\mathsf{T}\mathbf{m}_{i^\star}$. The i-th point is detected as ground point if $|q_i| > \tau$, where τ is a positive threshold. During implementation, the threshold τ is set to 0.3 experimentally. After removing the ground points, the rest are building points, which are denoted as $\mathcal{X}_B$ (see Fig. 2).

3.3 Region Growing for Building Segmentation

The region growing method is used to segment candidate buildings. Specifically, after performing region growing algorithm, the building points $\mathcal{X}_B$ are clustered into C non-overlapped clusters, that is,

$$\mathcal{X}_B = \bigcup_{c=1}^{C} \mathcal{X}_B^c, \ \ \mathcal{X}_B^i \cap \mathcal{X}_B^j = \emptyset, 1 \leq \ i \neq j \leq C \ , \tag{9}$$

where $\mathcal{X}_B^c$ is the c-th candidate building. We remove the small clusters with the number of points less than a threshold ξ. During implementation, the threshold ξ is set to 1000.

3.4 Building Classification

In practice, the tilt angles between abnormal and normal buildings are different. Based on this fact, we first compute a vertical feature for each candidate building. For example, the vertical feature v_c for the c-th candidate building $\mathcal{X}_B^c$ is calculated by

$$v_c = \frac{1}{N_c} \sum\nolimits_{j=1}^{N_c} \mathbf{m}_{i^\star}^\mathsf{T} \mathbf{n}_j \ , \tag{10}$$

where $N_c = |\mathcal{X}_B^c|$ is the number of points on c-th building and $\mathbf{m}_{i^\star}$ is the ground normal obtained by the constrained K-means (see Subsection 3.2). Then, a thresholding operation is used to classify all candidate buildings into two classes, that is, normal building and abnormal building, given by

$$l_c = \begin{cases} \text{normal building,} & v_c \geq \beta \\ \text{abnormal building,} & \text{otherwise} \end{cases} , \tag{11}$$

where l_c is the label of the c-th candidate building and β is a positive threshold (refer to Fig. 2 as example).

4 Experiments

In this section, we evaluate our building segmentation and classification method on the real dataset, which is obtained from Aerial LiDAR. Our algorithm is implemented in C++ on platform Intel(R) Core(TM) i3-2100 CPU @ 3.10GHz with 4GB RAM.

Table 1. The description on the dataset.

	data1	data2	data3	data4
Num. of 3D Points	335172	209353	441796	534735
Num. of Buildings	7	14	10	15
Num. of Abnormal Buildings	2	6	3	6

4.1 Parameter Setting

Three major parameters, i.e., γ, θ and β are used for building segmentation and classification. During implementation, they are experimentally set as follows: 1) parameters γ and θ are set to $\gamma = 7500$ and $\theta = 25$ for MRF-based planar object classification; 2) in building classification, the threshold of β is set to $\beta = 0.08$.

4.2 Dataset Description

The dataset obtained from Aerial LiDAR is created to evaluate the effectiveness of our method (refer to Fig. 2 and Fig. 3 for visual perception). The 3D point numbers of them are 335172, 209353, 441796 and 534735, respectively. They contain 5, 8, 7 and 9 normal buildings, as well as 2, 6, 3 and 6 abnormal buildings (refer to Table 1). We manually labeled the ground truth for quantitative evaluation. There are several difficulties, which make building segmentation and classification problems challenging.

- Buildings are built on mountains. On the one hand, the buildings to be segmented are often shaded by the other objects, which results in severe data missing. On the other hand, the elevations of them are different with each other. Hence, they do not locate on the same ground plane.
- Some abnormal buildings are seriously destroyed. In this case, it is very hard to distinguish them from clutter objects, such as trees. For some abnormal buildings, the inclination angle is small, which causes that it is very hard to distinguish them from the normal buildings.

4.3 Visual Results

The visual comparisons of our method with the ground truth are shown in Fig. 2 and Fig. 3. With the local planer feature and ground normal estimation, the buildings are segmented and classified accurately. It is very close to the ground truth. Although our method can provide satisfactory segmentation and classification results, it still has some small errors. As shown in Fig. 3, in **data3**, one building is segmented into two buildings (see the white rectangle in the third row). In **data4**, two buildings are not segmented out. The main reason is that

Table 2. The building classification results on the dataset.

	data1	data2	data3	data4	Total
Num. of Buildings	7/7	14/14	9/10	13/15	**93.48%**
Num. of Abnormal Buildings	1/2	5/6	2/3	6/6	**82.35%**

the numbers of building points are very small. However, in practice, our method works well in building segmentation and classification, and it can be used in real world application.

4.4 Quantitative Results

To further evaluation the effectiveness of our method, the TPR criterion is adopted to evaluate the segmentation and classification results quantitatively. The criterion TPR is defined as follows:

$$TPR = \frac{TP}{TP + FN}, \tag{12}$$

where TP is the number of correct segmented (or classified) buildings, while FN is the number of missed segmented (or classified) buildings. Table 2 demonstrates the results of our method on the dataset. As illustrated in this table, our method achieves high TPR in building segmentation and classification.

5 Conclusion

In this paper, we propose a novel method for building segmentation and classification via local planar feature. The core idea is to detect planar objects from clutter 3D points. To evaluate the effectiveness of our method, the dataset is created, which contains major difficulties in building detection and classification. Experimental results on this dataset demonstrate the effectiveness of our method.

However, our method still has some limitations. For example, the very small building may be miss-detected, and two near buildings would be detected as one building. In addition, as some abnormal buildings are destroyed very seriously, we can only detect a part of them. In the future, we will solve the above problems to make our method more general.

References

1. Charaniya, A.P., Manduchi, R., Roberto M., Lodha, S.K.: Supervised parametric classification of aerial lidar data. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop, pp. 25–32 (2004)

2. Lalonde, J.-F., Unnikrishnan, R., Vandapel, N., Hebert, M.: Scale selection for classification of point-sampled 3-d surfaces. In: Fifth International Conference on 3-D Digital Imaging and Modeling, pp. 285–292 (2005)
3. Lalonde, J.-F., Vandapel, N., Huber, D., Hebert, M.: Natural terrain classification using three-dimensional ladar data for ground robot mobility. Journal of Field Robotics **23**(10), 839–861 (2006)
4. Secord, J., Zakhor, A.: Tree detection in aerial lidar and image data. Tech. Rep. (2006)
5. Carlberg, M., Gao, P., Chen, G., Zakhor, A.: Classifying urban landscape in aerial lidar using 3d shape analysis. In: Proceedings of the International Conference on Image Processing, pp. 1701–1704 (2009)
6. Lafarge, F., Mallet, C.: Creating large-scale city models from 3d-point clouds: A robust approach with hybrid representation. International Journal of Computer Vision **99**(1), 69–85 (2012)
7. Weinmann, M., Jutzi, B., Mallet, C.: Semantic 3d scene interpretation: a framework combining optimal neighborhood size selection with relevant features. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences **II–3**, 181–188 (2014)
8. Forlani, G., Nardinocchi, C., Scaioni, M., Zingaretti, P.: Complete classification of raw lidar data and 3d reconstruction of buildings. Pattern Anal. Appl. **8**(4), 357–374 (2006)
9. Matei, B.C., Sawhney, H.S., Samarasekera, S., Kim, J., Kumar, R.: Building segmentation for densely built urban regions using aerial lidar data. In: Computer Vision and Pattern Recognition, pp. 1–8 (2008)
10. Wang, J., Kai, X., Liu, L., Cao, J., Liu, S., Zeyun, Y., Xianfeng David, G.: Consolidation of low-quality point clouds from outdoor scenes. Comput. Graph. Forum **32**(5), 207–216 (2013)
11. Li, S.Z.: Markov Random Field Modeling in Image Analysis, 3rd edn. Springer Publishing Company, Incorporated (2009)
12. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11), 1222–1239 (2001)

2DPCANet: Dayside Aurora Classification Based on Deep Learning

Zhonghua Jia[1,2], Bing Han[1,2(✉)], and Xinbo Gao[1]

[1] School of Electronic Engineering, Xidian University, Xi'an 710071, China
bhan@xidian.edu.cn, xbgao@mail.xidian.edu.cn
[2] State Key Laboratory of Remote Sensing Science, Beijing 100101, China

Abstract. The mysterious and beautiful aurora represents various physical meaning, thus the classification of aurora images have significant scientific value to human beings. Principal component analysis network (PCANet) has achieved good results in classification. But when using PCANet to extract the image features, it transform original image into a vector, so that the structure information of the image are missing. Compared with PCA, 2DPCA is based on 2D image matrices rather than 1D vectors so that 2DPCA can use the structure information of original image more efficiently and reduce the computational complexity. Based on PCANet, we develop a classification method of aurora images, 2-dimension PCANet (2DPCANet). To evaluate 2DPCANet performance, a series of experiments were performed on two different aurora databases. The classification rate across all experiments was higher using 2DPCANet than PCANet. The experiment results also indicated that the classification time is shorter using 2DPCANet than PCANet.

Keywords: Aurora image · Deep learning · Principle component analysis · PCANet · 2DPCANet

1 Introduction

Aurora is the magnificent light when the solar wind travels the magnetosphere of high altitude areas near the north and south poles of the earth. The aurora is not only related to the earth's atmosphere and geomagnetic field, but also related to the solar eruption of high-speed charged particles. When charged particles are emitted by the sun toward the earth into the scope of the earth magnetic field, they travel along the earth's magnetic field lines into the upper atmosphere near the north and south poles under the influence of the magnetic field, and then inspire visible light after proton collisions, and finally become a high-profile, we call it aurora.

Aurora phenomenon is not only a simple optical phenomenon, but also an important way for understanding the atmospheric physics. Different forms of the aurora imply different physical meanings. Therefore, the highly efficient classification of aurora images has very important value in scientific research.

From 1964 till now, Akasofu [1], Hongqiao Hu [2], and the Chinese polar research center [3] have divided the aurora into different types. For a long time, the aurora was divided into arc aurora and corona aurora. The corona aurora was further divided

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 323–334, 2015.
DOI: 10.1007/978-3-662-48570-5_32

into drapery corona, radial corona and hot-spot corona. In 2015, on the basis of arc aurora and corona aurora, the Chinese polar research center considers that corona aurora contains two types aurora: drapery corona and radial corona. Hot-spot corona aurora is regarded as anther aurora. So now the most significance aurora types are arc aurora, drapery corona aurora, and radial corona aurora. According to this, we classify these three types aurora effectively in this paper to discover the mechanism from them and provide an effective analysis for aurora physics research.

In 2004,Syrjasuoet al. [4] have firstly introduced the image processing and machine visual technology into the classification of the aurora image, and employ the texture feature of the aurora image to classify its shape as arc, block, omega and south-north shape. Since the texture structure of the aurora image in the complex background is not clear, the classification accuracy of the aurora image is not high. In 2007, Qian Wang et al. [5] have extracted the gray scale features of the aurora image by the principal component analysis (PCA) to divide it into the arc type, the corona type and the hybrid types, to achieve better classification efficiency. In 2008, Lingjun Gao et al. [6] have proposed the classification method of the aurora image based on Gabor transform to reduce the characteristic redundancy and to ensure the characteristic effectiveness and the classification. In 2009, Rong Fu et al. [7] has combined the analysis of the aurora image with the morphology to greatly improve the classification accuracy between the arc-like and corona aurora images. In 2010, Yuru Wang et al. [8, 9] have obtained the classification algorithm of the aurora image based on X-GLAM characteristics by modifying the GLAM field, in order to improve the classification precision, but its computational complexity is high. In 2013, Bing Han, et al. [10] employed the Salient Coding method to classify the features of aurora images, and Bing Han, et al. [11] proposed an aurora image classification method based on latent dirichlet allocation with saliency information, which improves the classification accuracy of the arc aurora.

According to the analysis of the existing algorithms, there are several critical problems on the highly effective classification of the aurora image, how to effectively extract the aurora feature, how to reduce the algorithm complexity and to improve the algorithm efficiency simultaneously.

With the development of the science and technology, convolutional deep neural network architecture [12], [13] consists of multiple trainable stages stacked on top of each other, followed by a supervised classifier. PCANet cascading two-layer PCA in [14] has been referred as a simple deep learning network to obtain the well efficiency in classification owing to that PCA in [15] has been referred as the data processing method with the advantages of revealing the essence of things and simplifying complex problems. However, in the PCA-based classification technique, the 2-dimention image matrices must be previously transformed into a vector, which ignores its structural information and leads to a high-dimensional image vector space. It is difficult to evaluate the covariance matrix accurately due to its large size and the relatively small number of training samples. While, 2-Dimention PCA (2DPCA) in [16] is a straightforward image projection technique for image feature extraction. Compared with PCA, 2DPCA is based on 2D image matrices rather than 1D vectors so that

2DPCA can use the structure information of original image more efficiently and reduce the computational complexity.

In this paper, we propose deep-learning classifications of the aurora image, two-staged 2DPCANet by full employing its structure information in order to increase the classification accuracy and to reduce the elapsed time.

The remainder of this paper is organized as follows: In section 2, the background of PCANet is reviewed. The ideal of the proposed two-staged 2DPCANet are described in section 3. In section 4, experimental results and analysis are presented. The final one is conclusion.

2 The PCA Network

PCANet is a deep learning network for image classification in [14]. It comprises components: cascaded PCA, binary hashing, and block-wise histograms and employed to learn multistage filter banks. This architecture can be designed and learned extremely easily and efficiently.

Assume that N input training images $\{\mathcal{I}_i\}_{i=1}^{N}$ of size $m \times n$ are given and trained in the PCANet system. The patch size is $k_1 \times k_2$ at all stages and suppose that the number of filters in layer i is L_i .

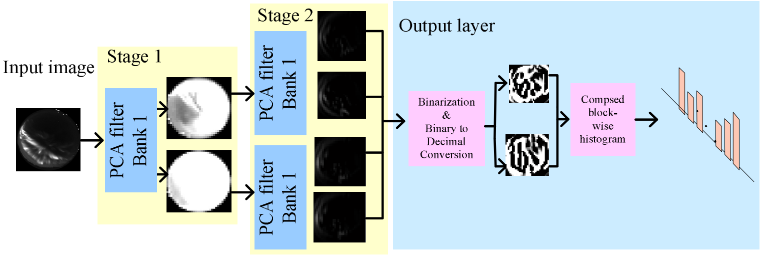


Fig. 1. Illustration of how the proposed PCANet extracts features from an image through three simplest processing components

2.1 The First Stage: PCA

1. Take a $k_1 \times k_2$ patch around each pixel and collect all patches of the i th input image: $x_{i,1}, x_{i,2}, \cdots, x_{i,\mathrm{mn}} \in R^{k_1 \times k_2}$,where $x_{i,j}$ denotes the j th patch in the i th input image;
2. Calculate patches mean and compute mean-removed patch from each patch to get $\overline{X}_i = \left[\overline{x}_{i,1}, \overline{x}_{i,2}, \cdots, \overline{x}_{i,\mathrm{mn}}\right]$, where $\overline{x}_{i,j}$ is a mean-removed patch.
3. By dealing all input image with the same steps and putting them together, we obtain $X = \left[\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_N\right] \in R^{(k_1 \times k_2) \times Nmn}$;
4. PCA minimizes the reconstruction error within a family of orthonormal filters, i.e., $\min_{V \in \mathbb{R}^{k_1 k_2 \times L_1}} \| X - VV^T X \|_F^2, s.t. V^T V = I_{L_1}$, where I_{L_1} is identity matrix of size $L_1 \times L_1$;

5. The PCA filters can be expressed as $W_l^1 = mat_{k_1,k_2}(q_l(XX^T)) \in R^{k_1 \times k_2}$, $l = 1,2,\cdots,L_1$, where $mat_{k_1,k_2}(v)$ is a function that maps v to a matrix W .

2.2 The Second Stage: PCA

Just like repeating the process as the first stage, the l th filter output of the first stage can be $\mathcal{I}_i^l = \mathcal{I}_i * W_l^1, i = 1,2,\cdots,N$. Just as the first stage, collect all the patches of $\mathcal{I}_i^l$, compute mean-removed patch from each patch and form $\overline{Y}_i^{\,l} = [\overline{y}_{i,l,1}, \overline{y}_{i,l,2}, \cdots, \overline{y}_{i,l,\mathrm{mn}}]$. The matrix collecting all mean-removal patch of the l th filter output can be defined as $Y^l = [\overline{Y}_1^l, \overline{Y}_2^l, \cdots, \overline{Y}_N^l] \in \mathrm{R}^{(\mathrm{k}_1 \times \mathrm{k}_2) \times Nmn}$, and concatenate Y^l for all the filter outputs as $Y = [Y^1, Y^2, \cdots, Y^{L_1}] \in \mathrm{R}^{(\mathrm{k}_1 \times \mathrm{k}_2) \times L_1 Nmn}$. Then, obtain the PCA filters of the second stage as $W_l^2 = mat_{k_1,k_2}(q_l(YY^T)) \in R^{k1 \times k2}$, $l = 1,2,\cdots,L_2$. We can achieve L_2 outputs for each input $\mathcal{I}_i^l$ of the second stage: $O_i^l = \left\{\mathcal{I}_i^l * W_l^2\right\}_{l=1}^{L_2}$.

2.3 Output Layer

For the outputs from the second stage, we binaries these outputs and get $\{H(\mathcal{I}_i^l * W_l^2)\}_{l=1}^{L_2}$. Convert binary bits of each pixel from each outputs bake into a decimal number. Then, we can get single integer-valued “images”: $\mathcal{T}_i^l = \sum_{l=1}^{L_2} 2^{l-1} H(\mathcal{I}_i^l * W_l^2)$. Partition each of the images into B blocks. Then, concatenate all the B histograms into one vector after compute the histogram of the decimal values in each block and denote as $Bhist\left(\mathcal{T}_i^{\,l}\right)$. After encoding process, the feature of the i th input image is defined as: $f_i = [Bhist\left(\mathcal{T}_i^1\right), \cdots, Bhist\left(\mathcal{T}_i^{L_1}\right)]^{\mathrm{T}} \in R^{\left(2^{L_2}\right) L_1 B}$.

3 The 2DPCA Network

2DPCA is based on 2-dimention image matrices rather than 1-dimention vectors so that the structure information can be fully considered and the dimension can be reduced. In this section, we adopted the cascaded 2DPCA , the binary hashing, and block histograms to classify the aurora images. The proposed 2DPCANet model is illustrated in Fig. 2.

Suppose that N input training images $\{\mathcal{I}_i\}_{i=1}^N$ of size $m \times n$ are given and the number of filters in layer i is L_i .

3.1 The First Stage: 2DPCA

1. For each image, subtract image mean from each image to obtain $\overline{X}_i$ and putting them together to get $X = \left[\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_N\right] \in R^{(m \times n) \times N}$;

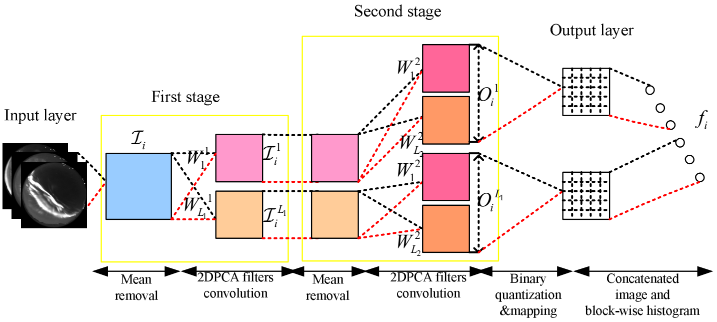


Fig. 2. The detailed block diagram of the proposed (two-stage) 2DPCANet.

2. 2DPCA minimizes the reconstruction error within a family of orthonormal filters, i.e., $\min_{V\in\mathbb{R}^{(m\times n)\times L_1}} \| X - VV^TX \|_F^2, s.t. V^TV = I_{L_1}$, where I_{L_1} is identity matrix of size $L_1 \times L_1$;
3. The 2DPCA filters can be expressed as $W_l^1 = mat_{m,n}(q_l(XX^T)) \in R^{m\times n}$, $l = 1,2,\cdots,L_1$, where $mat_{m,n}(v)$ is a function that maps v to a matrix W, and $q_l(XX^T)$ is the lth primal eigenvectors of matrix XX^T.

3.2 The Second Stage: 2DPCA

Almost repeating the same process as the first stage, set the lth filter output of the first stage is $\mathcal{I}_i^l = \mathcal{I}_i * W_l^1, i = 1,2,\cdots,N$. Before convolving $\mathcal{I}_i$ with W_l^1, the boundary of $\mathcal{I}_i$ is zero–padded. Just as the first stage, we define $Y^l = [\bar{Y}_1^l, \bar{Y}_2^l, \cdots, \bar{Y}_N^l] \in \mathrm{R}^{(m\times n)\times N}$ for the matrix collecting all mean-removed image of lth filter output, and concatenate Y^l for all the filter outputs as $Y = [Y^1, Y^2, \cdots, Y^{L_1}] \in \mathrm{R}^{(m\times n)\times L_1 N}$ .Then, obtain the 2DPCA filters of the second stage as $W_l^2 = mat_{m,n}(q_l(YY^T)) \in R^{m\times n}$, $l = 1,2,\cdots,L_2$. We will have L_2 outputs for each input of the second stage: $O_i^l = \left\{\mathcal{I}_i^l * W_l^2\right\}_{l=1}^{L_2}$.

One can simply repeat the above process to build more (2DPCA) stages.

3.3 Output Layer

Binaries the outputs from the second stage and get $\left\{H(\mathcal{I}_i^l * W_l^2)\right\}_{l=1}^{L_2}$. Convert binary bits of each pixel from each outputs bake into a decimal number to get $\mathcal{T}_i^l = \sum_{l=1}^{L_2} 2^{l-1} H(\mathcal{I}_i^l * W_l^2)$, whose every pixel is an integer in the range $\left[0, 2^{L_2-1}\right]$. Partition each of the images into B blocks. Then, concatenate all the B histograms into one vector after compute the histogram of the decimal values in each block and denote as

$Bhist\left(\mathcal{T}_i^l\right)$. After encoding process, the feature of the ith input image is defined as: $f_i = [Bhist\left(\mathcal{T}_i^1\right),\cdots,Bhist\left(\mathcal{T}_i^{L_1}\right)]^{\mathrm{T}} \in R^{\left(2^{L_2}\right)L_1 B}$.

4 Experiment Results and Analysis

In this section, the proposed 2DPCANet classification scheme is evaluated by conducting several classification and comparison experiments, and the experiment results are exhibited and analyzed then.

4.1 Dataset

The aurora data were obtained from the all-sky imagers at Yellow River Station (YRS) in Ny-lesund, Svalbard. The optical instruments at YRS capture photoemissions at 427.8, 557.7, and 630.0nm, and the time interval between every two images is 10 second. Dayside aurora images are divided into two categories: arc aurora, and corona aurora. The corona aurora can be further divided into drapery corona and radial corona. Sample images from the three aurora shape categories are shown in Fig. 3.

Experiments in this article, we adopt two kinds of aurora database. They were concentrated on the dayside aurora and selected from December 2003 to January 2004 to constitute the two different databases. Only auroras at 557.7 nm were adopted in consideration of the image characteristics.

The database I where images are unrelated in time domain contains 2400 aurora images which have 800 arc aurora images, 1600 corona aurora images (800 drapery corona and 800 radial corona). The aurora images of this database are all similar with standard aurora in morphology. The database II where images are related in time domain is larger than the database I and contains 11133 images. In database II, images are selected due to not only their morphology similarity, but also their physical development of an aurora event. That is to say, the time interval of images with same class is so small in the database II.

Compare two kinds of database on time and show in Fig. 4. We select the images form an aurora event. And different images with different time from this aurora event belong to the database I and the database II. The numbers over each aurora image represent the time when the image was captured. When the blank space in the row of the database I are filled with a tick, it illustrates that the aurora image over the blank space are belonged to the database I. And the same to the row of the database II.

On the basis of two different aurora databases, we randomly select different numbers of aurora images used for training and testing. In addition, the numbers of aurora images for training are three times than the numbers of aurora images for testing. The labels of arc, drapery corona and radial corona are 1, 2 and 3, respectively.

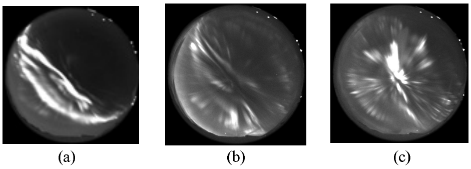
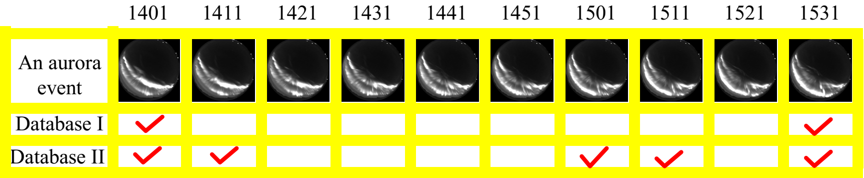

(a) (b) (c)

Fig. 3. Typical categories of aurora. Columns from left to right are: (a) arc, (b) drapery corona, and (c) radial corona.



	1401	1411	1421	1431	1441	1451	1501	1511	1521	1531
An aurora event										
Database I	✓									✓
Database II	✓	✓					✓	✓		✓

Fig. 4. An aurora event (arc aurora).

4.2 Classification Experiment for Parameters Setting

In the experiments, we evaluate the performance of 2DPCANet-3, 2DPCANet-2, 2DPCANet-1, PCANet-2, and PCANet-1. We select 600 aurora images used in the experiments. The numbers of arc, drapery corona, and radial corona are 200, respectively. We select randomly 150 aurora images form each category used for training and the rest used for testing. In this part, the experiments are handled used this data set.

We deal with the image classification used SVM classifier. Hence, the parameters in the SVM classifier should first be selected before the classification experiments are carried out. We conduct ten times cross validation. The selection results are shown in Fig.5. According to the performance of ten times cross validation, training and testing sets with optimal SVC parameters will be employed for constructing ROC curves.

Then, we should found the optimal number of filters in the different stages in different layers. We vary the number of filters in the first stage L_1 from 2 to 12 for one-staged networks. When considering two-staged networks, we set $L_2 = 8$ and vary L_1 from 4 to 24. At last, we set $L_2 = 14, L_3 = 8$ and vary L_1 from 4 to 20 to adjust the number of filters at the first stage in three-staged networks. The results of the number of filters are shown in Fig. 6. One can see that 2DPCANet achieves better results than PCANet. Moreover, the accuracy of 2DPCANet and PCANet (for all staged networks) increases for larger L_1. However, three-staged 2DPCANet achieves so lower classification accuracy rat.

The optimal number of filters and the optimal parameters in SVM will improve the performance of feature analysis methods.

For verifying the performance of the proposed method and the other models, we take a group of comparison experiments on the data set according to the optimal parameters that we select. Due to the performance of 2DPCANet-3, we only compare 2DPCANet-2, 2DPCANet-1, PCANet-2, and PCANet-1. The testing results of ROC curves are shown in Fig.7.

It can be seen from Fig. 7 that 2DPCANet-2 possesses the biggest area under curve which shows the best performance in classification. For getting more statistical and persuasive results, Table 1 shows the average classification accuracy rate of different methods. More intuitive results can be found in Table 1. Each method is conducted 200 times and the accuracy is the mean results of the 200 times classification procedures. We also find that 2DPCANet-2 performed better than others.

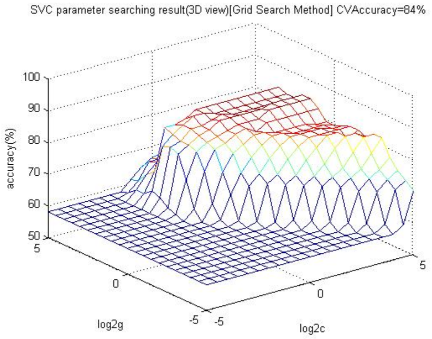


Fig. 5. SVM parameter selection with three dimensional view of (two-stage) 2DPCANet

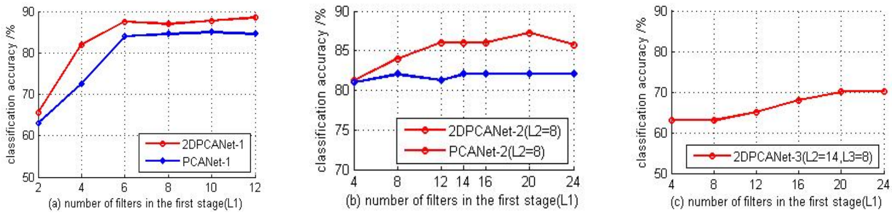


Fig. 6. Classification accuracy of 2DPCANet and PCANet for varying number of filters in the first stage

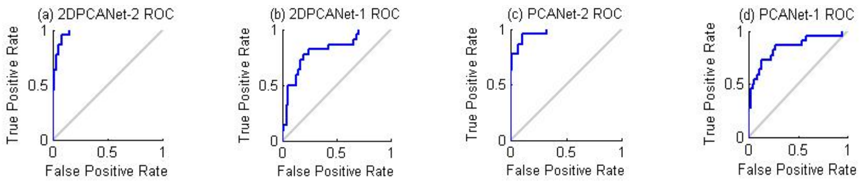


Fig. 7. The ROC curves of different classification methods

Table 1. The accuracy and generalized running time of different methods

Classification method	Feature dimensions	Accuracy (%)	Generalized running time(s)
2DPCANet-2	75264	**83**	**1.00**
2DPCANet-1	5376	69	**0.52**
PCANet-2	75264	80	1.10
PCANet-1	5376	64	0.61

4.3 Classification Experiment on the Database I

In order to evaluate the validity of the proposed method on the database I, experiments are designed and conducted. First, we select randomly different numbers of aurora images for experiments. The image datasets are shown in Table 2. 2DPCANet-2 is compared with 2DPCANet-1, PCANet-2, and PCANet-1.

We set the optimal parameters of all method and SVM classifier. Then, the experiments are conducted 200 times, and the final results are the average of them. Fig. 8 shows the performance of our method and the other methods with different numbers dataset. We observe that, the classification accuracy of our proposed method is higher than the other methods. And 2DPCANet-2 acquired smooth faster than others. In addition, we compared our method with PCANet-2 in the generalized running time, as shown in Table 3. It illustrates that our method spend less time than others and the advantage of our method are obviously with the increasing of the numbers of dataset.

Table 2. Dataset of aurora images

Num of data-set	2400	1800	1200	600	300	150	60
Arc	800	600	400	200	100	50	20
Hot-spot	800	600	400	200	100	50	20
Corona	800	600	400	200	100	50	20

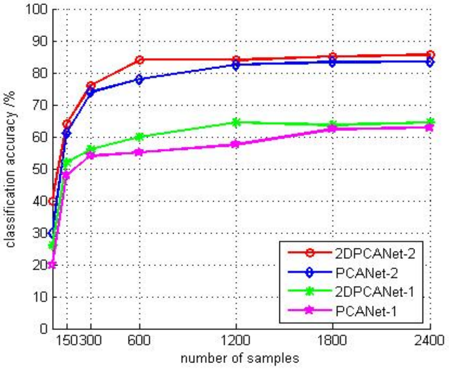


Fig. 8. Average classification accuracy of different representations on the database I

Table 3. The generalized running time between 2DPCANet and PCANet

Num of dataset	2400	1800	1200	600	300	150	60
2DPCANet-2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PCANet-2	1.20	1.19	1.17	1.03	1.06	1.07	1.02

4.4 Classification Experiment Based on the Database II

Imitate the experiments on the database I. We also select randomly different numbers of aurora images consisting datasets. As well as selecting the 8 sets from the database I, we select three more datasets from the database II to experiment. Compared 2DPCANet-2 with 2DPCANet-1, PCANet-2, and PCANet-1, we obtain the classification accuracy shown in Fig. 9 and generalized running time shown in Table 3. Obviously, the classification accuracy of 2DPCANet-2 is higher and the running time is smaller than other methods.

In contrast to PCANet, 2DPCANet has two important advantages. First, it reduces the data dimension of the aurora image by fully considering its structure information. Second, it required less time to determine the corresponding eigenvectors.

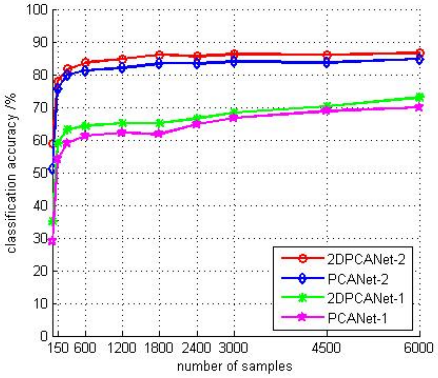


Fig. 9. Average classification accuracy of different representations on the database II

Table 4. The generalized running time between 2DPCANet and PCANet

Num of dataset	6000	4500	3000	2400	1800	1200	600	300	150	60
2DPCANet-2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PCANet-2	1.30	1.27	1.12	1.09	1.08	1.13	1.14	1.09	1.06	1.03

5 Conclusions

In this paper, the structure of the cascaded 2DPCA ,the binary hashing, and block histograms has been employed to express the features of the aurora images, and the 3-type classification operation of the aurora image has been executed in terms of the features inputted into the SVM classifier. From the experimental results, the proposed deep learning model, 2DPCANet, has increased the classification accuracy of the aurora image and reduced the running time. With the increasing number of aurora image, 2DPCANet can contribute to the research on aurora.

Although the 2DPCANet processing of the aurora image increases the classification accuracy, the 2DPCA just employs part structure information of the aurora image, which must be solved in the future works. In addition, our future work also includes applying our method to other more databases.

References

1. Akasofu, S.I.: The development of the auroral substorm. Planetary and Space Science **12**(4), 273–282 (1964). doi:10.1016/0032-0633(64)90151-5
2. Hu, H.Q., Liu, R.Y., Wang, J.F., Yang, H.G.: Statistic characteristics of the aurora observed at Zhongshan Stantion. Antarctica. Chinese Journal of Polar Research **11**(1), 8–18 (1999). (in Chinese with English abstract)
3. Hu, Z.J., Yang, H., Huang, D., Araki, T., Sato, N., Taguchi, M., Seran, E., Hu, H., Liu, R., Zhang, B., Han, D., Chen, Z., Zhang, Q., Liang, J., Liu, S.: Synoptic distribution of dayside aurora: Multiple-Wavelength all-sky observation at Yellow River Station in Ny-Ålesund, Svalbard. Journal of Atmospheric and Solar-Terrestrial Physics **71**(8–9), 794–804 (2009). doi:10.1016/j.jastp.2009.02.010
4. Syrjäsuo, M.T., Donovan, E.F.: Diurnal auroral occurrence statistics obtained via machine vision. Annales Geophysicae **22**(4), 1103–1113 (2004). doi:10.5194/angeo-22-1103-2004
5. Wang, Q., Liang, J.M., Gao, X.B., Yang, H.G., Hu, H.Q., Hu, Z.J.: Representation feature based aurora image classification method research. In: Proc. of the 12th National Solar-Terrestrial Space Physics Academic Conf., vol. 71 (2007)
6. Gao, L.J.: Dayside aurora classification based on gabor wavelet transformation. MS. Thesis. Xi'an: Xidian University (2009)
7. Fu, R., Li, J., Gao, X.B., Jian, Y.J.: Automatic aurora images classification algorithm based on separated texture. In: Proc. of the 2009 IEEE Int'l Conf. on Robotics and Biomimetics (ROBIO), pp. 1331−1335. IEEE (2009). doi:10.1109/ROBIO.2009.5420722
8. Wang, Y.R., Gao, X.B., Fu, R., Jian, Y.J.: Dayside corona aurora classification based on X-gray level aura matrices. In: Proc. of the ACM Int'l Conf. on Image and Video Retrieval, pp. 282−287. ACM (2010). doi:10.1145/1816041.1816083
9. Wang, Y.R.: Dayside aurora classification based on X-gray level aura matrices. MS. Thesis. Xi'an: Xidian University (2011) (in Chinese with English abstract)
10. Han, B., Qiu, W.L.: Aurora images classification via features salient coding. Journal of Xidian University **40**(6), 1001–2400 (2013). doi:10.3969/j.issn.1001-2400.2013.06.030
11. Han, B., Yang, C., Gao, X.B.: Aurora image classification based on LDA combining with saliency information. Ruan Jian Xue Bao Journal of Software **24**(11), 2758–2766 (2013). (in chinese) http://www.jos.org.cn/1000-9825/4481.html

12. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. In: ICML (2013)
13. Kavukcuoglu, K., Sermanet, P., Boureau, Y., Gregor, K., Mathieu, M., LeCun, Y.: Learning convolutional feature hierarchies for visual recognition. In: NIPS (2010)
14. Jolliffe, I.T.: Principal component analysis, 2nd edn. Springer, New York (2002)
15. Chan, T.-H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y.: PCANet: A Simple Deep Learning Baseline for Image Classification? Posted on arXiv.org and submitted to TIP, August 2014
16. Yang, J., Zhang, D., Frangi, A.F., Yang, J.: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans. Patt. Anal. Mach. Intell. **26**(1), 131–137 (2004)

An Accelerated Two-Step Iteration Hybrid-norm Algorithm for Image Restoration

Yong Wang[1(✉)], Wenjuan Xu[1], Xiaoyu Yang[1], Qianqian Qiao[1], Zheng Jia[1], and Quanxue Gao[2]

[1] School of Electronic Engineering, Xidian University, Xi'an 710071, China
ywangphd@xidian.edu.cn
[2] School of Telecommunication Engineering, Xidian University, Xi'an 710071, China

Abstract. Linear inverse problem is an important solution frame to solve image restoration. This paper develops an accelerated two-step iteration hybrid-norm reconstruction algorithm, exhibiting much faster convergence rate and better image than iteration shrinkage/thresholding based L1 norm algorithm. In the proposed method, hybrid norm model is built for image restoration objective function. Two-step iteration accelerates objective minimization optimization. Two-step iteration hybrid-norm algorithm converges to a minimizer of hybrid-norm objective function, for a given range of values of its parameters. Numerical examples are presented to validate that the effectiveness of the proposed algorithm is experimentally confirmed on problems of restoration with missing samples.

Keywords: Hybrid-norm · Image restoration · Compressive sensing · Two-step iteration

1 Introduction

Image restoration is still an important image processing research field and has played an important role in medical and astronomical imaging, image and video coding, remote sensing, radar imaging and many other applications [1-3].

Image restoration is to recover an image from distortions to its original image. These distortions usually are twisting, noising, blurring and so on. Image restoration can be described as an inverse problem [2,4]. Let $x \in \mathbb{R}^{n^2}$ be an original $n \times n$ image, $A \in \mathbb{R}^{m \times n}$ be an operator, and $y \in \mathbb{R}^m$ be an observation which satisfies this relationship:

$$y = \mathbb{N}\left(Ax\right) \in \mathbb{R}^m \tag{1}$$

Where $\mathbb{N}(\cdot)$ is an operation process that represents a noise contamination or corruption procedure. In this inverse problem, the goal is to estimate an unknown image

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 335–343, 2015.
DOI: 10.1007/978-3-662-48570-5_33

for observation. In detail, given A, image restoration is a procedure that extract x from y, which is either under-determined or ill-conditioned problem. When A is a linear operator, it is called a linear inverse problem (LIP). Approaches to LIP define a solution $\hat{x}$ (a restored image) as a minimizer of an objective function f. Given by

$$f = \min_{x}\left\{\Re_{reg}(x) + \mu\Re_{fid}(Ax - y)\right\} \tag{2}$$

Where $\Re_{reg}(\cdot)$ promotes solution regularity such as sparseness, $\Re_{fid}(\cdot)$ fits the observed data by penalizing the difference between Ax and y. μ balances the two terms to minimization.

For regularization term $\Re_{reg}(\cdot)$, sparseness is an important measurement which used in image restoration. According to sparseness definition, $\Re_{reg}(\cdot)$ should be L0 norm minimization problem which is NP-hard problem. L0 norm minimization is only the ideally accurate solutions. But it is hard to obtain by solve the L0 concave solutions. Conventional L1-norm minimization is to solve convex optimization that is able to guarantee stable solutions to acquire reconstructions. From this view point of accurate and stable reconstruction, the motivation of the proposed hybrid-norm is to balance both aspects.

The purpose of this paper is to develop a new fast two-step iteration hybrid-norm algorithm (TIH) for restoring x from observation y, where A is a general linear operator. In the section II, hybrid-norm model and two-step iteration solver is introduced which also contains the central theorem of the paper. Finally, experimental results are reported in section III. Conclusions are drawn in the final section.

2 Method

According to our design, restoration procedure consists of several parts shown as Fig. 1. When an original image transmits in transmission channel, it is usually corrupted by some noises or disturbances. Then we use hybrid-norm to build restoration model. In leading to the restoration objective function, two-step iteration solver is employed to solve the hybrid-norm restoration model. When iteration conditions have meet, the final results will be obtained. Hybrid-norm model and two-step iteration method are focused on following sections.

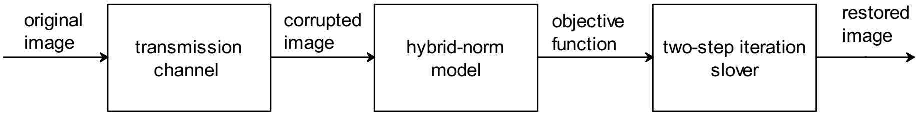


Fig. 1. Flowchart of restoring image based on two-step iteration hybrid-norm (TIH)

According to restoration objective function, a framework for image restoration is hybrid-norm based minimization which is a special case of image restoration where the linear operator is an identity matrix. Denote f as the image restoration objective function:

$$f = \arg\min_{x} \frac{1}{2}\|Ax-y\|_F^2 + \lambda \cdot H(x) \tag{3}$$

where the $H(\cdot)$ regularizer can be homotopic L0 norm which balances between L0 norm and L1 norm describing in hybrid-norm model.

2.1 Hybrid-norm Model

Given x is a sparse and measurement matrix is A, and then the restoration problem can be given by

$$\min_{x} H(x) \qquad s.t. \quad Ax = y \tag{4}$$

where $H(\cdot)$ is hybrid-norm that is transformed to unstrained equation as formula (3).

Hybrid-norm model is defined by

$$H(x) = \sum_{\Omega} g(x) \tag{5}$$

$$g(u) = \begin{cases} a|u|/\tau, & |u| < \tau \\ \dfrac{\left||u|-b\right|}{\left||u|-b\right|+\varepsilon}, & |u| \geq \tau \end{cases} \tag{6}$$

where $H(\cdot)$ means hybrid-norm operator of the proposed method. Constants $a = \dfrac{\sqrt{\varepsilon^2+4\tau\varepsilon}-\varepsilon}{\sqrt{\varepsilon^2+4\tau\varepsilon}+\varepsilon}$ and $b = \tau - \dfrac{\sqrt{\varepsilon^2+4\tau\varepsilon}-\varepsilon}{2}$ are chosen to make the function continuous and differentiable at $|u| = \tau$. Parameter τ is a threshold and $0 < \tau < 1$ is introduced to provide stability. Functional g is related to parameter τ . $\varepsilon > 0$ is to avoid problems due to non-differentiability of hybrid-norm function around intersection point. Profile of hybrid-norm function is shown in Fig. 2. Meanwhile, to be convenient for comparison and understanding of profile functions, profiles of L1, L0, Lp(0<p<1) are added to Fig. 2.

It can be shown from Fig. 2, for any fixed value of ε and τ, hybrid-norm function curve consists of two sections. The first section in the small absolute u is straight with bigger slope than L1. The second section is a conic that is close to L0 under the control of ε. Intersection between two sections keeps smooth and differential when building variables a and b. In the section of L1, this function is strictly convex over $\mathbb{R}^+$. A unique and exact solution to the sparse reconstruction can be acquired. In the section of approaching to L0, solution is the sparsest reconstruction. Hybrid-norm metric combines the merits of L0 and L1 and keeps stable and accurate in reconstruction. In addition, the curve of hybrid-norm is smooth and continuous and its difference is existent and convergent.

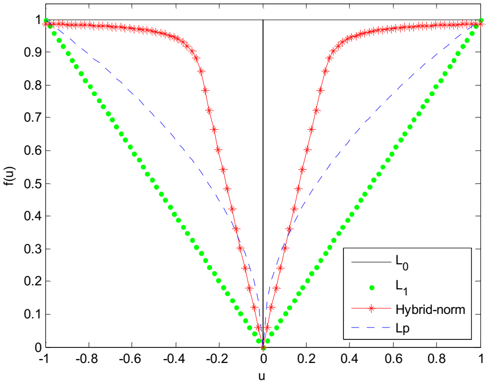


Fig. 2. Hybrid-norm function curve profile comparison with L0, L1 and Lp (0<p<1)

Furthermore, it is seen that the proposed hybrid-norm function includes L1 norm function as a special case when $\tau = 1$. As τ approaches zero, hybrid-norm becomes the L0 of signal. For any $0 < \tau < 1$, hybrid-norm mixes characteristic of both L1 and L0. The value of τ controls the contributions from L1 or L0 respectively. For large τ, hybrid-norm function is closer to a convex function and thus has better convergence to the global minimum. For small τ, it can acquire more accurately solutions because of the profile approaching to L0 norm. Therefore, an optimal τ would best compromise between these two cases.

2.2 Two-Step Iteration Solver

Two-step iteration solver is a fast and effective solver to solve linear inverse problem which developed in fundamental of iterative shrinkage/thresholding (IST)[5,6]. It has been recently used to handle high-dimensional convex optimization problems arising in image inverse problem. In the (k+1)-th iteration, the Two-step iterative solver is as follows.

$$\begin{cases} x_1 = \Gamma_\lambda(x_0) \\ x_{k+1} = (1-\alpha)x_{k-1} + (\alpha - \beta)x_k + \beta\Gamma_\lambda(x_k) \\ \Gamma_\lambda(x) = \Psi_\lambda\{x - A^*(A(x) - y)\} \end{cases} \tag{7}$$

where Ψ_λ is a denoising operator such as wavelet transformation. A* is an adjoint operator of A. α and β are two parameters. The convergence of the two-step iteration algorithm has been well established in [7,8]. Some details also can be found in ref [7,8]. From formula (7), hybrid-norm based restoration problem for equation (3) should be solved for each iteration of two iteration method. In real applications, this subproblem can be solved only approximately, resulting in non-monotonic decrease of the objective function value.

3 Experimental Results

This section reports some experiments to validate image restoration quality and the convergence speeds of the proposed two-step iteration hybrid-norm algorithm (TIH).

We conduct extensive experiments in some examples. Due to the limitation of writing space, we only show two groups of tests in this paper. The goals of these experiments are to present restoration effectiveness from missing samples. The observed images are obtained by convolving the well-known "phantom"and"cameraman" images with a 9*9 uniform blur and then adding noise with variance 40dB below that of the blurred image. The evolution of the objective function and convergence performance are shown using iterative shrinkage/thresholding (IST) and the proposed TIH method in the results.

Example 1. In this group, test object is phantom that comes from typical medical test image. Table 1 lists its results in mean square error (MSE)and CPU time. Figure 1 shows the observed image and the restored image produced by IST and TIH. Figure 2 shows convergence processing of IST and TIH.

Table 1. Experimental results for phantom

	IST	Proposed TIH
MSE	0.17306	0.026431
CPU time	35.537028	33.633816

Quantitative index MSE and CPU time are shown in Tab1. In this table, IST and TIH take 0.17306 and 0.026431 of MSE. TIH improves image quality approximate one power of magnitude from 0.17306 to 0.026431. The proposed method is super to IST. In consuming time, TIH consumes 33.633816s and IST has 35.537028s. The proposed method improves little faster than IST. The reason is that the phantom is an ideal sparse image which we can not dig more sparse information. These factors decide iteration times.

In Fig. 3, subimage (a) is original image. (b) is corrupted image by noisy and blurred factor. (c) stands for the restored image using the proposed algorithm. (d) is the restored image using IST method. Restored image using TIH reduces noisy and blurred factor, which gets sharp boundaries and clear contents in several important part such as gray circle, two black ellipses. The white circle boundary is sharp and clear. In subfigure (d), IST method restores image which has many pseudo artifacts like dummy circle. White circle boundary of subfigure (d) is little blurred. In visual effectiveness of restored images, TIH is clearer and neater than IST algorithm.

Fig. 4 has two subfigures. The above is a curve representing the relationship between objective function and CPU time. The blow curve represents relationship between restored error MSE and CPU time. TIH method converges very faster than IST and obtains lower restored error MSE. IST spends 35.537028 seconds to gets 0.17306 error and TIH needs 33.633816 seconds to have 0.026431 restored errors. Fig. 4 shows that TIH converges considerably faster and more excellent than IST.

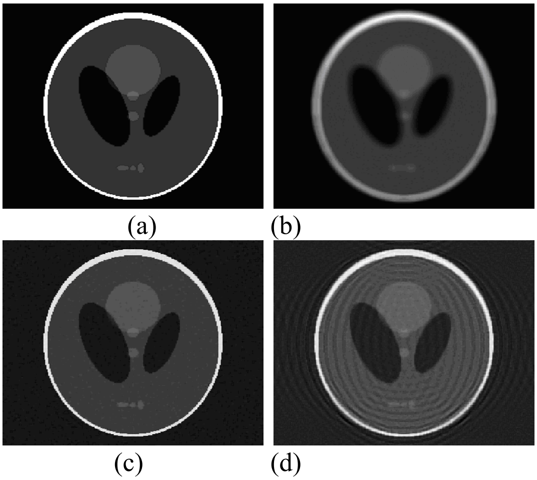


Fig. 3. Image restoration results for phantom. (a)Original image ;(b)Noisy and blurred image; (c)TIH restored image; (d)IST restored image

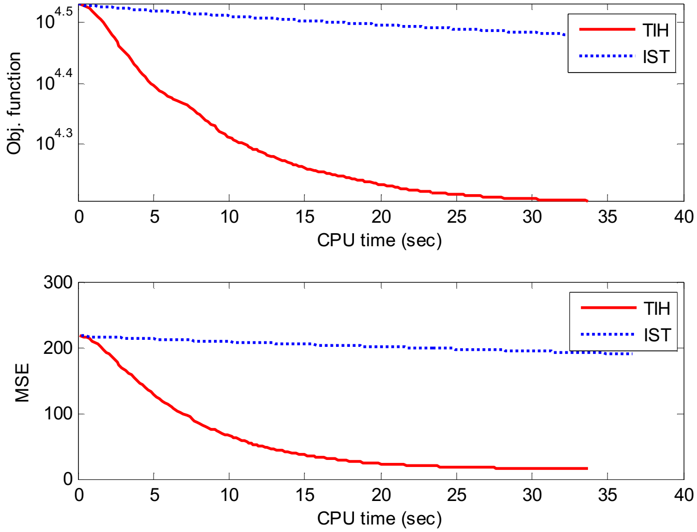


Fig. 4. Convergence behavior for phantom restoration. Above: objective; below: relative error MSE. In both plots, the horizontal axes denote CPU time in seconds.

Example 2. In the second experiment, we apply "cameraman" image to test effectiveness of the algorithm. In this group experiment, noisy and blurred image is obtained using the same method in example 1. "cameraman" image is a real natural image which is not completely different from phantom image in 1st experiment.

Table 2. Data results for cameraman

	IST	TIH
MSE	0.067087	0.027247
CPU time	39.078251	16.224104

According to Table 2, TIH obtains 0.027247 in MSE and 16.224104 seconds in CPU time. IST has 0.067087 MSE and 39.078251 seconds in CPU time. In restoration quality, TIH has 0.027247 errors of original image and restored image. IST only has 0.067087 differences between the original image and restored image. Restored image of TIH considerably is better than that of IST. Also, TIH is largely faster than IST.

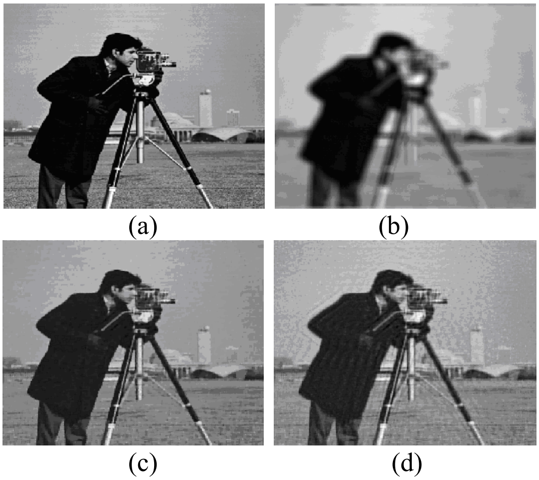

(a) (b) (c) (d)

Fig. 5. Image restoration results for phantom. (a) Original image; (b) Noisy and blurred image; (c) TIH restored image; (d) IST restored image

Some results of "cameraman" restored image are shown in the Fig. 5. In the same statements as in first group experiment. Subfigures (a) to (d) are original image, noisy and blurred image, TIH restored image and IST restored image separately. Differences of restored images in subfigure (c) and subfigure (d) are distinguished apparently. From the view of vision, subfigure (c) is clearer and neater than subfigure (d). Image in subfigure (d) has large amount of artifacts and alias. Restored image in subfigure (c) has little drawbacks. But it can be seen clearly. Why we can not restore an image as same as original image. Restoration is an anti-process that can not completely restore image as original image in the condition of loss of some information. From data of experiments, MSE in subfigure (c) is 0.027247 and subfigure (d) only takes 0.067087. The proposed algorithm improves apparently both in vision and experimental data.

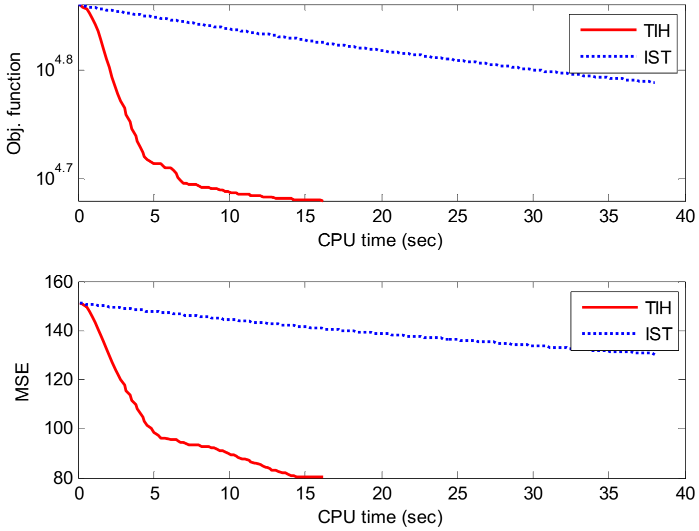


Fig. 6. Convergence curves for "cameraman" image restoration

In the figure 6, convergence curves of "cameraman" image restoration are shown. Relationship between objective function and CPU time is shown in the above subplot in Fig. 6 and restored image error is shown in the below subplot in Fig. 6. IST needs 39.078251 seconds to up to MSE value 0.067087 and TIH only requires 16.224104 seconds to obtain 0.027247 restored images. The two curves using TIH decrease sharper than that of IST. In other words, TIH converges rapidly and consumes little times.

Though different object images are employed in the two groups of experiments, nearly same conclusions are drawn that the proposed TIH method is superior to IST in both image quality and restoring speed.

4 Conclusion

This paper proposed a fast two step iteration hybrid-norm image restoration method to solve fast and high image restoration. The proposed method combined fastness of two-step iteration and effectiveness of hybrid-norm model which is a homotopical L0 norm method. Two groups of experiments in phantom and natural images give evidences of high image quality and fast speed in restoring image.

Acknowledgments. The main work in this paper is supported by "the Fundamental Research Funds for the Central Universities"(JB150218).

References

1. Andrews, H., Hunt, B.: Digital Image Restoration. Prentice-Hall, Englewood Cliffs (1977)
2. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. Bristol, UK (1998)
3. Katsaggelos, A.: Digital Image Restoration. Springer, New York (2012)
4. Archer, G., Titterington, D.: On Bayesian/regularization Methods for Image Restoration. IEEE Trans. Image Process. **4**(3), 989–995 (1995)
5. Daubechies, I., Defriese, M., De Mol, C.: An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
6. Beck, A., Teboulle, M.: A Fast Iterative Shrinkage-thresholding Algorithm for Linear Inverse Problems. SIAM J. Imaging Sciences **2**(1), 183–202 (2009)
7. Bioucas-Das, J., Figueiredo, M.: A New Twist: Two Step Iterative Shrinkage/thresholding Algorithms for Image Restoration. IEEE Trans. Image Process. **16**(12), 2992–3004 (2007)
8. Bioucas-Dias, J., Figueiredo, M.: Two-step Algorithms for Linear Inverse Problems with Non-quadratic Regularization. In: IEEE Int. Conf. Image Process., San Antonio, TX (2007)
9. Wang, Y., Liang, D., Chang, Y., Ying, L.: A Hybrid Total-Variation Minimization Approach to Compressed Sensing. IEEE International Symposium on Biomedical Imaging, Barcelona, Spain, pp. 74–77 (2012)
10. Chen, X., Michael Ng, K., Zhang, C.: Non-Lipschitz Lp-regularization and Box Constrained Model for Image Restoration. IEEE Trans. on Imaging Process. **21**(12), 4709–4720 (2012)
11. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally Centralized Sparse Representation for Image Restoration. IEEE Trans. on Imaging Process. **22**(4), 1620–1630 (2013)
12. Liang, D., Ying, L.: A Hybrid L0-L1 Minimization Algorithm for Compressed Sensing MRI. In: Proceedings of International Society of Magnetic Resonance in Medicine Scientific Meeting (2010)

SiftKeyPre: A Vehicle Recognition Method Based on SIFT Key-Points Preference in Car-Face Image

Chang-You Zhang[1,2(✉)], Xiao-Ya Wang[1,2], Jun Feng[2], and Yu Cheng[3]

[1] Laboratory of Parallel Software and Computional Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
changyou@iscas.ac.cn
[2] School of Infomatics Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
[3] Institute of Applied Mathematics, Hebei Academy of Sciences, Shijiazhuang 050081, China

Abstract. Vehicle recognition from images produced in roads bayonet provides important clues to solve vehicle crime cases. Its accuracy is not enough to meet the requirement in real conditions. We proposed a vehicle recognition method, SiftKeyPre, based on SIFT(Scale-invariant feature transform) key points preference for car-face images. Firstly, SiftKeyPre choices the SIFT key points following the DualMax algorithm to get a DualMax set. Meanwhile, Lowe set is defined as another one following Lowe algorithm. Secondly, we define a DL set under an intersection operation on DualMax set and Lowe set. For positive examples training images, we count the appearance times of each key point of DL set to compute the attention degree of each key point in base image. Finally, matching degree between the base image and a target image is evaluated with the attention degree of each matched points. SiftKeyPre method confirms a testing image based on its matching degree. Experiments results show that, under a given recall constraints, the precision of SiftKeyPre method is better than FLANN and Lowe. SiftKeyPre's computational complexity is closed to that of Lowe. Comparing with other algorithms based on training, SiftKeyPre is of lower training intensity.

Keywords: Car-face image · SIFT key points · Preference · Attention degree · Matching degree · Recognition

1 Introduction

In modern societies, the insecurity and threat events are increasing. Vehicle recognition from high-definition vehicle images produced in roads bayonet is an important source of clues on which public security departments solve cases of vehicle crime relies. For the phenomenon of faking and sheltering vehicle license plate, vehicle recognition system can not recognize car types correctly just based on the license plate. In a medium-sized city, more than 10000 images are captured at each major road bayonet by high-definition cameras per hour. At present, policeman selects out certain type of vehicle, for example black PASSAT, by "eyes of human".

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 344–358, 2015.
DOI: 10.1007/978-3-662-48570-5_34

This retrieval process is non-efficiency and tedious. It is urgency for investigation on automatic recognition algorithms to identify suspect vehicle.

The basic problem in computer vision research is object classification and detection. The image object recognition is an important branch with more than fifty years history [1]. Image object recognition algorithms are divided into two basic categories, algorithms based on low-level feature and deep learning model. In algorithms based on visual feature, low-level features are extracted from images, and then these features obtained from a variety of features extraction algorithms are encoded. Finally, the appropriate classifier is designed to get classification results. Deep learning model [2] is another kind of image object recognition algorithm. Its basic idea is to learn hierarchical feature representation in supervised or unsupervised way. The objects are described from bottom to top.

Classifier design is the key step in object recognition based on low-level features. Classical classifier based on visual feature include FLANN, Lowe (classifier in the SIFT algorithm) etc. FLANN algorithm is based on key points in base image and calculates Euclidean distance one by one with key points of each testing image. It selects minimum distance and takes key-points number exceeded the given threshold as matched points. It determines whether the image object is the same according to the number of matched points. Lowe also calculates Euclidean distances between one key point of base image and all key points of testing image. If the distance is smaller than a given value, Lowe algorithm selects this key-point pair as a high quality match [3]. The object similarity in Lowe is still based on the number of matched key points. Training-based classifiers include neural networks, support vector machines, k-nearest neighbor, random forests, and so on. These algorithms need to manually mark a large number of training samples to improve the classification quality.

Big data brings huge challenges to the traditional learning algorithms. Deep learning model has powerful ability to express data naturally, so it will impact on image object detection methods. However, there are some problems like poor interpretation, high model complexity, optimization difficulty, computing intensity etc [1]. Main deep learning models include automatic coder (Auto encoder) [4], restricted Boltzmann machines (Restricted Boltzmann Machine, RBM) [5], deep belief networks (Deep Belief Nets, DBN) [6], Convolutional neural networks (Convolutional Neural Networks, CNN) [7], bio-heuristic model [8], etc. Deep learning models rely on huge amounts of training samples, and it is of high training intensity.

Classification algorithms based on training require a lot of manual marked samples, or huge amounts of training samples. It is hardly to be used in sample-limited scenarios. Missing match rate is high in classic FLANN and Lowe algorithm when the distance is close between a pair of key-points. We proposed a SiftKeyPre method based on SIFT key-points preference in vehicle image. SiftKeyPre makes a compromise by taking advantages of both the classical linear classifier and training classifiers with high computing intensity. It extracts car-face as region of interest. In SiftKeyPre method, the intersection set of Lowe's preferring key points set and DualMax's preferring key points set is used as the preferring key points set. SiftKeyPre set is used to calculate attention degree of base image. Matching degree between the base image and testing images is evaluated. SiftKeyPre is essentially a linear classifier with low-intensity training algorithms.

The structure of the rest of this paper is organized as follows. Section 2 introduces some related works on image recognition. Principle of SiftKeyPre is described in section 3. Section 4 tests the effect of SiftKeyPre through experiment. Section 5 further analyses the results and algorithm parameter to enhance the practicability of SiftKeyPre. Section 6 summaries our works, and discusses SiftKeyPre's limitations and some further promising researches.

2 Related Works

The variation of angle and distance between the running cars and high-definition camera causes image differences in scale and orientation. Thus, we use SIFT algorithm to extracting features from images to guarantee the invariant of image scale and rotation.

Researchers investigated some auto-algorithms of car detection and vehicle recognition. Wei [9] developed a complexity-aware criterion to balance the separating capacity and retrieval efficiency based on strip feature of car in static images. This algorithm only detected the existence of vehicle, not identified its types. Some approaches classify vehicles into generic classes (vans, SUVs and bus etc.) [10-11]. These methods were not accurate enough for finding crime vehicles. Vehicles need to be classified into specific 'make and model' classes (MMR, Make and Model Recognition) [12].

Some scholars put forward some ideas and algorithms of vehicle recognition at early years. Sun Ze-hang [13] proposed an algorithm of vehicle recognition using Haar Wavelet decomposition for features extraction and Support Vector Machines (SVM) for classification. David Santos [14] introduced a vehicles recognition algorithm relying on the analysis of car external features. These features included shape of the car's rear view and the car back lights. Through comparing with the features in database, system determined whether both images were matched or not. Daniel Marcus Jang [15] demonstrated a recognition application based on the SURF (Speeded Up Robust Features) algorithm. In its vehicle database, the images of each type of vehicle were photographed from 16 views. It cost approximately 16 times computing workload. Wang Quan [16] presented a MDS(multi-dimensional scaling) feature learning framework in which MDS is applied on high-level pairwise image distances to learn fixed-length vector representations of images. Images need to be uniformed caused information loss. Ferencz [17] built a classification cascade for visual recognition from one example and proposed an approach for vehicle recognition. The main contribution of this work is a classification cascade built by arranging information-rich hyper-features extracted from a single vehicle exemplar image. For running vehicles, images of exact front and lateral views are very rare.

So for from 2004, WOB (Word of Bag) are the mainstreem algorithm in image recognition. The main idea of these algorithms is clustering the features by employing K-means clustering algorithm to construct the visual vocabulary. These clustering centers are regarded as visual words. Then, they use the histogram described by appearance frequency of the visual words to represent the content of the image. By regarding the visual words histogram of each image as features vector, the classification model was abtained through SVM training [18-20]. In these algorithms, the differences of the categories are obvious. And the training of SVM was based on enough selected samples of images. In the scenario of just 100 car-face images, the effect of SVM training is limited.

3 SiftKeyPre Recognition Method

The vehicle recognition processes of SiftKeyPre consist of five steps. They are designing data structure of key points, constructing key-point pairs, preferring key-point pairs, calculating attention degree of key points, and calculating matching degrees of target images. The algorithms components are illustrated in Fig. 1.

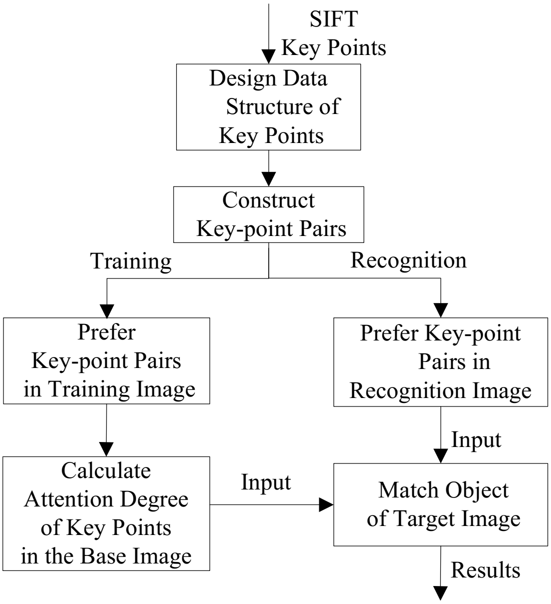

SIFT Key Points
Design Data Structure of Key Points
Construct Key-point Pairs
Training
Recognition
Prefer Key-point Pairs in Training Image
Prefer Key-point Pairs in Recognition Image
Input
Calculate Attention Degree of Key Points in the Base Image
Input
Match Object of Target Image
Results

Fig. 1. Framework of SiftKeyPre method.

We call the template image of a given type of vehicle as *base image*, and the image to be matched as *target image*. In SiftKeyPre algorithms, there is just a single base image and multiple target images. For convenience, we assume that there are *m* SIFT key points in base image, and *n* SIFT key points in any one target image. Each key point is described as a vector with 2 float numbers and 128 integers. Euclidean distance is used to measure the similarity of 2 key points in the same key-point pair.

3.1 Data Structure of Key Points

Data structure of SIFT key points is consist of some parameters such as octave, scale, σ, x, y, and so on. Some of them are no contribution to SIFTKeyPre algorithm. So, we just reserve the pixel position parameter (x, y), and the key points descriptor (128 integers) to construct the data structure of key point. The key point descriptor *kp* is designed as a sequence.

kp = sequence of { x, y, d_0,d_1, …, d_i, …, d_{127} }

Where,
x, y -- is the pixel coordinate of *kp*;
d_i -- i-th component of *kp* ($0 \leq k \leq 127$).

3.2 Construct Key-Point Pairs

The similarities between key points of base image and that of target image is the foundation of image recognization. Let A stand for the key points set of the base image, and B stand for the key points set of a target image. Cartesian product of A and B builds the key-point pairs set C. As mentioned above, there is *n* key points in set A, and *m* key points in set B, then there are n × m key-point pairs in set C.

Let $A = \{A_0, A_1, \ldots, A_i, \ldots, A_{n-1}\}$,
$B = \{ B_0, B_1, \ldots, B_j, \ldots, B_{m-1}\}$.

Then,

A_0 and B produces key-point pairs $<A_0, B_0>$, $<A_0, B_1>$, ……,$<A_0, B_{m-1}>$;
A_1 and B produces key-point pairs $<A_1, B_0>$, $<A_1, B_1>$, ……,$<A_1, B_{m-1}>$
……
A_{n-1} and B produces key-point pairs $<A_{n-1}, B_0>$, $<A_{n-1}, B_1>$, ……,$<A_{n-1}, B_{m-1}>$

3.3 Key-Point Pairs Preference

Distance is the classic measurement method for evaluation. To evaluate the matching quality of a pair of points in a single key-point pair, we construct a distance matrix H with size of n × m based on their distances.

We get distances of two key points $<A_i, B_j>$ as formula (1)

$$dist(A_i,B_j) = \sqrt[2]{\sum_{t=2}^{129}(A_i[t]-B_j[t])^2} \tag{1}$$

Where,
A_i -- the *i*-th key point data in set A, ($0 \leq i < n-1$);
B_j -- the *j*-th key point data in set B, ($0 \leq j < m-1$);
$A_i[t]$ -- the *t*-th element of A_i, ($2 \leq t < 129$);

$B_j[t]$ -- the *t*-th element of B_j, $(2 \le t < 129)$.

These distances are all filled into a matrix H. Row of matrix H corresponds to a certain key point of the base image, and column of matrix H corresponds to that of one target images. It is said that H_{ij} denotes the distance between i-th key point in the base image and j-th key point in a target image.

According to common sense, when the distance of a key-point pair is larger than a special value, we think the both points are not similar. Their similarity is approximately 0. To reduce the computing intensity, we transform distance in H to similarity following the rules that,

(1) Smaller distance mapping to bigger similarity,
(2) If a distance is bigger than a given threshold D, similarity is 0,
(3) Similarity value range is from 0 to 1.

According to the above rules, matrix H is transformed to matching quality matrix EV according to formula (2).

$$EV_{ij} = \begin{cases} 0, & H_{ij} \ge D \\ \dfrac{D - H_{ij}}{D}, & H_{ij} < D \end{cases} \qquad \textbf{(2)}$$

Where,
EV_{ij} – similarities in matching quality matrix;
H_{ij} – distances in H;
D – given threshold by experiments on image samples.

From formula (2), there are some zero elements in EV. These elements are no chance to be choose as matched key-point pairs. It is said that the corresponding key points of target image are out of matching.

After above pre-processing, the vital step in SiftKeyPre is to prefer real matched key-point pairs by Dual-direction evaluation. We call this matching selection as DualMax optimization. DualMax follows these steps.

```
(1) Let i=0;
(2)Traverse the i-th row of EV, select the maximum element,
        $ema_{ij} == \max\{ e \mid e_{ij}, 0<=j<=m \}$;
        mark j-th column.
(3) Traverse the j-th column, if $ema_{ij}$ is the maximum value,
        $ema_{ij} == \max\{ e \mid e_{ij}, 0<=i<=n \}$;
        mark the corresponding key-point pair as a DualMax matching, and put it
    into set Q.
(4) Change all values of i-th row and values of j-th column to zero.
(5) Else i++; goto (2);
(6) If i >= n, finished.
```

The elements in set Q are preferred key-point pairs.

3.4 Attention Degree of Key Points in the Base Image

Human vision system often focuses visual attention on some special objects of the scene when processing a relative complex scene. It processes these special objects in priority so as to get main information of the scene in minimum time cost [21]. For different aims, these special objects are different. More deeply, there must be a few attentional points to represent the object.

According to this character of human vision system, we speculate that the SIFT key points extracted from car-face are of different attention degrees. To investigate into this, we give each SIFT key point an attention degree by means of a statistic on being preferred times based on a set of positive example images.

Let set L denote key-points set preferred by Lowe matching algorithm. As mentioned above, Q denotes key-points set preferred by DualMax algorithm. To further enhance the preference quality, we build set LQ as an intersection set of L and Q. All the key points in set LQ is deemed as being preferred once. We repeat this operation on a positive sample image set to get the preferred-times of each key-point.

We assume that there are S samples in a given training set. As mentioned above, A is the key point set of base image, and there are n key points in set A. Let AN denote the times of key-point preferred. The attention degrees are calculated in accordance with the following steps.

```
(1) i = 0; i < $S$;
(2) $LQ_i = L_i \cap Q$;
(3) for each element $kp \in LQ_i$, if kp == A[i], AN[i]++;
(4) i++, goto (1)
(5) output AN, finish.
```

Finally, we normalize AN according to the formula (3).

$$AD[i] = \frac{AN[i]}{\sum_{0}^{n-1} AN[i]} \qquad (3)$$

Where, AD is a vector of the attention degrees of key points in base image.

3.5 Matching Degree of Target Image

Matching degree is a comprehensive evaluation which compounds matched key points number and their attention degrees. If the matching degree of target is bigger than the given threshold, this target image is matched with the base image.

Let vector AQ denotes the sequence of flag for key points in the base image with n components. AQ is initiated with zero. Vector AV denotes the sequence of key points in base image, and BV denotes the sequence of key points in target image. As mentioned in section 3.4, set Q denotes the preferred key points in target image. We get the matching degree in the following steps.

```
(1) i = 0;
(2)if key-point pairs <AV[i],BV[j]>$\in$Q, AQ[i] = 1;
(3) i++;
(4) if i < n, goto (2);
(5) matching degree $v = AQ \bullet AD$ ;
(6) if v >=V, target image is matched, finish.
```

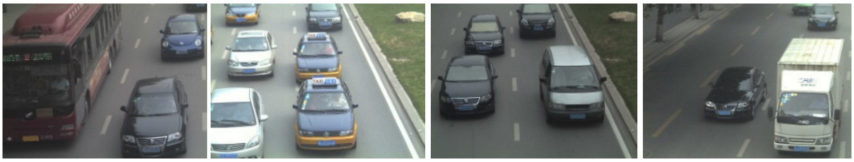
Matching threshold V depends on the need of specific application.

4 Experiments

4.1 Experimental Setup

Platform: CPU-Phenom II 960T 3.0GHz* quad-core; RAM-DDR3 3.25G; Ubantu 12.04 OS; openCV library.

Data: The testing images are HD images produced at real road intersections of a city in China. There are total 1000 images, where 100 positive samples (BLACK PASSAT). Typical original images are illustrated in Fig. 2. These images are created in various angles and different distance.



* license numbers are blurred for privacy protection

Fig. 2. Examples of original images.

4.2 Experimental Results

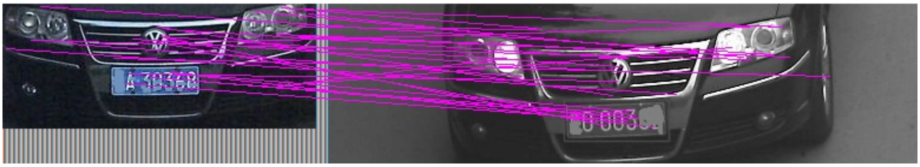
To investigate the precision and performance of our SiftKeyPre algorithm, we compare both indices among the three typical algorithms (FLANN, Lowe and SiftKeyPre) at a given recall.

In this experiment, ROI is car face extracted from original images. SiftKeyPre selects key points of high quality to determine whether the car in a target image of the same type as that in the base image or not. One of the target images' matched key-point pairs are illustrated in Fig. 3.

In Fig. 3, the base image is on the left, and the target image is on the right. Each matched key-point pair is illustrated with a line.

(1) Precision

Effectiveness of SiftKeyPre algorithm is evaluated with two indicators: precision and recall. The precision rate has negative relation with recall. It is said that the improvement



(base image) (target image)

Fig. 3. The matched key-point pairs in SiftKeyPre algorithm.

of precision followed with a drop of recall. We compared the three algorithms (FLANN, Lowe and SiftKeyPre) in their precision and recall. The results are shown in figure 4.

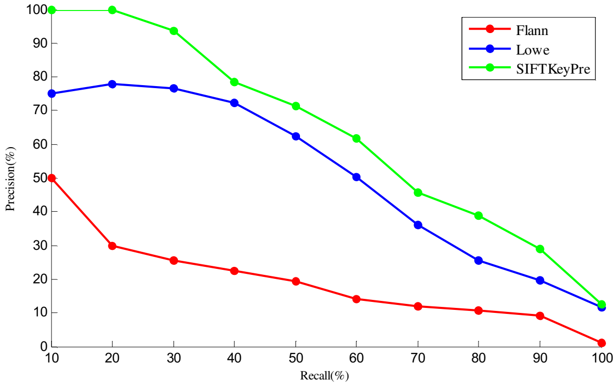


Fig. 4. Comparison between SiftKeyPre/FLANN/Low algorithms.

In Fig. 4, the abscissa denotes recall indicator, and ordinate denotes precision indicator. The experiments test a range of recall from 10% to 100% and the corresponding precision. As shown in Fig. 4, the precision of SiftKeyPre is significantly higher than that of FLANN and Lowe at a given recall. For instance, at the point of recall = 90%, the precision of SiftKeyPre is 27.95%, that of Lowe is 19.65%, and that of FLANN is 9.29%.

Precision are different between SiftKeyPre and the other two algorithms at a various recall from 10% to 100%. These differences are listed in Table 1.

Table 1. Precision differences with FLANN and Lowe algorithms.

differences with	Max	Min	Average
Lowe	+25.00%	+1.04%	+12.46%
FLANN	+35.86%	+1.29%	+16.69%

As shown in Table 1, compared with Lowe algorithm, SIFTKeyPre achieved a maximum +25% improvement in retrieval accuracy. Meanwhile, compared with FLANN algorithm, a maximum 35.86% improvement was obtained. Obviously, SiftKeyPre performs better than the other two algorithms.

(2) Performance

SiftKeyPre algorithm is consist of two key processes (training and recognition). In training process, we get algorithm parameters such as D, attention degree et al. Training process needs only once in advance. Training with a sample image costs 0.177961s in average. For a given practical application, users seemingly unconcerned about the time cost on training.

Users more concerned about the response efficiency of recognition process. We investigated into the response time for a single target image. The results comparing with FLANN and Lowe are listed in Table 2.

Table 2. Comparison of response time with FLANN and Lowe.

Algorithm	FLANN	Lowe	SIFTKeyPre
average response time (s)	0.122930	0.081694	0.083226

From table 2, SiftKeyPre saves 32.30% than FLANN in response time. SiftKeyPre costs a little longer time than Lowe. Even so, it is well worth to exchange a performance loss of 1.88% for a precision improvement of 35.86%.

5 Analysis and Discussions

In this section, we analyze parameters of SiftKeyPre and discuss the training intensity. To be sure that the parameters of FLANN and Lowe are adjusted carefully to achieve their best precision on testing images.

5.1 DualMax Threshold

In formula (2), D is a key parameter in DualMax preferring process. In fact, D is a critical value to determine whether a distance of key-point pair maps to zero or not. A bigger D means higher quality of key-point pairs. And there are much more 0 in matrix EV. There will be less key-point pairs in DualMax set Q. Meanwhile, this will loss more key point information which contribute to the final recognition.

To balance this compromise, we develop 2 principles. (1) Gold section number is graceful to be used as the dividing line between the zero similarity and non-zero similarities; (2) For the same target image, the number of key-point pairs in Q and LQ should be roughly equal. Accordance with both principles, we determine D in the following steps. As mentioned above, matrix H has n rows and m columns. And the gold section number is 0.618.

(1) Let i =0;
(2) For i-th row, if $H_{ij} = \min\{ H_{ij} \mid H_{ij}<H_{i*}, 0<j<m \}$, $V[i] =H_{ij}$; i++;
(3) if i<n, goto (1);
(4) $R = round\ (n \times (1 - 0.618\ ))$
(5) choose the R-th bigger number in V, let D_k=V[R], (0<=k<K)
(6) For each image in training set of K images, D is valued as average of Dks.

$$D = (\sum_{k=1}^{K} D_k) / K \tag{4}$$

5.2 Training Intensity

SiftKeyPre is a linear classifier with low-intensity training. This training is the important reason for the improvement of precision. We define the training intensity as the minimum number of training samples when the vital parameters of SiftKeyPre are convergent. To investigate into the convergence, we do three experiments from various views. (1) Overview all attention degrees of key points; (2) the transferring curves of typical key points with significant value changing; (3) the impact on recognition precision.

The aim of training is to get the attention degree of each key point in the base image. We do experiments with samples of 10, 20, 30, ..., 90 and 100 and draw the corresponding attention degree values together in the same coordinate system. The inflection point of these curves are the alternative training intensity. The attention degree changing curves are illustrated in Figure 5.

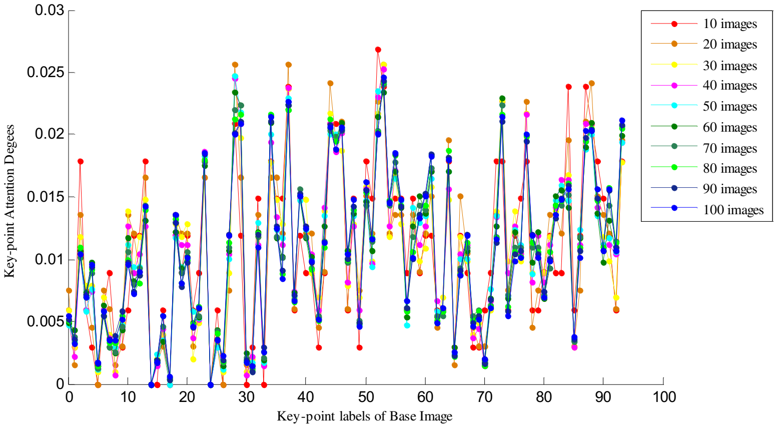


Fig. 5. Attention degree changing curves of all key points.

In figure 5, abscissa is the label of key points in base image, and the ordinate is the attention degree of these key points. As the figure legend, the training times from10 to 100 mapping to the colors from red to blue. The majority key points attention degree are converged to a stable value illustrated in a "cooler" color.

Investigating into figure 5, we find some key points (such as id=47, 52, 65 and 91) whose attention degree value fluctuates more dramatically. To show the trend more clearly, we select these 4 key points and draw the changing path in an unfolded view, as shown in figure 6.

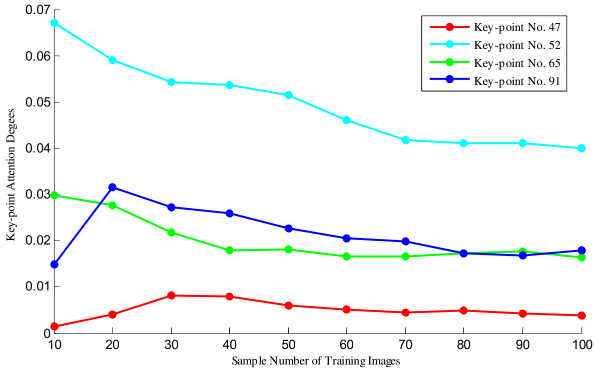


Fig. 6. Changing path of selected attention degrees in unfolded view.

In figure 6, the changing trends show that the attention degree value of key points will be stable under a certain number of training samples. This number range from 50 to 70. Then, a new question is coming. Is there any significant influence on the final precision under the training intensity of 50 and 70?

We developed another experiment on the training intensity of 0, 50, and 100. The changing trend of precision with incremental recall is illustrated in Figure 7.

From figure 7, we find that the precision under no any training is much lower than that under 50 samples' training. It is said that training process improved the precision of SiftKeyPre. Meanwhile, when the training intensity enhanced from 50 or 100 samples, the both precisions become no obvious difference. It is said that 50 is a critical point of training intensity from the view of precision effect. The attention degrees of key points reach to convergent values.

5.3 Attention Degree

To view the attention degree of key points more clearly, we draw these points on the base image with various radius and colors according to its pixel position of (x, y) and attention degree.. The bigger radius denotes bigger attention degree. Their colors range from blue to red, mapping from the smallest to biggest attention degrees. These 2 pictures in figure 8 illustrate the changing trace of attention degrees under training intensity of 0, 50.

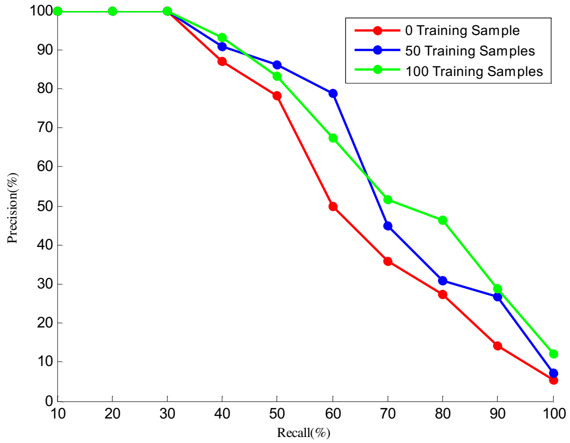
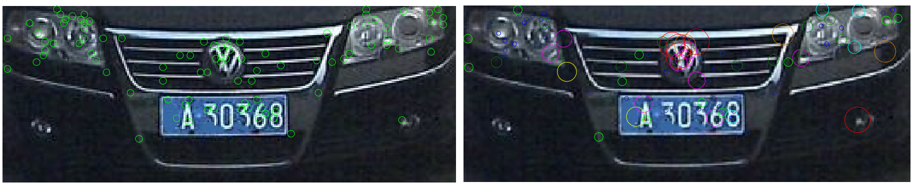


Fig. 7. Trend of precisions with incremental recall



(a)under no training

(b)under training intensity of 50

Fig. 8. View of attention degrees of key points.

In the intuition of human perception system, part of key points is of significant contribution in the recognition decision. These key points are more important than others. From figure 8, we found that the most important point concentrate on the region of car logo, car light, and some distinct texture. In figure 8(a), all key points are without training, so the importance are almost in average. In figure 8(b), key points are trained under 50 positive samples. The most important points become significant in size. Training more than 50 samples do not contribute significantly to the size of key points. That proves again that the attention degrees convergence to a stable value.

Some key points with significant attention degree locate on license plate. In fact, this is a wrong matching because license plate is not a inherent part of a car. These points should be removed from the key points set.

6 Conclusion and Future Works

With the analysis on the property of vehicle images on road, we proposed a vehicle recognition algorithm SiftKeyPre based on low-level feature extraction in SIFT algorithm. SiftKeyPre consists of five steps: Design data structure of SIFT key points, construct key-point pairs, prefer key-point pairs, calculate attention degree of key points in base image, and match object of target image. Under the given recall rate, SiftKeyPre achieved obvious improvement of precision comparing with both FLANN and Lowe. As for time-consumption, SiftKeyPre algorithm cost less computing time than that of FLANN in 32.30% and almost equal to that of Lowe.

There are spaces to improve this method. For example, finer pre-processing is helpful to higher precision. Combined SiftKeyPre with support vector machines (SVM) based on WOB, neural networks, and deep learning algorithms will be a promising field in vehicle recognition system. With the huge amount of images streaming into the system, high performance algorithms are the future direction in image recognition and retrieval systems.

Acknowledgement. This paper is supported by the Natural Science Foundation of China (61379048), Hebei Province Natural Science Foundation (F2013210109). Special Project of National CAS Union – The High Performance Cloud Service Platform for Enterprise Creative Computing, Special Project of National CAS Union – Car Image Retrieval on Many core platforms, Special Project of National CAS Union (Hebei, No.14010015)–The Performance Optimization of Key Algorithms in Senile Dementia Data Processing. The authors also would like to express appreciation to the anonymous reviewers for their helpful comments on improving the paper.

References

1. Huang, K.-Q., Ren, W.-Q., Tan, T.-N.: A Review on Image Object Classification and Detection. Chinese Journal of Computers **37**(6), 1225–1240 (2014)
2. Liu, J.-W., Liu, Y., Luo, X.-L.: Research and development on deep learning. Application Research of Computers **31**(7), 1921–1942 (2014)
3. Lowe, G.D.: Distinctive Image Features from Scale-Invariant Key points. International Journal of Computer Vision **60**(91–110), 1–28 (2004)
4. Bourlard, H., Kamp, Y.: Auto-association by multilayer perceptions and singular value decomposition. Biological, Cybernetics **59**, 291–294 (1988)
5. Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. In: Processing of the Processing of the Parallel Distributed: Explorations in the Microstructure of Cognition, Foundations, vol. 1, Chapter 6. MIT Press (1986)
6. Hinton, G.E., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation **18**(7), 1527–1554 (2006)
7. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

8. Huang, Y.-z., Huang, K.-q., Tao, D.-c., et al.: Enhanced biologically inspired model for object recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **41**(6), 1668–1680 (2011)
9. Zheng, W., Liang, L.: Fast Car Detection Using Image Strip Features, pp. 2703–2710. IEEE (2009)
10. Chen, Z., Pears, N.E., Freeman, M., Austin. J.: Road vehicle classification using support vector machines. In: IEEE Int. Conf. Intelligent Computing and Intelligent Systems, pp. 214–218 (2009)
11. Hua, L., Xu, W., Wang, T., Ma, R., Xu, B.: Vehicle Recognition Using Improved SIFT and Multi-View Model. Journal of Xi'an Jiaotong University **47**(4), 92–99 (2013)
12. Pearce, G., Pears, N.: Automatic make and model recognition from frontal images of cars. In: 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, pp. 373–378 (2011)
13. Sun, Z., Miller, R., Bebis, G., et-al.: A real-time precrash vehicle detection system. In: Sixth IEEE Workshop on Applications of Computer Vision, pp. 171–182, December 3–4, Orlando, Florida (2002)
14. Santos, D., Correia, P.L.: Car recognition based on back lights and rear view features. In: 2009 10th Workshop on Image Analysis for Multimedia Interactive Services, pp. 137–140, May 6–8, London, United Kingdom (2009)
15. Jang, D.M., Turk, M.: Car-rec: a real time car recognition system. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 599–605, January 5–7, Kona, HI, USA (2011)
16. Quan, W., Boyer, K.L.: Feature learning by multidimensional scaling and its applications in object recognition. In: 2013 XXVI Conference on Graphics, Patterns and Images, pp. 8–15, August 5–8, Arequipa, Peru (2013)
17. Ferencz, A., Learned-Miller, E.G., Malik, J.: Building a classification cascade for visual identification from one example. In: ICCV 2005, Wash., DC, pp. 286–293 (2005)
18. Gao, H., Dou, L., Chen, W., Sun, J.: Image classification with bag-of-words model based on improved SIFT algorithm. In: 2013 9th Asian Control Conference. ASCC 2013, pp. 1–6, June 23–26, Istanbul, Turkey (2013)
19. Bai, S.: Sparse code LBP and SIFT features together for scene categorization. In: Audio, Language and Image Processing (ICALIP), pp. 200–205 (2014)
20. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proceedings of the 8th European Conference on Computer Vision (ECCV 2004), Prague, Czech, pp. 1–22 (2004)
21. Chen, Z., Zou, B.: Research on Region of Interest Extraction. Central South University, May 2012

RGB-D Salient Object Detection via Feature Fusion and Multi-scale Enhancement

Peiliang Wu(✉), Liangliang Duan, and Lingfu Kong

School of Information Science and Engineering, Yanshan University,
Qinhuangdao, Hebei Province, China
peiliangwu@ysu.edu.cn

Abstract. Salient object detection, especially for multi-object detection in complex scene, is a very challenging issue in computer vision. With the emergence and promotion of somatosensory sensors such as Kinect, RGB-D data jointing color and depth information can be obtained easily and inexpensively. This paper focuses on the RGB-D salient object detection. Firstly, the RGB image is converted into Lab color space and superpixels are segmented according to color and merged according to depth. Then, color contrast features and depth contrast features are calculated to construct an effective multi-feature fusion to generate saliency map. Finally, multi-scale enhancement is operated on the saliency map to further improve the detection precision. Experiments on the public data set NYU depth V2 show that the proposed method can effectively detect each salient object in multi-object scenes, and can also highlight the each object entirely.

Keywords: RGB-D · Superpixel segmentation · Salient object detection · Multi-feature fusion

1 Introduction

The human visual system has the capability to locate the most interest region in a cluttered visual scene, this selective attention mechanism allows us to effectively capture prey and escape predators, which is a very important survival skill for human. Due to its biological importance, many research efforts have been made to find the essence of attention mechanism. A variety of calculation models are proposed to simulate such biological mechanisms in the recent period of time. Visual saliency detection is a very important research in the field of visual attention mechanism. Currently, visual saliency detection is roughly divided into two aspects. One is using High-level visual prior knowledge to mimic human's top-down saliency computational model, whose basic idea is firstly to cluster image pixels into block feature, and through some prior knowledge to simulate the human eye's ability which can identify different objects. The other is the bottom-up detection model based on low-level visual features, whose basic idea is based on visual feature of gray, color or direction for forming the feature map of each feature dimension and merging into final saliency map. Computing visual saliency has very important applications such as image segmentation [1], image classification [2], object

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 359–368, 2015.
DOI: 10.1007/978-3-662-48570-5_35

recognition [3] and other fields, and can optimize the allocation of various computing resources.

At present, many RGB-D sensors, such as Bumbbee camera, PMD camera, especially Kinect have been developed. As the perfect combination of visual camera and ultrasonic sensor, this kind of sensor can obtain scene and object RGB image and depth image simultaneously and is becoming a simple, cheap, convenient environment data acquisition equipment. Studies have been launched, but mainly focus on color RGB-D point cloud data registration [4], 3D reconstruction [5], etc.

Considering the color and depth information are important external data obtained by human vision, RGB-D data will be an important role in promoting research on human visual attention mechanism. In most recent years, salient object detection for RGB-D data gains much attention. Referring to the working mechanism of the human visual system, we propose a saliency calculation framework for RGB-D salient object detection. First of all, superpixels are segmented with Lab color and merged with depth. Then color contrast features and depth contrast features are calculated and fused based on background contrast. Finally, RGB-D image saliency calculation framework is proposed and improved with multi-scale saliency enhancement.

2 Related Work

Recently, RGB image saliency detection has been studied widely and deeply, in which low-level image contrast features play a very important role. The most influential model was proposed by Itti[6] in 1998, by combining low-level image feature such as color, edge and direction etc. and center-surround difference to calculate the image salient region. Harel et al. [7] developed Itti method to generate saliency map and perform the normalized operation based on graph method. Hou et al. [8] proposed a method based on the calculation of the residual spectrum, using the amplitude spectrum information generated by the Fourier transform of the image. Achanta[9] proposed a frequency-tuned approach, in which the distance between the image pixel and the average values of image are calculated as the pixel saliency. Cheng et al. [10] extended color histogram to 3D color space and proposed saliency analysis method based on the color region histogram. Perazzi et al. [11] combine color contrast and color distribution information for image saliency analysis. Margolin et al. [12] proposed a method that combined pattern and color into a model. The above methods can get good results when processing simple images. But when dealing with images containing complex background and several objects, the detection results are bad. Therefore, more saliency factors need to be integrated to solve this problem. RGB-D sensors collecting the color and depth information of the scenes at the same time, is expected to provide depth saliency factor in addition to color. But in terms of salient object detection based on RGB-D data, although several prior works[13-16] aim to explore the saliency analysis of RGB-D, they are still at the initial stage.

3 RGB-D Salient Object Detection

3.1 RGB-D Salient Object Detection Framework

The framework of our RGB-D salient object detection is shown in Fig. 1. For the input RGB-D image, firstly converte RGB into CIELab space and normalize depth into [0-255], secondly, segment superpixels according to Lab color. Thirdly, considering each superpixel as a processing unit, calculate the average depth of each processing unit and merge Lab-based superpixels according to their difference value of average depth(In this paper two superpixel will be merged when their difference value of average depth<10). Then, fuse Lab contrast features and depth contrast features of each merged superpixel to get the global saliency map, and finally, the multi-scale enhancement is designed to improve the detection precision.

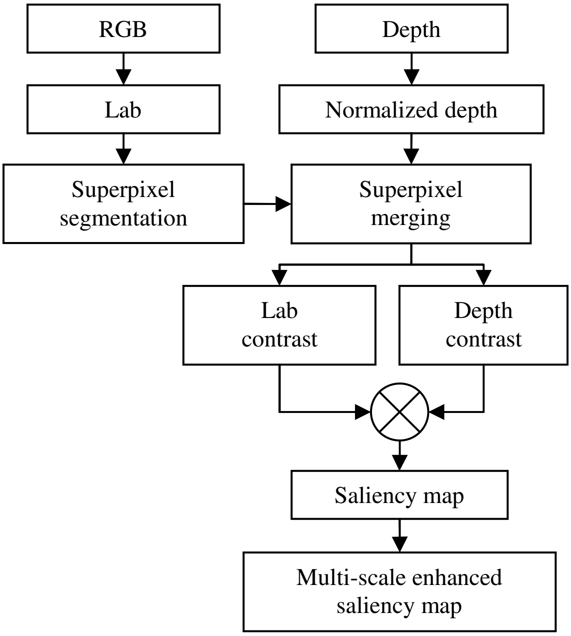


Fig. 1. The framework of RGB-D salient object detection

3.2 RGB-D Superpixel Segmentation

Early salient object detection methods are mainly based on pixel or regularized image unit, and the detection results is unsatisfied. Current methods based on irregular image unit (superpixel) become very popular, including graph cutting [17], Mean shift [18] and SLIC [19], these methods can significantly improve the saliency detection results and generate saliency map with high quality. This paper develops graph cutting segmentation [17] to handle RGB-D images, The details are as follows.

(1) Convert RGB to Lab color, segment the Lab into several disjoint regions $\{O_i\}_{i=1,2,...,N}$ according to [17], Where N is the number of segmented region. $f_c^{(i)}$ denotes the color feature of i^{th} region, where $f_c^{(i)} = \sum_{p\in O_i} V_p \big/ U_i$ denotes the average Lab color of all the pixels in this region, p is the pixel within the region of O_i, V_p denotes Lab feature vector of pixel p, U_i is the number of pixel within region O_i.

(2) Normalize raw depth data. and mapping the normalization results to the range of 0~255. In the segmented regions obtained from (1), calculate the average depth of each region $f_d^{(i)} = \sum_{p\in O_i} D_p \big/ U_i$, where D_p denotes normalized depth of p with value in [0-255].

(3) Merge adjacent segmentation regions when their depth difference<10. And then, recalculate the number of regions N, as well as color feature $f_c^{(i)}$ and depth feature $f_d^{(i)}$.

3.3 RGB-D Contrast Features

Contrast is the most important factor in the low visual saliency calculation. Because the size of each superpixels segmented above is obvious different, we need to consider the size factor to calculate RGB-D contrast of each segmented region $\{O_i\}_{i=1,2,...,N}$.

Color Contrast

Considering the color difference between the segmented image regions(In the Lab color space), the distance of between two regions (in depth channel) and the size of segmented region, the global color contrast feature of a segmented region is defined as follows.

$$S_c^{(i)} = \sum_{k=1,k\neq i}^{N} \varphi(O_i,O_k)\cdot\tau(O_i,O_k)\cdot U_k \tag{1}$$

where $\varphi(O_i,O_k)$ is the color difference between segmented region O_i and O_k (measured in the CIELab color space), the definition as shown in (2).

$$\varphi(O_i,O_k) = \left\| f_c^{(i)} - f_c^{(k)} \right\| \tag{2}$$

$\tau(O_i,O_k)$ denotes smooth term measuring the distance between the different segmented regions of image, which is used to balance the impact of saliency between different positions within the image space.

$$\tau(O_i,O_k) = 1 - D(O_i,O_k) \tag{3}$$

where $D(O_i, O_k) = \left\| c^{(i)} - c^{(k)} \right\|$ denotes spatial distance between region O_i and O_k. When calculating color contrast of region, it has a great impact on adjacent neighbor regions, on the contrary, it has a little impact on long distance regions.

Depth Contrast

Considering the depth difference between two segmented regions (in the depth channel) and the size of the segmented regions, the depth contrast feature of the segmented region O_i is defined as follows.

$$S_d^{(i)} = \sum_{k=1,k\neq i}^{N} \phi(O_i, O_k) \cdot \tau(O_i, O_k) \cdot U_k \tag{4}$$

where $\phi(O_i, O_k)$ is the depth difference between two segmented region O_i and O_k of image.

$$\phi(O_i, O_k) = \left\| f_d^{(i)} - f_d^{(k)} \right\| \tag{5}$$

$\tau(O_i, O_k)$ denotes smooth term measuring the distance between the different segmented regions of image which is defined as equation (3).

3.4 Saliency Features Fusion

When the scene image contains complex background and a variety of objects, it is difficult to detect salient objects accurately only use one single cue. Saliency cues of both color contrast and depth contrast reflect image saliency from different perspectives. Simple linear fusion may make saliency detection bad[20], so it is necessary to design an effective strategy to integrate these saliency cues. In order to highlight each salient object uniformly in a complex and multi-object scene, we use the following feature fusion approach.

$$S_{original}^{(i)} = \exp(S_c^{(i)}) \times S_d^{(i)} \tag{6}$$

So far, the saliency map is obtained by multi-feature fusion with both color and depth channels.

3.5 Multi-scale Saliency Enhancement

Under a single scale, saliency image analysis are often not comprehensive [6,21]. When changing the resolution of the image, the image structure will show different features, so it is very necessary for saliency analysis under multiple scales.

In this paper we use the multi-scale representation of the image to further enhance the saliency detection results, and achieve the goal that highlight each salient object uniformly. In this paper, we down sample RGB-D images into four different scales. Finally, the definition of fusion type of saliency image at multi-scales is defined in (7).

$$S_{final}(I) = \overset{H}{\underset{h=1}{\oplus}} S_{original}(I^h) \tag{7}$$

where I^h are images at different scales, the image of the original scale is I whose h=1. $S_{original}(I^h)$ is the saliency detection result in the single h-scale based on above section. We normalize $S_{final}(I)$, and get the final multi-scale salient object detection results.

4 Experiments and Analysis

We chose NYU Depth V2 as data set, which is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect.

4.1 Comparison of Superpixel Segmentation Results

In order to evaluate the advantages of superpixel segmentation jointing color and depth, we compared graph-cutting based superpixel segmentation on (1) depth, (2) Lab color, and (3) depth + Lab. The three segmentation results are shown in Fig. 2.

It can be seen from the figure that for superpixel segmentation only based on depth, the segmented regions are too large, and adhesion appears easily at the bottom of objects placed on table. For superpixel segmentation only based on color, the segmented regions almost too small, and the whole object is segmented into many small parts, which is called as over-segmentation. Over-segmentation always happens on the texture objects such as boxes and right-side cap in Fig. 2. For our RGB-D superpixel segmentation, because the combination of both color cue and depth cue, superpixels are segmented neither too large nor too small, and the boundary of each object is distinct.

4.2 Salient Object Detection Results

In order to evaluate the advantages of jointing depth and color data to detect salient object, we compared salient object detection at a single scale and at multi-scales respectively based on (1) depth, (2) Lab image, and (3) depth + Lab. The obtained saliency maps are shown in Fig. 3.

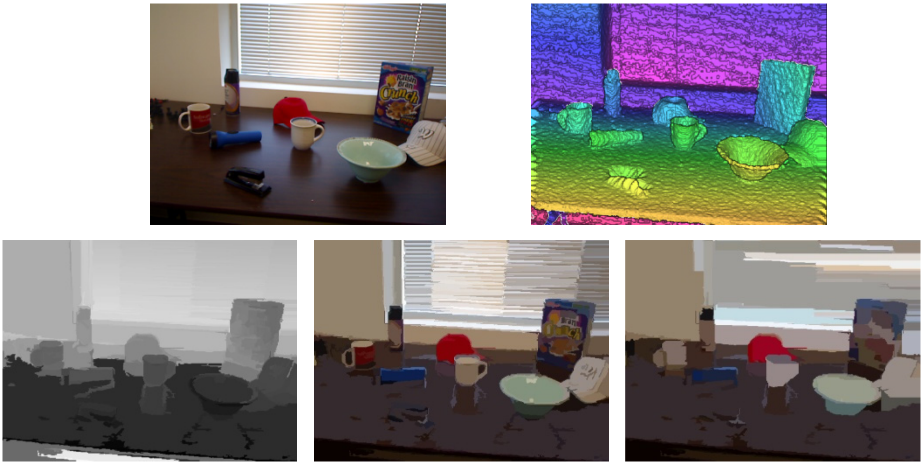


Fig. 2. Graph-cut based Superpixel segmentation on three type of data (top-left: raw RGB, top-right: raw depth; bottom-left: segmentation on depth, bottom-middle: segmentation on Lab, bottom-right: segmentation on Depth + Lab)

As can be seen from Fig. 3, (1) the result of salient object detection is unsatisfied under a single scale, while after multi-scale enhancing, the object regions are highlighted. (2) Over-segmentation in saliency map is distinct when only using color cue, suppression of background is not good when only using depth cue. While the salient object detection jointing color and depth can detect almost every salient object and meanwhile highlight the outline of each object.

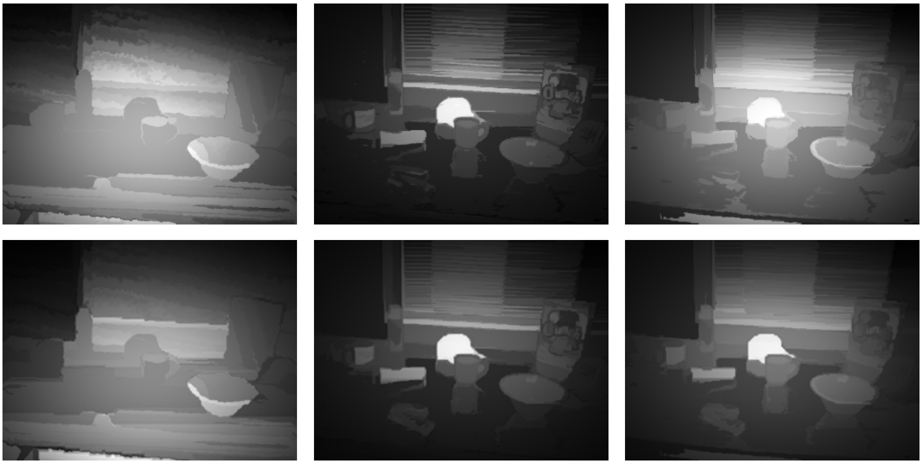


Fig. 3. Salient object detection on three type of data (top: salient object map under single scale, top-left: only depth cue, top-middle: only Lab color cue, top-right: Lab+depth cues. bottom: salient object map with multi-scales enhancement, bottom-left: only depth cue, bottom -middle: only Lab color cue, bottom-right: Lab+Depth cues)

4.3 Contrasts of Salient Object Detection Results

In order to evaluate the proposed salient object detection method jointing depth and color cues, we compared our salient object detection results with early work [16]. The experimental RGB-D images are chosen form four scenes (including desk, kitchen_small and meeting_small, table) from NYU Depth V2. The obtained saliency maps are shown in Fig. 4.

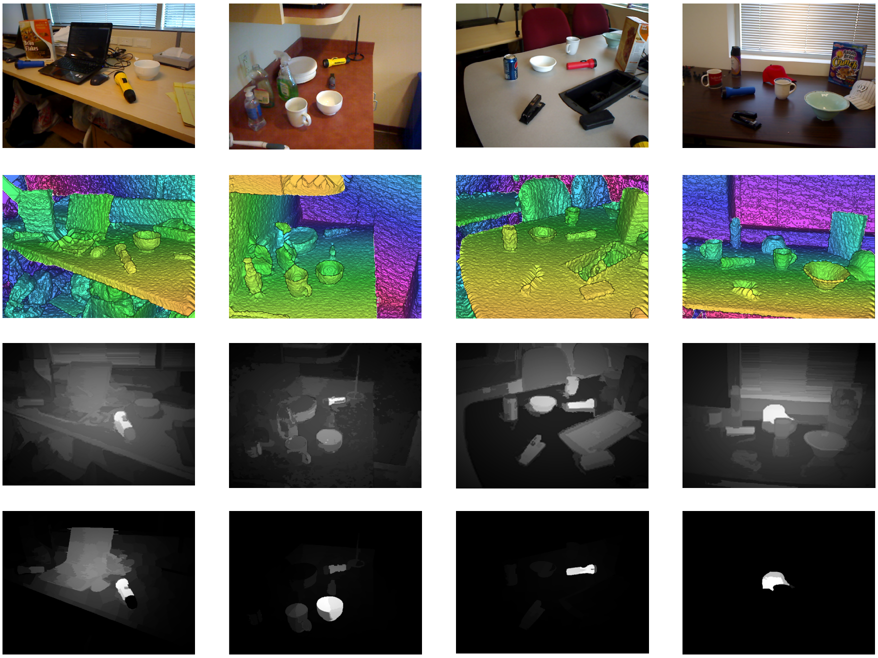


Fig. 4. Contrasts of RGB-D salient object detection on four type of scenes (top: original RGB image; second row: the corresponding depth image; Third row: salient map of our approach; bottom: salient map of [16].)

As we can see from Fig. 4, on contrary, our approach perform better on multi-object scenes, and detect almost each of the salient objects. While [16] only detects one or two objects in the middle part of image.

5 Conclusion

In this paper, we propose a multi-feature fusion framework for RGB-D salient object detection. As a preprocessing stage, RGB-D superpixels are segmented based on graph-cut algorithm. Then color contrast feature and depth contrast feature are extracted and integrated from each superpixel. Under different scales, multi-scale enhancement is designed. The proposed method can produce high quality salient object map which not only highlights each salient object in the multi-object scene, but also can effectively alleviate the over-segmentation.

Because of the complex of the scenes, although our approach can detect almost each of the salient objects, but some background is also highlighted. The next step of our work is to reduce the impact of complex background and improve detection accuracy.

Acknowledgments. This work is supported by National Natural Science Foundation of China No. 61305113.

References

1. Chang, K.Y., Liu, T.L., Lai, S.H.: From co-saliency to co-segmentation: an efficient and fully unsupervised energy minimization model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2136 (2011)
2. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3506–3513 (2012)
3. Ren, Z., Gao, S., Chia, L.T., et al.: Region-based Saliency Detection and Its Application in Object Recognition. IEEE Transactions on Circuits and Systems for Video Technology **24**(5), 769–779 (2014)
4. Hao, M., Biruk, G., Kishore, P.: Color point cloud registration with 4D ICP algorithm. In: IEEE International Conference on Robotics and Automation. Shanghai, China, pp. 1511–1516 (2011)
5. Zollhofer, M., Niebner, M., et al.: Real-time Non-rigid Reconstruction using an RGB-D Camera. ACM Transactions on Graphics **33**(4), 1–12 (2014)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998)
7. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Annual Conference on Neural Information Processing Systems, pp. 545–552 (2006)
8. Hou, X.D., Zhang, L.Q.: Saliency detection: a spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE, Minneapolis (2007)
9. Achanta, R., Hemami, S., Estrada, F., et al.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604. IEEE, Miami Beach (2009)
10. Cheng, M.M., Zhang, G.X., Mitra, N.J., et al.: Global contrast based salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–416. IEEE, Colorado Springs (2011)
11. Perazzi, F., Krahenbuhl, P., Pritch, Y., et al.: Saliency filters: contrast based filtering for salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 733–740 (2012)
12. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1139–1146 (2013)
13. Ciptadi, A., Hermans, T., Rehg, J.M.: An in Depth View of Saliency. In: BMVC, pp. 1–11 (2013)
14. Desingh, K., Krishna, K.M., Jawahar, C.V., Rajan, D.: Depth really matters: improving visual salient region detection with depth. In: BMVC, pp. 1–11 (2013)

15. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth Matters: Influence of Depth Cues on Visual Saliency. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 101–115. Springer, Heidelberg (2012)
16. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD Salient Object Detection: A Benchmark and Algorithms. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part III. LNCS, vol. 8691, pp. 92–109. Springer, Heidelberg (2014)
17. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59**, 167–181 (2004)
18. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**, 603–619 (2002)
19. Achanta, R., Shaji, A., Smith, K., et al.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**, 2274–2282 (2012)
20. Gopalakrishnan, V., Hu, Y., Rajan, D.: Salient region detection by modeling distributions of color and orientation. IEEE Transactions on Multimedia. **11**(5), 892–905 (2009)
21. Goferman, S., Zelnik-Manor, L., Tal, A.: Context- aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(10), 1915–1926 (2012)

Object Color Constancy for Outdoor Multiple Light Sources

Liangqiong Qu, Zhigang Duan, Jiandong Tian, Zhi Han,
and Yandong Tang(✉)

State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese
Academy of Sciences, Shenyang 110016, China
{quliangqiong,duanzhigang,tianjd,hanzhi,ytang}@sia.cn

Abstract. Color is one of the mostly applied features for object recognition and tracking. Most work for color constancy is often based on the assumption of spatial uniformity or smooth illuminant transaction, which is not always true due to the presence of multiple light sources. In this paper, without these assumptions, we deal with the problem of color constancy in multiple light sources by computing the color constancy on a given object rather than the whole image. It keeps the color constancy for a given object under different outdoor lighting conditions, especially for an object under different shadows. We first calculate a transfer vector based on the given object and the illuminants ratio vector. This vector is then added to the original image to make the object be perpendicular to the illuminants ratio vector. Finally, an object color constant image is obtained by performing an orthogonal decomposition along the illuminants ratio vector on the new image. Compared with color constancy on whole image, this proposed method can reduce color distortion in the object and keep mostly color constancy for an object to be recognized and tracked regardless of lighting conditions. Both quantitative and qualitative experiments validate our method.

Keywords: Object color constancy · Illumination invariant · Outdoor multiple light sources · Object detection

1 Introduction

Although color is commonly experienced as an indispensable feature in describing the world around us, the color variation caused by different lighting conditions often introduces undesirable effects in digital images. It may negatively affect the performance of computer vision methods for different applications such as object recognition, tracking and surveillance [1,2]. Consider, for example, an object recognition application which identifies the DARK SKIN checker of Macbeth ColorChecker by color in Fig. 1. It may successfully identifies the DARK SKIN checker in Fig. 1 (a) but fails when the ColorChecker partly lies in shadow (Fig. 1 (b)) and totally lies in shadow (Fig. 1 (c)). This is because the change in the illumination affects object color and further hampers the robustness of

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 369–378, 2015.
DOI: 10.1007/978-3-662-48570-5_36

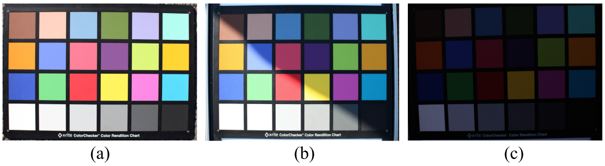


Fig. 1. Sequence of images under different lighting conditions

object recognition and tracking. Therefore, recovering the object color invariant to changing lighting conditions (color constancy) is necessary and worthwhile.

A majority of methods advanced so far for illuminant invariant and color constancy are usually based on the assumption of spatial uniformity. Assuming that the spectral distribution of a light source is uniform across scenes, these methods (such as grey-world [3], white-patch [4], and gamut mapping [5]) get a color constant image by a color correction on original image after globally estimating the color of the light source [2]. More recently, Gijsenij et al. [6] proposed an photometric edge weighting color constancy algorithm based on photometric properties of different edges. Although this assumption works well in most cases, it is often violated as there might be more than one light source illuminating the scene [7].

Retinex theory [8], which assumes that an abrupt change in chromaticity is caused by a change in reflectance properties, is considered as one of the first color constancy methods for multiple light sources. It implies that the illuminant varies smoothly across the image. More specifically, the shadow removal problem [9–11] can be considered as a category of color constancy problem involving two light sources. Even though these shadow removal methods exhibit impressive results for shadow regions, they cannot yield an identical color consistent result regardless of lighting condition (e.g., Fig. 1(c)). Recently, Gijsenij et al. [7] proposed a color constancy method for multiple light sources by applying color constancy locally on small sampled patches. Greatly affected by the effectiveness of sampling method, this method may fail when the distribution of the lighting source is varying.

In order to deal with complex multiple illuminants successfully, these previous mentioned methods either resort to spatial uniformity assumption or smooth illuminant transaction assumption, which are not often true in real situation. In some applications, such as object recognition and tracking, the color constancy on a whole image maybe not necessary, but only the color constancy on the given object is required. In this paper, without spatial uniformity or smooth illuminant transaction, we deal with the color constancy problem for outdoor multiple light sources by computing the color constancy on a given object rather than the whole image. It keeps the color constancy for a given object under different outdoor ligting conditions, especially for an object under different shadows.

This work is based on our previous research on shadow linear model [10] and the color illumination invariant image [12] from the view of atmosphere transmittance effects. As be compressed, our previous color illumination invariant image has some color distortions. In order to make the color of the object keep the same as the canonical color, in this paper, we first calculate a transfer vector based on the given object and illuminants ratio vector. Then we add this transfer vector to the original Log-RGB image to obtain a new transferred image. This will make the object in the new transferred image be perpendicular to the illuminants ratio vector. Finally, the object color constant image is obtained by performing an orthogonal decomposition on the transferred image. Compared with our previous color illuminant invariant on whole image, this method can reduce color distortion in orthogonal decomposition processing and keep mostly color constancy for an object to be recognized and tracked. Both the quantitative and qualitative experiments and comparisons with other methods demonstrate that the color information of our object color constant image can serve as a stable feature for object recognition.

2 Background and Our Previous Work

In this section we first give a brief introduction of the formation of an outdoor image [10] and then we present our pixel-wise orthogonal decomposition for color illumination invariant image [12].

Light emitted from the sun will scattered by atmospheric transmittance effects that causes the incident light to be split into direct sunlight and diffuse skylight. It's revealed that the sRGB tristimulus values of a surface illuminated by daylight are proportional to those of the same surface illuminated by skylight in each of the three color channels [10], i.e.,

$$\log(F_H) = \frac{\log(K_H)}{2.4} + \log(f_H) \tag{1}$$

where F_H denotes the RGB values of a surface in non-shadow area and f_H denotes the RGB values for the same surface in shadow area, $H = \{R, G, B\}$. The proportional coefficients K_H are independent of reflectance and are approximately equal to constants determined by Eq. 2.

$$K_H = \arg\min \sum_{\lambda=400}^{700} |Q_H(\lambda) \cdot (E_{day}(\lambda) - K_H \cdot E_{sky}(\lambda))| \tag{2}$$

Expanding Eq. 1 and letting $\boldsymbol{u} = (u_R, u_G, u_B)^T$ defines a Log-RGB value vector of a pixel, $u_H = log(v_H)$, we have

$$A\boldsymbol{u} = \boldsymbol{I} \tag{3}$$

where $A = \begin{bmatrix} 1 & 1 & -\beta_1 \\ 1 & -\beta_2 & 1 \\ -\beta_3 & 1 & 1 \end{bmatrix}$ and $\boldsymbol{I} = (I_1, I_2, I_3)^T$. $\boldsymbol{I}$ represent a shadow invariant for a pixel in a image [12]. The β_1, β_2 and β_3 in matrix A are calculated as following,

$$\beta_1 = \frac{\log(K_R) + \log(K_G)}{\log(K_B)}, \beta_2 = \frac{\log(K_R) + \log(K_B)}{\log(K_G)}, \beta_3 = \frac{\log(K_G) + \log(K_B)}{\log(K_R)} \tag{4}$$

According to the definitions and calculations of β_1, β_2 and β_3, we have $rank(A) = 2$. Then for a Log-RGB value vector $\boldsymbol{u}$, from algebraic theory, we can obtain an orthogonal decomposition (for more information please refer to [12]):

$$\boldsymbol{u} = \boldsymbol{u_p} + \alpha \boldsymbol{u_0} \tag{5}$$

where $\boldsymbol{u_0}$, satisfying $A\boldsymbol{u_0} = 0$ and $\|\boldsymbol{u_0}\| = 1$, is the normalized free solution of Eq. 3; $\alpha \in R$ and $\boldsymbol{u_p}$ is a particular solution of Eq. 3 such that $\boldsymbol{u_p} \perp \boldsymbol{u_0}$. The symbol $\|\cdot\|$ denotes L^2 norm. Here the free solution has no relationship with the image itself but is determined by matrix A, i.e. illumination condition. $\boldsymbol{u_p}$, $\boldsymbol{u_p} \perp \boldsymbol{u_0}$, is only determined by illumination invariant $\boldsymbol{I}$ and $(\beta_1, \beta_2, \beta_3)^T$. It means that for a pixel with Log-RGB value vector $\boldsymbol{u}$, no matter how different the values of the pixel are with different lighting conditions (within shadow, without shadow or other illuminating), $\boldsymbol{u_p}$ is invariant and only α reflects the variation of pixel RGB values caused by shadow or different illuminating.

Shown in Fig. 2, we use three Macbeth ColorCheckers taken in outdoor scenes at different times on a sunny day to verify these illuminants invariant. It shows that although the three original images are different largely, their color illumination invariant images are almost the same (Fig. 2 (r2)). However, even this color illumination invariant image eliminates the influence of illumination, there still exist some color distortions, which may bring some wrong results for computer vision algorithms, such as object recognition and tracking.

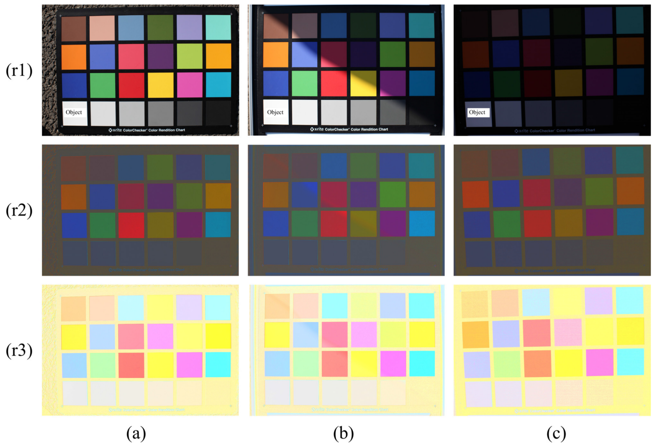


Fig. 2. Orthogonal decomposition and object color constancy. (r1) Original images under different lighting conditions (the WHITE check marked with **object** is the object needs to keep color constancy); (r2) Our color illumination invariant images; (r3) Our object color constant images.

3 Object Color Constancy

For a good prerequisite processing method for object recognition and tracking, it is expected that it can keep the similarity of the object in different lighting conditions meanwhile eliminate or diminish the similarity between the object and the background. In this section, from this point of view, we will introduce a color constancy algorithm for a given object, which will make the color of the object in different lighting conditions keep constant.

Even though the previous color illumination invariant image eliminates the influence of illumination, there still exist some color distortions. In order to make the object have no color distortion, a transfer vector should be added to the object's Log-RGB value vector to make this vector be perpendicular to the illuminants ratio vector. Let the Log-RGB value vector of the object we want to keep color constancy be $\boldsymbol{u}$ and its normalized illumination invariant vector $\boldsymbol{u}_{\boldsymbol{p}}^{t}$ can be calculated according to Eq. 5. Then this transfer vector can be calculated as following,

$$\boldsymbol{T} = \|\boldsymbol{u}\| \cdot \boldsymbol{u}_{\boldsymbol{p}}^{t} - \boldsymbol{u} \tag{6}$$

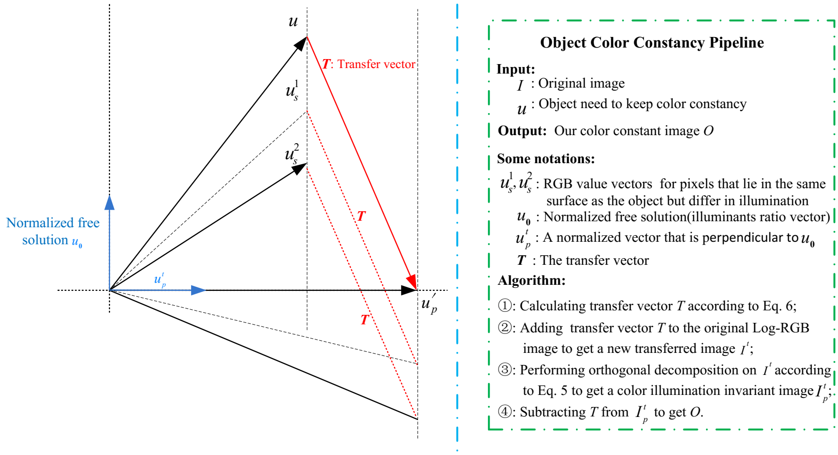
After this transfer, the Log-RGB value vector of this object is perpendicular to the illuminants ratio, which will make the object have no color distortion in our orthogonal decomposition operation.

Given the object we want to keep color constancy in an image, the overview color constancy algorithm can be calculated in the following four steps:

1) Calculating the transfer vector $\boldsymbol{T}$ according to Eq. 6;

2) Adding the transfer vector $\boldsymbol{T}$ to original Log-RGB image to get a new transferred image $\boldsymbol{I}^{t}$;

3) Making an orthogonal decomposition on the new transferred image $\boldsymbol{I}^{t}$ according to Eq. 5 to get a new color illumination invariant image $\boldsymbol{I}_{\boldsymbol{p}}^{t}$. Since



Object Color Constancy Pipeline

Input:
I : Original image
u : Object need to keep color constancy

Output: Our color constant image O

Some notations:
u_s^1, u_s^2 : RGB value vectors for pixels that lie in the same surface as the object but differ in illumination
u_0 : Normalized free solution(illuminants ratio vector)
u_p^t : A normalized vector that is perpendicular to u_0
$\boldsymbol{T}$: The transfer vector

Algorithm:
①: Calculating transfer vector T according to Eq. 6;
②: Adding transfer vector T to the original Log-RGB image to get a new transferred image I^t;
③: Performing orthogonal decomposition on I^t according to Eq. 5 to get a color illumination invariant image I_p^t;
④: Subtracting T from I_p^t to get O.

Fig. 3. An illustration of our object color constancy method.

adding the same transfer vector $\boldsymbol{T}$ on original Log-RGB image does not change the physical properties of the image, performing orthogonal decomposition on $\boldsymbol{I}^t$ will still get an color illumination invariant image like previous section. This can be shown more clearly in Fig. 3. Here, $\boldsymbol{u}$ denotes the pixel value of the pixel that we need to keep color constancy (lies in the canonical lighting condition). $\boldsymbol{u_s^1}$ and $\boldsymbol{u_s^2}$ denote the pixel values of the same pixel lie partly in shadow and totally in shadow, respectively. It shows that, after be added with the transfer vector $\boldsymbol{T}$, these pixels can still be projected along the vector $\boldsymbol{u_0}$ (illuminants ratio vector) into an illumination invariant vector, $\boldsymbol{u_p'}$. Also, as these newly obtained pixels are perpendicular to the illuminants ratio vector, the orthogonal decomposition operation will no longer cause color distortion on these pixels.

4) Subtracting transfer vector $\boldsymbol{T}$ from $\boldsymbol{I_p^t}$ to get the object color constant image.

In Fig. 2, we show our object color constancy method for WHITE checker under different lighting conditions. Unlike the color illumination invariant images in Fig. 2 (r2), the object color constant images (Fig. 2 (r3)) maintain the color information of the original WHITE checker. An more accurate experiment with quantitative analysis will be shown in our experiment section.

4 Experiment

In our experiment, we applied our proposed method for object color constancy on both Macbeth ColorCheckers and real images. We first compare our method with Grey-World method [3] and Weighted Grey-Edge method [6] respectively. And then a set of object recognition experiments based on our results of object color constant images will show the utility of our method.

4.1 Analysis on ColorCheckers

Similar to the previous experiment for object color constancy, in this section we will further give a more accurate experiment with quantitative analysis on those outdoor Macbeth ColorCheckers. A comparison with Grey-World method and Weighted Grey-Edge method will show the effectiveness of our method.

We use the angular error to evaluate the performance of our object color constancy algorithm for its frequent use in the literature [13]. As the angular error is computed pixel by pixel throughout the object, the overall metric of performance of an algorithm for that set of objects can be the mean of errors. For accuracy, in this paper we calculate both mean and median as well as the max error as our measurement to compare different color constancy algorithms. In Fig. 4, we give some examples of object color constant images based on Grey-World, Weighted Grey-Edge methods and our method. The first checker (DARK SKIN checker) marked with **Object** in Fig. 4 (r1,a) is the object we need to keep color constancy. We use the color of this **Object** in daylight as the canonical object color. It can be seen from Fig. 4 (b), the color of the object based on our method are almost the same as the canonical color regardless of lighting

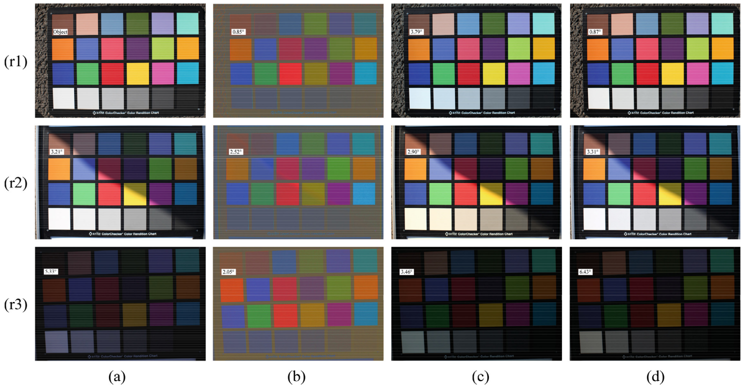


Fig. 4. Examples of object color constant images based on our method, compared to Grey-World method [3] and Weighted Grey-Edge method [6], along with their mean angular error compared to the canonical object color. The first checker in (r1,a) (DARK SKIN checker) marked with **Object** is the object we need to keep color constancy. The color of this DARK SKIN checker is used as the canonical color (ground truth color). The mean angular error is indicated in the left bottom corner of the object. For columns: (a) Original images taken under different lighting conditions, (b) Object color constant images by our method, (c) Color constant images by Grey-World method [3], (d) Color constant images by Weighted Grey-Edge method [6].

conditions. Whereas, even though the Grey-world method [3] and the Weighted Grey-Edge method [6] yield a pleasing result when the input image is illuminated by a uniform illuminant (Fig. 4 (r1)), they cannot deal with images with multiple varying lighting conditions (Fig. 4 (r2, r3)). The relevant quantitative measurement is given in Tab.1. Both the qualitative and the quantitative measurements demonstrate that our object color constant images are considerably closer to generate an canonical object color regardless of lighting condition than both the Grey-World method and Weighted Grey-Edge method.

Table 1. Angular errors for the ColorCheckers in terms of mean, median and max errors for several color constancy algorithms.

	Fig. 4 (r1)			Fig. 4 (r2)			Fig. 4 (r3)		
Methods	Mean	Median	Max	Mean	Median	Max	Mean	Median	Max
Do Nothing	-	-	-	3.21°	2.08°	13.16°	5.33°	4.04°	17.12°
Grey-World	3.79°	3.77°	8.09°	2.90°	2.60°	8.90°	3.48°	3.20°	14.00°
Grey-Edge	0.87°	0.80°	6.54°	3.34°	2.32°	17.47°	6.43°	6.81°	14.27°
Ours	0.86°	0.78°	4.94°	2.52°	2.57°	8.23°	2.05°	2.02°	6.30°

4.2 Applications of Our Object Color Constant Image

As a concrete test of the utility of our calculated object color constant image, we carried out a set of object recognition experiments which identify the object purely by color. Fig. 5 gives one example of this application. In this experiment, we choose the book with bluish green envelope as the object for recognition. These original images were imaged under three different illuminants, one without shadow, one partly in shadow and one totally in shadow. In our experiment, for original images, we adopt angular error to measure the color similarity of different objects. Besides, for comparison, we also evaluate an object recognition experiment based on an illumination and intensity invariant color descriptor: ***rg*histogram** [14] . In our method for object recognition, the root mean square error (RMSE) is used to measure the color similarity.

As shown in Fig. 5, the detection results shows that the using of angular error and the ***rg*histogram** color descriptor are variant to the illumination changes. Therefore, the recognition of this object on original images fails when the object lies partly in shadow (Fig. 5 (r2, b)) or totally in shadow (Fig. 5 (r2, c)). Whereas the object recognition on our object color constant image works quite well (Fig. 5 (r4)). These two experiments on real images show that our proposed method properly gets a color constancy for the given object in the presence of outdoor multiple light sources and can be directly applied to object recognition or tracking.

In addition to the above qualitative experiments, we also give a quantitative result on our proposed new dataset of five objects by comparing the recognition results with the ground truth identified objects. Shown in Fig. 6, our dataset contains 50 images, each of which consists of an original image and a manually marked object image (ground truth identified object). The five objects marked with "**Object1**, **Object2**, **Object3**, **Object4** and **Object5**" are the objects

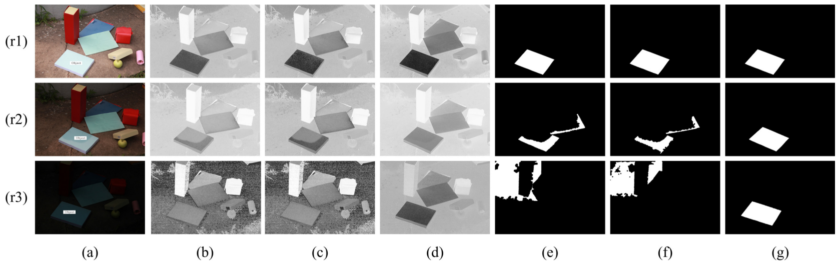


Fig. 5. Object recognition based on our object color constant images and the comparison with original images using angular error and one color descriptor (***rg*histogram**), respectively. For columns: (a) Original images under different lighting conditions. The book with bluish green envelope marked with **Object** is the object we want to identify and its color is used as the canonical color; (b), (c) and (d) The color difference of the original color and the canonical color using angular error, ***rg*histogram** color descriptor and our method, respectively; (e), (f) and (g) The recognition results based on angular error, ***rg*histogram** color descriptor and our object color constant image, respectively.

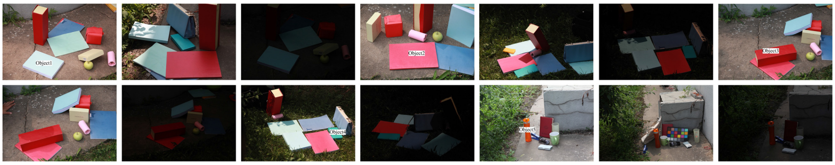


Fig. 6. Example images from our proposed dataset of five object viewed under different illuminants.

Table 2. Comparison of different methods for object recognition on our proposed dataset.

	Angular error			*Rg*histogram			Ours		
Object dataset	PPV	TPR	F1	PPV	TPR	F1	PPV	TPR	F1
Object1	0.6375	0.7332	0.6688	0.6099	0.7027	0.6394	0.9873	0.9367	0.9608
Object2	0.6339	0.4687	0.5375	0.6304	0.5742	0.6000	0.9968	0.9232	0.9570
Object3	0.6505	0.6145	0.6152	0.6655	0.6841	0.6739	0.9963	0.9911	0.9936
Object4	0.8305	0.4011	0.5212	0.7121	0.5185	0.5886	0.9901	0.9394	0.9636
Object5	0.4913	0.4529	0.4708	0.5828	0.5980	0.5866	0.9998	0.9839	0.9917
Mean	0.6487	0.5341	0.5627	0.6401	0.6155	0.6177	0.9941	0.9548	0.9734

that we use to evaluate our method. Each of the object were imaged with ten different illuminants. The precision rate (PPV), recall rate (TPR) and F1 score (F1) are used as the measurement to evaluate the recognition performance.

Table. 2 give the comparison of different methods for object recognition on our dataset of five object. The mean recognition precision rate of the object recognition on original image (angular error) is only 64.87%. Even the so called illuminant invariant color descriptor ***rg*histogram** is applied, the precision rate is still 64.01%. It reveals that this ***rg*histogram** color descriptor isn't really illuminant and shadow invariant and it cannot improve the recognition performance regardless of lighting conditions. While, the precision rate of the object recognition based on our object color constant image has approached 99.41%. This experiment on object recognition dataset clearly suggests that the color of our object color constant image can serve as a stable feature for object recognition.

5 Conclusion

Approaches for color constancy on a whole image under single light source have made considerable progress. However, color constancy on a whole image under multiple light sources remains an open problem. Different from previous work deriving color constancy for the whole image, this paper settles this problem by focusing on the color constancy for a given object. It can keep the color constancy for a given object under different outdoor lighting conditions, especially for an object under different shadows. This proposed method for object color constancy can be directly applied to some applications such as object recognition and tracking and can improve the performance of these methods.

Acknowledgments. This work was supported by the Natural Science Foundation of China under Grant No.61333019 and 61102116.

References

1. Arend Jr., L.E., Reeves, A., Schirillo, J., Goldstein, R., et al.: Simultaneous color constancy: papers with diverse munsell values. JOSA **8**(4), 661–672 (1991)
2. Gijsenij, A., Gevers, T., Van De Weijer, J.: Computational color constancy: Survey and experiments. TIP **20**(9), 2475–2489 (2011)
3. Buchsbaum, G.: A spatial processor model for object colour perception. Journal of the Franklin institute **310**(1), 1–26 (1980)
4. Joze, H.R.V., Drew, M.S.:White patch gamut mapping colour constancy. In: Proc. ICIP, pp. 801–804 (2012)
5. Forsyth, D.A.: A novel algorithm for color constancy. IJCV **5**(1), 5–35 (1990)
6. Gijsenij, A., Gevers, T., Van De Weijer, J.: Improving color constancy by photometric edge weighting. TPAMI **34**(5), 918–929 (2012)
7. Gijsenij, A., Lu, R., Gevers, T.: Color constancy for multiple light sources. TIP **21**(2), 697–707 (2012)
8. Land, E.H., McCann, J.J.: Lightness and retinex theory. JOSA **61**(1), 1–11 (1971)
9. Finlayson, G.D., Hordley, S.D., Lu, C., Drew, M.S.: On the removal of shadows from images. TPAMI **28**(1), 59–68 (2006)
10. Tian, J., Tang, Y.: Linearity of each channel pixel values from a surface in and out of shadows and its applications. In: PROC. CVPR, pp. 985–992 (2011)
11. Tian, J., Wang, Z., Tang, Y.: Static shadow detection: A survey. Information and Control **44**(2), 215–222 (2015)
12. Qu, L., Tian, J., Han, Z., Tang, Y.: Pixel-wise orthogonal decomposition for color illumination invariant and shadow-free image. Optics Express **23**(3), 2220–2239 (2015)
13. Joze, H.R.V., Drew, M.S.: Exemplar-based color constancy and multiple illumination. TPAMI **36**(5), 860–873 (2014)
14. Van De Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. TPAMI **32**(9), 1582–1596 (2010)

Feature-Based 3D Reconstruction Model for Close-Range Objects and Its Application to Human Finger

Feng Liu[1(✉)], Linlin Shen[1], and David Zhang[2]

[1] School of Computer Science & Software Engineering, Computer Vision Institute,
Shenzhen University, Shenzhen, China
feng.liu@szu.edu.cn
[2] Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong, China

Abstract. This paper addresses the problem of feature-based 3D reconstruction model for close-range objects. Since it is almost impossible to find pixel-to-pixel correspondences from 2D images by algorithms when the object is imaged on a close range, the selection of feature correspondences, as well as the number and distribution of them, play important roles in the reconstruction accuracy. Then, features on representative objects are analyzed and discussed. The impact of the number and distribution of feature correspondences is analyzed by reconstructing an object with standard cylinder shape by following the reconstruction model introduced in the paper. After that, three criteria are set to guide the selection of feature correspondences for more accurate 3D reconstruction. These criteria are finally applied to the human finger since it is a typical close-range object and different number and distribution of feature correspondences can be established automatically from its 2D fingerprints. The effectiveness of the setting criteria is demonstrated by comparing the accuracy of reconstructed finger shape based on different fingerprint feature correspondences with the corresponding 3D point cloud data obtained by structured light illumination (SLI) technique which is taken as a ground truth in the paper.

1 Introduction

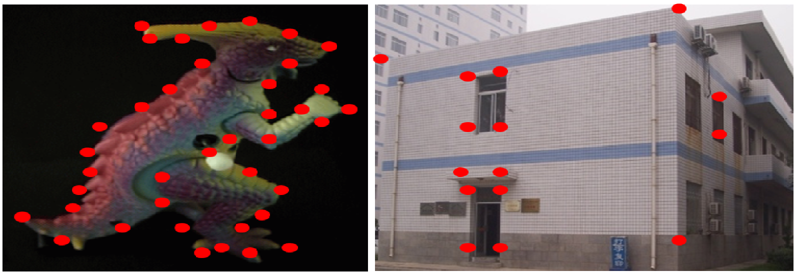
The 3D geometric shape and appearance of objects offer attributes that are invariant to the changes introduced by the imaging process. These attributes can facilitate recognition and assist in various applications, including graphical animation, medical applications, and so forth. Thus, how to obtain the 3D geometric models of real objects has attracted more and more attentions from researchers and companies [1-18]. In computer vision and computer graphics, the process of capturing the shape and appearance of real objects refers to 3D reconstruction. Currently, the existing 3D reconstruction techniques are divided into two categories: active and passive modeling. Active modeling creates the 3D point cloud data of geometric surface by interfering with the reconstructed objects, either mechanically or radiometrically [1-6], while the passive modeling uses only the information contained in the images of the scene to generate the 3D information, namely image-based reconstruction [7-17]. Each of these

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 379–393, 2015.
DOI: 10.1007/978-3-662-48570-5_37

two kinds of modeling has its own advantages and disadvantages. The active modeling reconstructs the 3D model of objects by devices directly with high accuracy but the used devices are costly and cumbersome [18]. The image-based reconstruction generates the 3D model of objects based on their 2D plain images captured by cameras which are challenged to achieve high reconstruction accuracy but the adopted capturing devices (cameras) are usually cheap and light weight [19]. Considering the cost and portability, as well as aiming to make breakthroughs to the reconstruction accuracy, image-based reconstruction is deeply investigated, as summarized in [20, 21].

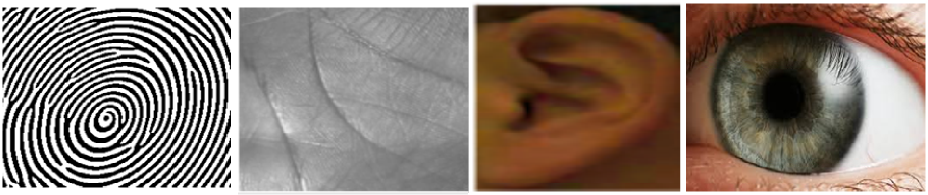
As summarized in [20], there are mainly five kinds of image-based reconstruction methods: shape from shading [7-9], photometric stereo [14-16], stereopsis [10,11], photogrammetry [22-24], and shape from video [12,13]. The shape-from-shading approaches recover the shape of an object from a gradual variation of shading in the image and only one 2D image is needed for depth calculation. Thus, they are the least on equipment requirements but at the price of accuracy and computational complexity [25]. Photometric stereo methods measure 3D coordinates based on different images of the object's surface taken under multiple non-collinear light sources. This kind of methods is an improved version of the shape from shading ones. Higher reconstruction accuracy is achieved due to the usage of more light sources and images [20]. The stereopsis approaches calculate the 3D depth by binocular disparity and two different images captured at the same time are necessary for 3D depth computation. This kind of methods provides better accuracy with less mathematical complexity but difficulty lies in establishing of feature correspondences in two different images automatically and making essential equipment calibrations [26]. Photogrammetry approaches use the same methods to compute the 3D coordinates as the stereopsis ones. Thus, they have similar merits and drawbacks. But, photogrammetry approaches usually use more than two images and produces good results in some types of applications. Typically, they have been successfully applied for modeling archaeological and architectural objects [20]. The shape-from-video approaches render the assumptions in all previous methods since a series of images can be parted from a video. But the problem still lies in the establishment of correspondences from 2D plain images. This kind of methods is usually used in reconstructing terrain, natural targets and buildings [21]. Among all of those methods, photogrammetry approaches are classical and well established ones. They have been around since nearly the same time as the discovery of photography itself [27]. Whereas, photogrammetrists are usually interested in building detailed and accurate 3D models from images. However, in the field of computer vision, work is being done on automating the reconstruction problem and implementing an intelligent human-like system that is capable of extracting relevant information from image data [28]. Thus, algorithms are usually specifically designed for different applications. Currently, the applications of 3D reconstruction approaches are mainly focus on the modeling of terrain, natural targets, as well as archaeological and architectural objects. The characteristics of those kinds of objects are imaged at a long distance and have contour points, as the examples shown in Fig. 1. The reconstruction of these kinds of objects made researchers ignored two important problems met by the reconstruction of close-range objects: one is that it is hard to find contour points or corner points for correspondences establishment on their 2D plain images of the close-range objects, the influence of the selection of feature correspondences, as well as the number of correspondences is increased for the reconstruction of close-range objects. The other one is minor depth difference corresponds to a significant rise of pixel difference on 2D plain images for the reconstruction of close-range objects. The effect of the distribution of correspondences is enlarged in this situation.

Currently, there are no proven results for close-range objects and irregular surfaces like human biometrics (see Fig. 2). Motivated by designing an effective method to model the shape of close-range objects without contour correspondences, a feature-based 3D reconstruction model is investigated in this paper. This 3D modeling was on the base of traditional binocular stereo vision theory. The methodology of the used reconstruction method is first introduced in this paper. Then, for the first time we analyzed the selection of feature points for correspondence establishment for close-range objects, as well as the impact of the number and distribution of feature correspondences on reconstruction accuracy by reconstructing an object with standard cylinder shape and of radius 10mm. The number and distribution of correspondences from two pictures of the cylinder were labeled and selected manually. After that, three criteria were set to guide the selection of feature correspondences on close-range objects for more accurate 3D reconstruction. These criteria were finally applied to the human finger since it is a typical close-range object and different number and distribution of feature correspondences can be established automatically from its 2D fingerprints. The effectiveness of the setting criteria was demonstrated by comparing the accuracy of reconstructed finger shape based on different fingerprint feature correspondences with the corresponding 3D point cloud data obtained by structured light illumination (SLI) technique which was taken as a ground truth in the paper.



(a) (b)

Fig. 1. Example images of archaeological and architectural objects labeled with contour points, (a) dinosaur, (b) buildings.



(a) (b) (c) (d)

Fig. 2. Example images of close-range small objects, (a) finger, (b) palm, (c) ear, (d) iris.

2 Feature-Based 3D Reconstruction Model

Based on the theory of binocular stereo vision [29], the 3D information of an object can be obtained from its two different plane pictures captured at one time. As shown in Fig. 3, given two images C_l and C_r simultaneously captured from two viewpoints, the 3D coordinate of V can be calculated if some camera parameters (e.g., focal length of the left camera f_l, focal length of the right camera f_r, principal point of the left camera O_l, principal point of the right camera O_r) and the matched pair $\left(\left(v_l\left(x_l, y_l\right)\right) \leftrightarrow \left(v_r\left(x_r, y_r\right)\right)$, where $v_*\left(*\right)$ represents a 2D point in the given images C_l or C_r ; x_* is the column-axis of the 2D image, and y_* is the row-axis of the 2D image) are provided. Thus, there are mainly three steps to obtain the 3D space coordinate of points from 2D images, namely camera calculation, correspondence establishment, and 3D coordinates calculation.

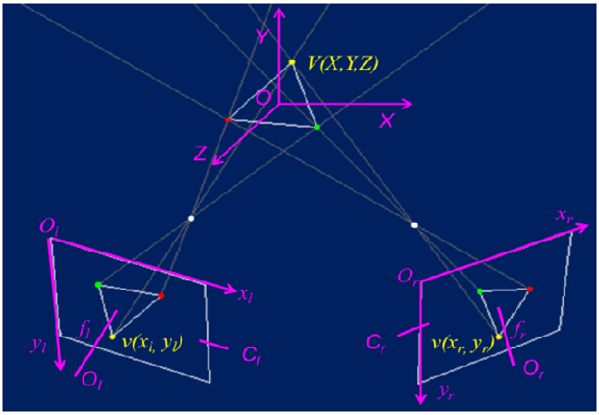


Fig. 3. 3D coordinates calculation on 3D space using binocular stereo vision theory.

Camera calibration refers to the calculation of camera parameters. It is the first step of reconstruction and provides the intrinsic parameters (focal length, principal point, skew, and distortion) of each camera and extrinsic parameters (rotation, translation) between cameras necessary for reconstruction. It usually implements off-line and the commonly used methods and codes are available [30, 31].

Correspondence establishment is of great importance and also a huge challenging problem to 3D modeling. The methods for correspondence establishment are categorized into two classes: feature-based approach and correlation technique [32-34]. Feature-based approach usually produces sparse depth maps by matching feature correspondences while correlation technique yields to dense depth maps by matching all pixels in the entire images. Each has merits and drawbacks. Feature-based approach is suitable when good features can be extracted from 2D images, relatively insensitive to illumination changes and faster than correlation technique. But it usually just provides sparse depth maps. While correlation technique is easier to implement than feature-based method and can provide a dense depth map. It does not work well when viewpoints are very different. Generally, feature-based approach is preferable than correlation technique by taking both accuracy and time complexity into account.

The 3D coordinate of each correspondence can be calculated by using the stereo triangulation method when given camera parameters and matched pairs between images of different views [31].

However, to obtain the 3D surface of an object, it is necessary to produce dense depth maps. They are two ways to realize 3D surface reconstruction by feature-based approach. One is to establish pixel-to-pixel correspondence by estimating the transformation model between 2D images based on feature correspondences (labeled by Framework I). The other one is to find representative feature correspondences from 2D images and given a prior shape model then reconstructing the 3D surface by interpolation (labeled by Framework II). The first framework of reconstruction technique is similar to the correlation-based one due to the establishment of pixel-to-pixel correspondence which has drawbacks of low accuracy and high time complexity. This paper thus studied reconstruction technique by following Framework II. Based on Framework II, this paper focused on investigating the influence of feature correspondences establishment to 3D reconstruction accuracy for close-range objects. The model of the proposed 3D reconstruction model is shown in Fig. 4.

3 Criteria for Close-Image Objects Reconstruction

Fig. 5 shows an example of the reconstruction result based on the model given in Fig. 4. It can be seen that the correspondences established on the objects are almost contour or corner points labeled manually. It is invalid for close-range objects without contour or corner points on them, which raises problems of the selection of representative features for correspondence establishment. Meanwhile, the number and distribution of feature correspondences also plays an important role in the 3D

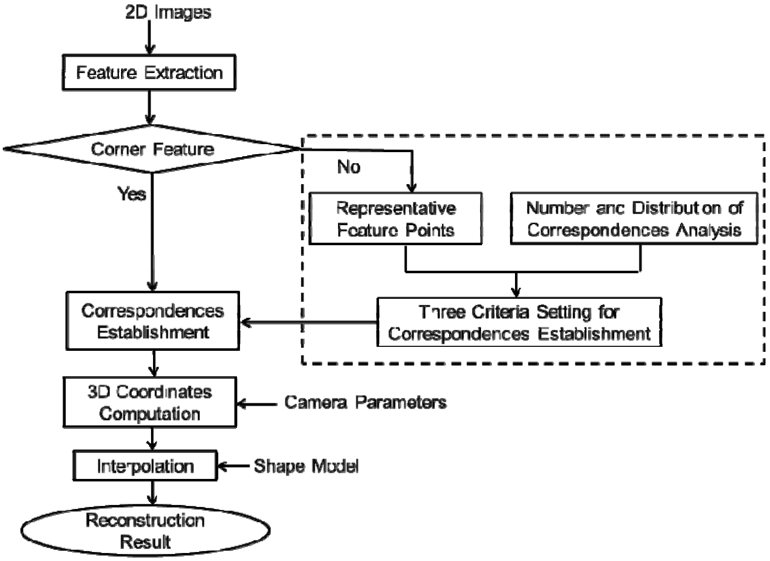
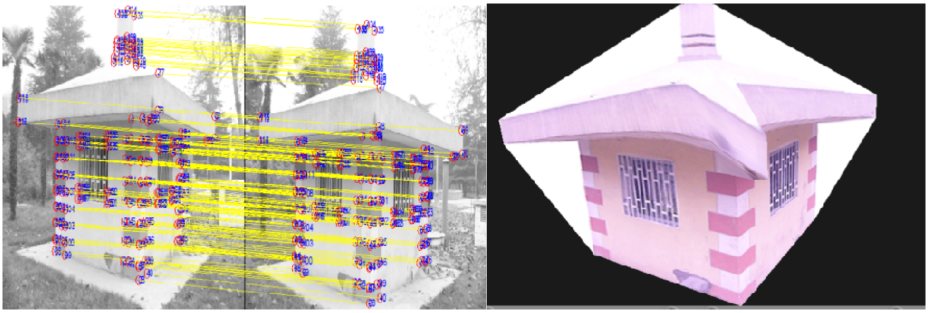

2D Images
Feature Extraction
Corner Feature
No
Yes
Representative Feature Points
Number and Distribution of Correspondences Analysis
Three Criteria Setting for Correspondences Establishment
Correspondences Establishment
3D Coordinates Computation
Camera Parameters
Interpolation
Shape Model
Reconstruction Result

Fig. 4. The proposed 3D reconstruction model in this paper.



(a) (b)

Fig. 5. Building reconstruction results by following the model shown in Fig. 4. (a) contour or corner correspondences establishment result, (b) 3D reconstruction result wrapped with texture image.

reconstruction accuracy. These two problems are studied in-depth in the following subsections. Finally, three criteria are set based on the previous analysis so as to guide feature correspondences establishment for 3D modeling of close-range objects.

3.1 Selection of Representative Features for Correspondence Establishment

Generally speaking, it is intuitive that corner points which refers to the intersection of two lines or point which located in two adjacent objects with different principle lines will be selected as representative features for correspondence establishment, as the points manually labeled in Fig. 5. However, there are no corner points in some objects, as the example images shown in Fig. 2. It is necessary to find representative feature points or corner-like points to instead corner points for correspondence establishment. By observing the images of objects in Fig. 2, we can see that lines or regions of variation are widespread in them. In this paper, we assume the points located in the position with changes as representative feature points. There are three typical situations (e.g. on line, between lines, between regions) we summarized as follows.

The first two situations are relative to lines, which are for a single line and between lines. These two situations are analyzed based on fingerprint images since they consist of lines. As the solid line labeled in the example fingerprint image shown in Fig. 6, for a single line, it can be seen that changes occur in the end of the line or the point where its direction changes largely. Generally, the end of a line is defined as termination point (triangle labeled in Fig. 6) and the point where lines' direction largely changed refers to local extreme point (circle labeled in Fig. 6). Thus, such two kinds of points are selected as representative feature points for a single line in this paper. In the situation of between lines (see dashed lines labeled in Fig. 6), change just exist in the intersection point of lines, the representative feature point is then defined as the intersection point between lines (rectangle labeled in Fig. 6).

The third situation is between regions. Besides lines, some objects consist of different regions, as the iris image shown in Fig. 7. It can be seen that it contains regions with different textures and colors. Similar to the situation between lines, changes occurs in the boundary between adjacent regions, as circles labeled in Fig. 7. Thus, in the third case, representative feature points are defined as the points in the boundary between adjacent regions.

Finally, we summarize the first criterion for feature correspondences establishment to the 3D reconstruction of close-range objects as: *Criterion 1: selecting representative feature points or corner-like points for correspondence establishment.*

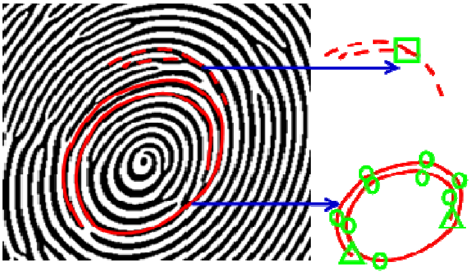


Fig. 6. Illustration of representative feature points of lines in a fingerprint image.

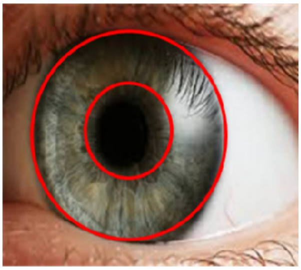


Fig. 7. Illustration of representative feature points between regions in an example iris image.

3.2 Influence Analysis of Correspondences to Reconstruction Accuracy

As known to everyone, it is extremely difficult to establish pixel-to-pixel feature correspondences manually or automatically, especially for close-range objects due to their small size and the unknown number and irregular distribution of feature points on them. To the best of our knowledge, there are no literatures available about the influence analysis of number and distribution of feature correspondences to 3D reconstruction accuracy. This subsection thus studied this problem by reconstructing a small object with standard cylinder shape and of radius 10mm. To facilitate the labeling of feature correspondences, we wrapped the cylinder with a grid paper, as shown in Fig. 8(a). As the representative feature points defined in the previous subsection, those feature points which are located in the boundary between adjacent regions are manually labeled for correspondence establishment. By following the method introduced in Section 2, the shape of the cylinder can be reconstructed. Here, a 3D software was used for the display and analysis of reconstruction results. This software is popularly used for 3D point cloud data display and analysis.

First, experiments were organized to analyze the effect of the number of correspondences on 3D reconstruction accuracy. Thus, the distributions of selected feature correspondences were all even. The largest number of correspondences between two 2D images shown in Fig. 8(b) is set to 40 for the reason that this is the largest area those points covered in the experiments. The number of feature correspondences both along horizontal axis and vertical axis was gradually reduced to get different reconstruction results. Table 1 lists the setting of parameters in the experiments, such as the number of feature correspondences, the distribution of correspondences, and the sampling interval and direction along decreasing number. The reconstruction results were also summarized in Table 1. Details of the corresponding feature correspondences establishment and reconstruction results were given in Fig. 9. From the results, we can see that the reconstruction accuracy dropped with the decreasing of correspondence number. From Tabel 1, we can see that there is a little influence of number decreasing on reconstruction accuracy for a line shape (vertical axis). For a curved shape (horizontal axis), the accuracy decreased when correspondence number decreasing and sampling interval increasing. For the same number of feature correspondences, the closer between correspondences, the larger the error may be due to the effect of errors resulted in 3D coordinate calculation for each correspondence.

Therefore, we set the second criterion for feature correspondences establishment to the 3D reconstruction of close-range objects as: *Criterion 2: Densely sampling of feature correspondences along the direction where depth changed quickly and sparsely sampling of feature correspondences along the direction where depth smoothly changed.*

We also conducted an experiment of 3D reconstruction by randomly selecting feature correspondences with irregular distributions. The selected feature correspondences and its reconstruction result are shown in Fig. 10, where the number of correspondences is around 20. By comparing the result with Enum-2, Enum-3 and Enum-4 shown in Fig. 8 (they have similar number of correspondences), we can see that better results were achieved with large sampling interval no matter the distribution of feature correspondences is even or not. Thus, the third criterion for feature correspondences establishment to the 3D reconstruction of close-range objects in this paper is: *Criterion 3: Establishing feature correspondences to cover the surface of object as large as possible.*

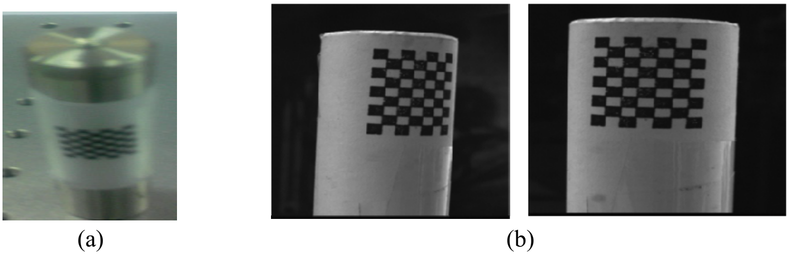
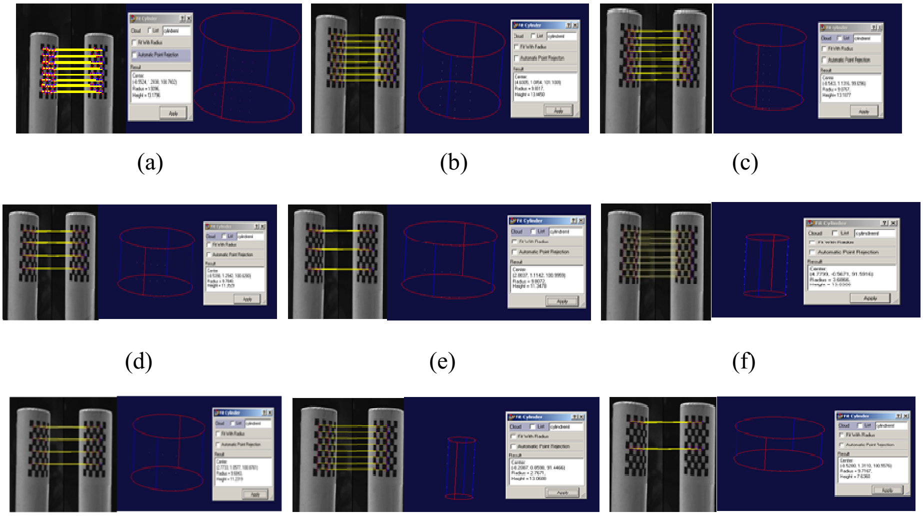


Fig. 8. (a) Original cylinder shape object wrapped with grid paper, (b) 2D images of (a) captured by the left and right cameras in the experiments.

Table 1. Setting of experimental parameters and the corresponding reconstruction results.

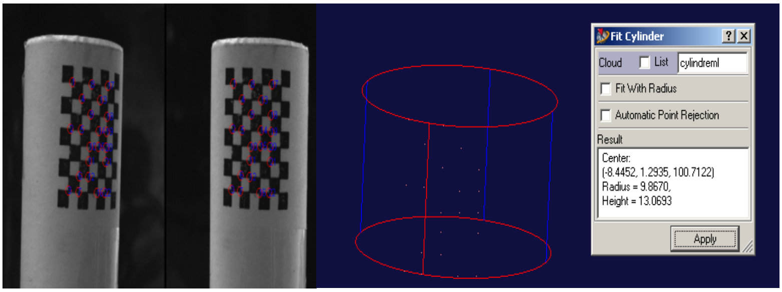
Title of experiments	Enum-1	Enum-2	Enum-3	Enum-4	Enum-5	Enum-6	Enum-7	Enum-8	Enum-9
Number of feature correspondences	40	24	24	20	15	12	12	12	10
Distribution of correspondences	Even	Even	Even	Even	Even	Even	Even	Even	Even
(Sampling interval, Direction along decreasing number)	(2mm, -)	(4mm, horizontal)	(2mm, horizontal)	(4mm, vertical)	(6mm, vertical)	(6mm, horizontal)	(8mm, vertical)	(8mm, horizontal)	(10mm, vertical)
Reconstruction Radius (standard 10mm)	9.91	9.85	9.08	9.78	9.80	2.77	9.68	3.69	9.71


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 9. Established feature correspondences and the reconstruction result for (a) Enum-1, (b) Enum-2, (c) Enum-3, (d) Enum-4, (e) Enum-5, (f) Enum-6, (g) Enum-7, (h) Enum-8, (i) Enum-9, listed in Table 1.



Result

Center:
(-8.4452, 1.2935, 100.7122)
Radius = 9.8670,
Height = 13.0693

Fig. 10. Established feature correspondences with irregular distribution and the reconstruction result.

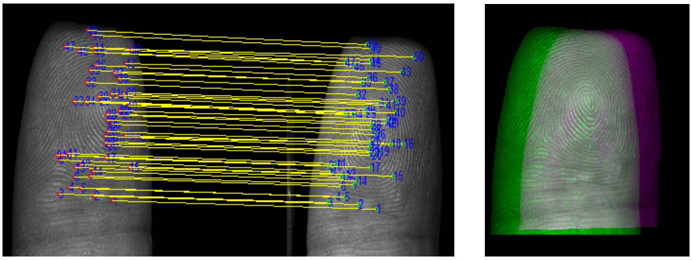
4 Case Study: Application to Human Finger

As we can see that human fingers are typical close-range objects. To verify the effectiveness of the proposed reconstruction model and the criteria to the reconstruction accuracy for close-range objects, this paper took the reconstruction of finger shape as a case study. The device used to capture 2D fingerprint images was the same as the one introduced in Ref. [35].

4.1 Effectiveness Validation of the Proposed Reconstruction Model

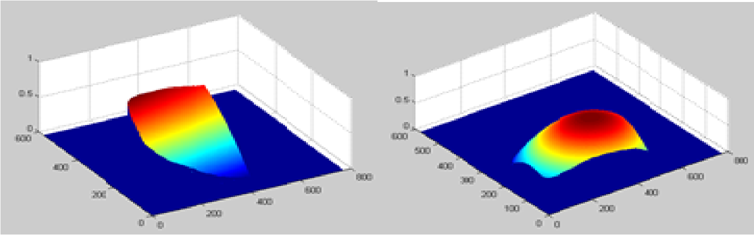
As mentioned in Section 2, there are two frameworks to realize 3D results by using feature-based reconstruction technique. The paper selected Framework II in the proposed reconstruction model. This subsection tries to demonstrate the effectiveness of the proposed model by reconstructing a human finger with two frameworks mentioned in Section 2. First, we manually labeled 50 representative feature correspondences on example fingerprint images by following the criteria set in Section 3, as shown in Fig. 11(a). Then, pixel-to-pixel correspondences were established by estimating the transformation model between images based on previously labeled feature correspondences. The result is shown in Fig. 11(b). Here, the rigid transform was selected as the model between images. After that, 3D reconstruction results can be achieved by following the procedures given in Section 2, as shown in Fig. 12. For better comparison, the depth of the reconstruction result is normalized to [0, 1] by MIN-MAX rule. From Fig. 12, we can see that the result obtained by the proposed model is closer to the appearance of human finger than the one generated by following the procedure of framework I.

Furthermore, we compared the reconstruction results with the 3D point cloud data of the same finger to verify the effectiveness of the model. The 3D point cloud data are defined as the depth information of each point on the finger. They are collected by a camera together with a projector using the Structured Light Illumination (SLI) method [36, 37]. Since this technique is well studied and proved to acquire 3D depth information of each point on the finger with high accuracy [36, 37], 3D point cloud data obtained using this technique are taken as the ground truth of the human finger in this paper. Compared our results in Fig. 12 with the ground truth shown in Fig. 13, it can be seen that the profile of the human finger shape reconstructed based on the proposed model is similar to the 3D point cloud data even though it is not that accurate. Meanwhile, the reconstruction result based on framework I shown in Fig. 12(a) is quite different from the 3D point cloud data. The real distances between the upper left core point and the lower left delta point of the reconstruction results in Fig. 12(a) and Fig 12(b), as well as of the ground truth in Fig. 13(a) were also calculated. The corresponding values are 0.431, 0.353 and 0.386, respectively. As a result, it is concluded that the proposed model is effective even though there is an error between the reconstruction result and the 3D point cloud data.



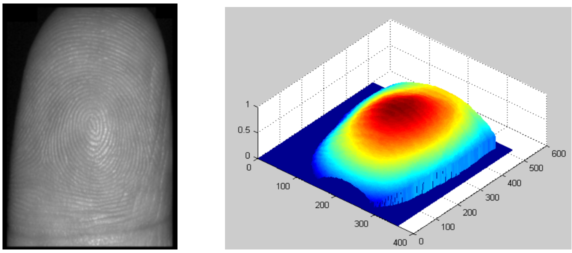
(a) (b)

Fig. 11. Correspondences establishment results. (a) manually labeled 50 presentative feature correspondences, (b) pixel-to-pixel correspondences (gray part in the center) after rigid transformation between fingerprint images.



(a) (b)

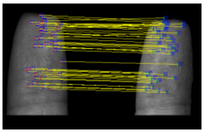
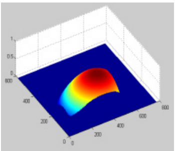
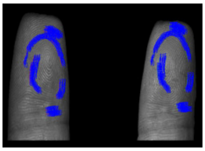
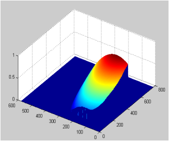
Fig. 12. 3D reconstruction results. (a) reconstruction result based on the model of Framework I, (b) reconstruction result based on our proposed reconstruction model.



(a) (b)

Fig. 13. Ground truth of the same finger of Fig. 11 captured by structured light illumination (SLI) technique. Comparison of 3D fingerprint images from the same finger but different acquisition technique: (a). original fingerprint image captured by the camera when collecting 3D point cloud, (b). 3D point cloud collected by one camera and a projector using the SLI method.

Table 2. Reconstruction results from different fingerprint feature correspondences of Fig. 11.

Results Used feature	Established correspondences	Reconstructed 3D fingerprint image
Minutiae		
Ridge feature		

4.2 Criteria Verification

This paper proposed three criteria to guide feature correspondences establishment for the 3D reconstruction of close-range objects. The effectiveness of such criteria was verified by analyzing the reconstruction accuracy based on different fingerprint feature correspondences. As studies in [38], there are two classical fingerprint features for low resolution fingerprint images, namely ridge feature and minutiae. Feature correspondences were first established automatically from the images shown in Fig. 11 by using the algorithms introduced in [35, 38, 39], and then the reconstruction results were generated based on those three different fingerprint feature correspondences, as illustrated in Table 2. It can be seen that the results are different corresponding to different feature matched pairs due to different numbers and distribution of established fingerprint feature correspondences and also the existence of false correspondences.

From the results shown in Table 2, it can be seen that the reconstruction result based on minutiae correspondences is better than the one based on ridge feature. The histograms of error maps between the results in Table 2 and the ground truth in Fig. 13(b) are shown in Fig. 14. Smaller errors were achieved between the minutiae-based reconstruction result and the ground truth. These results fully demonstrated the proposed criteria in this paper that: (1) minutiae, which refer to the ends or bifurcations of ridges, are satisfied the definition of representative feature points or corner-like points in the paper. However, ridge feature, which is the sampling of lines, provides too much insignificant information. Thus, it is better to select representative feature points or corner-like points for correspondence establishment; (2) poor result will be achieved if densely establishing correspondences along the direction with smoothly changed depth, like ridge feature correspondences. Therefore, we recommended to sparsely sampling of feature correspondences along the direction where depth smoothly changed; (3) the region covered by minutiae correspondences is larger than the one covered by ridge correspondences, better reconstruction result is achieved correspondingly. Hence, this paper set the third criterion for feature correspondences establishment to cover the surface of object as large as possible.

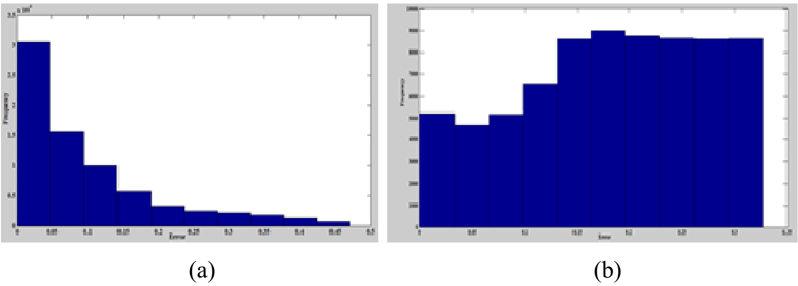


Fig. 14. Histogram of error maps between reconstructed results in Table 2 and Fig. 13(b). (a) histogram of err map between Fig. 13(b) and reconstruction result by using minutiae, (c) histogram of err map between Fig. 13(b) and reconstruction result by using ridge feature.

5 Conclusion

The issue of feature-based 3D reconstruction method for close-range objects was investigated in this paper. For close-range objects, it is very hard to found pixel-to-pixel correspondence from their 2D images. Thus, our study mainly focused on 3D modeling with limited feature correspondences. In this situation, the selection of representative feature correspondences, the number and distribution of the feature correspondences play an important role in the 3D reconstruction accuracy. Then, features on representative close-range objects were analyzed and the suitable features for correspondence establishment were indicated. Subsequently, the impact of the number and distribution of feature correspondences was analyzed by reconstructing an object with standard cylinder shape and of radius 10mm. Three criteria were set to guide the selection of features on close-range objects for more accurate 3D reconstruction. We finally took the reconstruction of human finger as a case study by applying our setting criteria. The effectiveness of the setting criteria was demonstrated by comparing the accuracy of reconstructed finger shape based on different fingerprint feature correspondences with the corresponding 3D point cloud data obtained by structured light illumination (SLI) technique which was taken as a ground truth in the paper.

Acknowledgement. The authors would like to thank the editor and the anonymous reviewers for their help in improving the paper. The work is supported by the NSFC fund 61403257 and Natural Science Foundation of SZU 201532.

References

1. Rusinkiewicz, S., Holt, O., Levoy, M.: Real-time 3D model acquisition. In: The Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, no. 3, vol. 21, pp. 438–446, July 2002
2. Bradley, B., Chan, A. , Hayes, M.: A simple, low cost, 3D scanning system using the laser light-sectioning method. In: IEEE International Instrumentation and Measurement Technology Conference Victoria, Vancouver Island, Canada, pp. 299–304, May 2002

3. Wu, Q., et al.: A 3D modeling approach to complex faults with multi-source data. Computers & Geosciences **77**, 126–137 (2015)
4. Wang, Y., Hassebrook, L., Lau, D.: Data acquisition and processing of 3-D Fingerprints. IEEE Transactions on Information Forensics and Security **5**(4), 750–760 (2010)
5. Stockman, G., Chen, S., Hu, G., Shrikhande, N.: Sensing and recognition of rigid objects using structured light. IEEE Control Syst. Mag. **8**(3), 14–22 (1988)
6. Hu, G., Stockman, G.: 3-D surface solution using structured light and constraint propagation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(4), 390–402 (1989)
7. Horn, B.: Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical Report No. 232, AI Lab., MIT (1970)
8. Zhang, R., Tsai, P., Cryer, J., Shah, M.: Shape from shading: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(8), 690–706 (1999)
9. Worthington, P.: Reillumination driven shape from shading. Computer Vision and Image Understanding **98**(2), 326–344 (2005)
10. Akimoto, T., Suenaga, Y., Wallace, R.: Automatic creation of 3D facial models. IEEE Computer Graphics & Applications **13**(5), 16–22 (1993)
11. Lao, S., Sumi, Y., Kawade, M., Tomita, F.: Building 3D facial models and detecting face pose in 3D space. In: Proc. of Second Int. Conf. on 3D Digital Imaging and Modelling, pp. 398–404 (1999)
12. Brodski, T., Fermuller, C., Aloimonos, Y.: Shape from video. IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2**, 1–151 (1999)
13. Strecha, C., Verbiest, F., Vergauwen, M., Van Gool, L.: Shape from video vs. still images. In: Proceedings Conference on Optical 3D Measurement Techniques, Zürich, Switzerland, vol. 2, pp. 168–175 (2003)
14. Woodham, R.: Photometric method for determining surface orientation from multiple images. Optical Engineering **19**(1), 139–144 (1980)
15. Rushmeier, H., Taubin, G., Gueziec, A.: Applying shape from lighting variation to bump map capture. In: Proc. of Eurographics Work. on Rendering, St. Etienne, France, pp. 35–44 (1997)
16. Malzbender, T., Wilburn, B., Gelb, D., Ambrisco, B.: Surface enhancement using real-time photometric stereo and reflectance transformation. In: Eurographics Workshop on Rendering, Switzerland, pp. 245–250 (2006)
17. Grün, A.: Semi-automated approaches to site recording and modeling. Int. Archives of Photogrammetry and Remote Sensing **33**(B5), 309–318 (2000)
18. Bradley, B., Chan, A., Hayes, M.: A simple, low cost, 3D scanning system using the laser light-sectioning method. In: Proceedings of the IEEE International Instrumentation and Measurement Technology Conference Victoria, Vancouver Island, Canada, pp. 299–304, May 2002
19. Paris, S.: Methods for 3D reconstruction from multiple images
20. Said, A., Halabi, H., Baharum, B.: Image-based modeling: a review. Journal of Theoretical and Applied Information Technology, 188–196 (2005)
21. Remondino, F., El-Hakim, S.: Image-based 3D modeling: A review. The Photogrammetric Record **21**(115), 269–291 (2006)
22. Divya, U.J., Kim, H.S., Kim, J.I.: An image-based approach to the reconstruction of ancient architectures by extracting and arranging 3D spatial components **16(**1), 12–27 (2015)
23. Grün, A., Remondino, F., Zhang, L.: Photogrammetric reconstruction of the Great Buddha of Bamiyan, Afghanistan. The Photogrammetric Record **19**(107), 177–199 (2004)

24. Xia, S., Zhu, Y.: 3D simulation and reconstruction of large-scale ancient architecture with techniques of photogrammetry and computer science. In: CIPA 2005 XX International Symposium, Torino, Italy (2005)
25. Zhang, R., Tsai, P., Cryer, J., Shah, M.: Shape-from-shading: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(8), 690–706 (1999)
26. Poggio, G., Poggio, T.: The analysis of stereopsis. Annual Review of Neuroscience **7**(1), 379–412 (1984)
27. Hartley, R., Mundy, J.: The relationship between photogrammetry and computer vision. In: Barrett, E.B., McKeown, D.M. (eds.) SPIE Proceedings. Integrating Photogrammetric Techniques with Scene Analysis and Machine Vision, vol. 1944, pp. 92–105. SPIE Press (1993)
28. Henrichsen, A.: 3D reconstruction and camera calibration from 2D Images. University of Cape Town (2000)
29. Hartley, R.: Multiple view geometry in computer vision. Cambridge Univ. Press, Cambridge (2000)
30. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. on Pattern Anal. Mach. Intell. **24**(11), 1330–1334 (2000)
31. Bouguet, J.: Camera Calibration Toolbox for Matlab
32. http://en.wikipedia.org/wiki/Correspondence_problem#Basic_Methods
33. Ogale, A., Aloimonos, Y.: Shape and the stereo correspondence problem. International Journal of Computer Vision **65**(3), 147–162 (2005)
34. Belhumeur, P., Mumford, D.: A bayesian treatment of the stereo correspondence problem using half-occluded regions. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 506–512 (1992)
35. Liu, F., Zhang, D., Song, C., Lu, G.: Touchless Multi-view Fingerprint Acquisition and Mosaicking. IEEE T. Instrumentation and Measurement **62**(9), 2492–2502 (2013)
36. Wang, Y., Hassebrook, L., Lau, D., Data acquisition and processing of 3-D Fingerprints, IEEE Transactions on Information Forensics and Security, **5**(4), 750–760 (2010)
37. Zhang, D., Kanhangad, V., Luo, N., Kumar, A.: Robust palmprint verification using 2D and 3D features. Pattern Recognition **43**(1), 358–368 (2010)
38. Choi, H., Choi, K., Kim, J.: Mosaicing touchless and mirror-reflected fingerprint images. IEEE Trans. On Information Forensics and Security **5**(1), 52–61 (2010)
39. Jain, A., Hong, L., Bolle, R.: On-line fingerprint verification. IEEE Trans. on Pattern Anal. Mach. Intell. **19**(4), 302–314 (1997)

Contour Extraction Based on Human Visual System

Xiaosheng Yang(✉) and Yinfeng Li

School of Electronic Engineering, Xidian University, Xi'an 710071, China
xiaoshengy@stu.xidian.edu.cn

Abstract. The contour extraction in image processing and computer vision is extremely an important image analysis method. On the basis of features of the primary visual cortex (V1 area) neurons which will inhibit or enhance the response to the specific area of the visual field, this paper improves the traditional Gabor function, establishes more effective mathematical models of visual receptive field and proposes an algorithm based on visual perception mechanism. Experiments demonstrate that the algorithm can extract the image contour efficiently.

Keywords: Contour extraction · Primary visual cortex · Gabor function · Mathematical models

1 Introduction

For humans, visual system is the most important and direct way to acknowledge the world, analyze the external environment and respond accordingly. With the development of computer performance and functionality, how to reveal and simulate the human visual system has been a research focus. A complex natural scene image contains a wealth of information and it's impossible for sight to give the same level of attention to every point in space. The human visual system experiments [1-4] demonstrate that the contour feature in images is particularly important. They retain the border (useful structure information) about the object, while greatly reducing the amount of data, thereby simplifying the forms of expression, so that the visual can handle the ever-changing inputs in a timely and effective manner. In many cases, the object can be identified according to the outline of objects. In the past few years, researches [5-6] on contour extraction based on visual attention mechanism have made tremendous progress, but how to extract significant contour features of complex images is still a pressing problem quickly and accurately.

The traditional edge detection method is the classical operator method, namely by means of a spatial difference operator, convolve the image with the template such as Gradient operator, Laplace operator, Canny operator and so on. In early 1980, Canny presented the canny edge detection operator from the point of signal processing, which is theoretically a relatively complete edge detection operator. Although the several operators are simple to achieve and fast to operate, they both failed to properly

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 394–405, 2015.
DOI: 10.1007/978-3-662-48570-5_38

deal with an edge noise interference brought by an actual image texture, which leads to precisely a result that extracted contour accuracy can't be guaranteed. Based on this, we need to address two problems existing in the traditional edge detection methods: firstly, inhibit the noise brought by the texture information; secondly, make the discrete edge pixels continuous.

Studies on the optic nerve [7-10] showed that: the retinal process includes forming the central- peripheral receptive field of the bipolar cells and ganglion cells. Other cells in the retina, particularly horizontal cells and amacrine cells transfer lateral information (transfer from one neuron to the same layer adjacent neurons) to form a more complex receptive fields, such as motion sensitive and color insensitive receptive field or color sensitive and motion insensitive receptive field. Related experiments [8, 10-13] demonstrated that when neurons in the visual cortex respond to stimulation with a specific space, receptive field plays an important role in combing and organizing the contour. Various representative models have been established based on this feature. Grossberg [14] et al proposed a boundary contour system to detect some false contour generated by the visual illusion. Li [15] proposed a significant edge detection method to locate edge information by detecting the edge orientation and homogeneous boundary point. These visual models are mainly used to explain how the human visual system to achieve a combination of contour and segmentation of boundary, mainly for processing synthetic images instead of natural images.

Knierim [16] et al, proposed environment suppression domain applied to the contour extraction of natural scenes, but the environment suppression domain is isotropic. On the basis of the properties of non-classical receptive field in the primary visual cortex, Cosmin Grigorescu [17] et al made comprehensive consideration of isotropic and anisotropic suppression, and proposed an effective algorithm to outstand the boundary and save the orphaned contour.

Environment suppression can reduce some edge-texture noise but leave behind many discrete and fracture edge segments. In order to detect a more complete edge of objectives, we also need to further combine and connect the edge segments. Geisler et al [18] thought that the visual cortex's stimulation response to the components possessing a consistent space agencies will be strengthened. And it has two characteristics: if local ingredients are smoother and closer, there will be a greater probability of being aggregated into global contours; if two local ingredients are close and touch the same circle, then this local contour will have a higher significance, which is e well applied to the algorithms for making the edge segments continuous.

According to problems existing in the traditional edge detection methods, basing on the physiological mechanism of visual saliency, combining with significant computer calculations, the paper presents an algorithm based on visual perception mechanism. The basic flow chart of the algorithm is shown below.

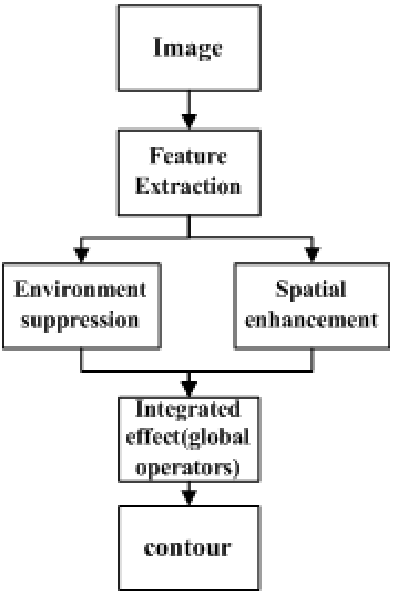


Fig. 1. The Basic Flow Chart of the algorithm

This paper will establish a more effective mathematical model (improved Gabor function) to describe human visual system in the feature extraction Module, and then use improved Gabor energy based on human visual system to achieve effects of environment suppression and spatial enhancement, finally use a global operator to obtain the goal of extracting the contour. Experiments demonstrate that this method has strong anti-interference, high precision and can meet the actual needs of the engineering survey compared with classical contour extraction methods.

2 Feature Extraction

Gabor function can simulate the structure of receptive field. It is possible to simulate the response to complex cellular by Gabor energy function to obtain the energy diagram of the visual characteristics.

However, studies [19-22] showed that in the process of the receptive field structure predicted by Gabor function gradually increasing as the center frequency of visual pathways, the receptive field center and the periphery will generate the phenomenon of alternate oscillation, which is inconsistent with most of the neurons well-known to us in the visual receptive field structure.

Longxiang You [23] et al thought that owing to the spectral distribution of each spatial frequency of the visual pathway channel having a certain overlap, the visual system information processing can't be equivalent to the compression and recovery process of the spectrum. Thus, Gabor function cannot predict the complex structure of receptive fields and the corresponding mathematical description of it as a visual receptive fields of neurons need to be improved.

Longxiang You [23] et al presented mathematical description of mathematical models of isotropic and anisotropic visual receptive fields under the premise of analyzing shortages of the existing mathematical models of visual information processing neuron receptive field. In addition to the relationship with the distribution parameters of receptive field models, they also studied its response to spatial frequency domain.

In this paper, learning from their models, we improve the Gabor function, which uses the Laplace transform of Gaussian function to optic spatial distribution model of nerves receptive fields, so as to achieve the purpose of extracting the contour of the target.

The calculation process is as follows:

For the human visual system, the process of extracting feature edges in spatial domain and spatial frequency spectrum can be expressed as following:

$$\mathrm{f}_\sigma(\mathrm{x}',\mathrm{y}') = \iint_{s_1} \mathrm{f}(\mathrm{x},\mathrm{y}) \times \mathrm{h}(\mathrm{x}' - x, \mathrm{y}' - y)dxdy \tag{1}$$

$$\mathrm{F}_\sigma(\mathrm{u},\mathrm{v}) = \mathrm{F}(\mathrm{u},\mathrm{v}) \times H(u,v) \tag{2}$$

Where $\mathrm{h}(\mathrm{x},\mathrm{y})$ is the system kernel, s_1 is the spatial domain, $\mathrm{f}(\mathrm{x},\mathrm{y})$ is the input, $\mathrm{f}_\sigma(\mathrm{x},\mathrm{y})$ is the output.

Depending on the difference of treatments, treatments will be divided into mathematical description to isotropic neurons and anisotropic neurons.

Laplace transform expression of Gaussian function is as follows:

$$\nabla^2 G(x,y) = \frac{-1}{\pi\sigma^4}\left(1 - \frac{x^2+y^2}{2\sigma^2}\right)\exp\,(-\frac{(x^2+y^2)}{2\sigma^2} \tag{3}$$

In engineering applications, we use the difference between different spatial distributions of two Gaussian functions to approximate. Expression is shown below.

$$\mathrm{h}(\mathrm{x},\ \mathrm{y}) = \frac{1}{2\pi\sigma_1^2}e\,\mathrm{xp}\left(-\frac{(x^2+y^2)}{2\sigma_1^2}\right) - \frac{1}{2\pi\sigma_2^2}e\,\mathrm{xp}\left(-\frac{(x^2+y^2)}{2\sigma_2^2}\right) \tag{4}$$

The system kernel is as follows:

$$\mathrm{h}(\mathrm{x},\mathrm{y}) = \mathrm{g}(\mathrm{x},\ \mathrm{y})\exp\,(j2\pi fx) \tag{5}$$

Gaussian function has good smoothness and locality in space domain and spatial frequency, so the Laplace transform is suitable as kernel function of neuronal receptive field. We will name the response to image by the system kernel as the improved Gabor energy, the expression is as follows:

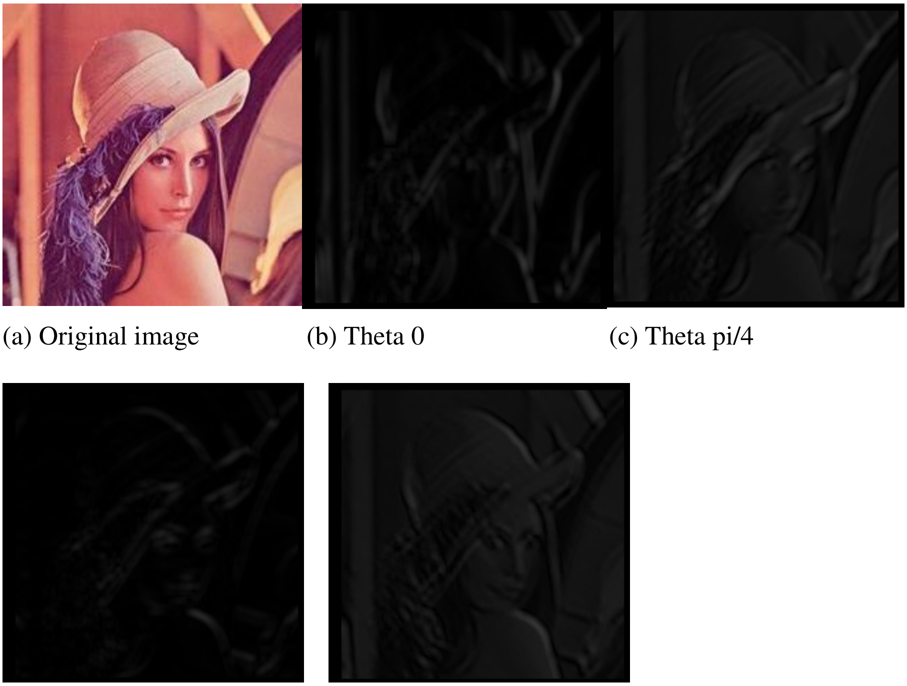
$$\mathrm{E}_\sigma(\mathrm{x},\mathrm{y}) = \sqrt{\mathrm{E}_e(\mathrm{x},\mathrm{y})^2 + \mathrm{E}_o(\mathrm{x},\mathrm{y})^2} \tag{6}$$

$$\mathrm{E}_e(\mathrm{x},\mathrm{y}) = \mathrm{f}(\mathrm{x},\mathrm{y}) * \mathrm{h}_e(x,y) \tag{7}$$

$$\mathrm{E}_o(\mathrm{x},\mathrm{y}) = \mathrm{f}(\mathrm{x},\mathrm{y}) * \mathrm{h}_o(x,y) \tag{8}$$

Where $\mathrm{h}_e(x,y)$, is the real part and $\mathrm{h}_o(x,y)$ is the imaginary parts.

We use the improved the Gabor filter in four directions to operate the Lena. The results are shown below.



(a) Original image (b) Theta 0 (c) Theta pi/4

(d) Theta pi/2 (e) Theta pi*3/4

Fig. 2. The Improved Gabor Filter

The improved Gabor energy can identify the chaotic texture the chaotic texture. However, the response of improved Gabor energy function is only partial orientation information. So to achieve effects of environmental suppression and spatial enhancement, we also need to design a global operator on the basis of the improved Gabor energy function.

3 Environment Suppression

Neurons in the primary visual cortex preferentially respond to stimulation with particular space. When the stimulation is larger than the area they feel, these neurons will be suppressed and have effects on aggregate perception of contours and lines. Studies [24-27] have shown that modulation of environment on neural response is considered to be the basis of many sensory phenomena.

Due to neurons possessing direction selectivity in the human primary visual cortex, these cells suppress non-classical receptive field around the area, and have the impact on aggregate perception of contours and lines. Cosmin Grigorescu [17], who was inspired to propose a significant computing model called the environment suppression in order to improve the detection of contours and boundaries of the target area of a natural scene image. It is calculated as follows:

(1) Environment suppression
Cosmin Grigorescu [21] operated the Gauss difference to simulate human primary visual cortex's suppression on the environment. DoG expression is shown below.

$$\mathrm{DoG}\sigma(\mathrm{x},\mathrm{y}) = \frac{1}{2\pi(4\sigma)^2}\exp\left(-\frac{\mathrm{x}^2+\mathrm{y}^2}{2(4\sigma)^2}\right) - \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\mathrm{x}^2+\mathrm{y}^2}{2\sigma^2}\right) \tag{9}$$

Custom expression of weighting function is shown below.

$$\mathcal{W}\sigma(\mathrm{x},\mathrm{y}) = \frac{\mathrm{H}(\mathrm{DOG}\sigma(\mathrm{x,y}))}{\|\mathrm{H}(\mathrm{DOG}\sigma)\|1} \tag{10}$$

$$\mathrm{H}(\mathrm{z}) = \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases} \tag{11}$$

$\|*\|\ 1$ represents the norm of L1.

(2) Isotropic and anisotropic suppression
The effect of isotropic suppression only considers the distance factor. The expression is shown below.

$$F_{env}(x,y) = \iint_{\Omega} \mathrm{E}_{\sigma}(\mathrm{x}-\mathrm{u},\mathrm{y}-\mathrm{v})\,\mathcal{W}_{\sigma}(\mathrm{u},\mathrm{v})dudv \tag{12}$$

Where $\mathrm{E}_{\sigma}(\mathrm{x},\mathrm{y})$ is the improved Gabor energy response, Ω is the spatial domain.

Compared to the isotropic suppression, the anisotropic suppression adds a suppression factor namely the direction factor. There are two points (x, y) and (x-u, y-v) and the expression of the direction factor is shown below.

$$\Delta_{\Theta,\sigma}(\mathrm{x},\mathrm{y},\mathrm{x}-\mathrm{u},\mathrm{x}-\mathrm{v}) = |\cos(\Theta_{\sigma}(\mathrm{x},\mathrm{y}) - \Theta_{\sigma}(\mathrm{x}-\mathrm{u},\mathrm{y}-\mathrm{v}))| \tag{13}$$

According to the above equation, if these two points are in the same direction and then the suppression is greatest. When the angle increases, the inhibition is reduced because of cos (0) = 1, When the two points are perpendicular to each other, the inhibitory will have minimum effects (cos (π / 2) = 0).

After adding the direction factor, the expression is shown below.

$$F_{env}(x,y) = \iint_{\Omega} \mathrm{E}_{\sigma}(\mathrm{x}-\mathrm{u},\mathrm{y}-\mathrm{v})\,\mathcal{W}_{\sigma}(\mathrm{u},\mathrm{v}) \times |\cos(\Theta_{\sigma}(\mathrm{x},\mathrm{y}) - \Theta_{\sigma}(\mathrm{x}-\mathrm{u},\mathrm{y}-\mathrm{v}))|dudv \tag{14}$$

4 Spatial Enhancement

Geisler et al [18] thought that the visual cortex's stimulation response to the components possessing a consistent space agencies will be strengthened. And it has two characteristics: if local ingredients are smoother and closer, there will be a greater probability of being aggregated into global contours; if two local ingredients are close and touch the same circle, then this local contour will have a higher significance, which is known as co-circular rules. Concyclic geometric relationship is shown in Figure 2. If an azimuth of the center position is α (0⩽ α < π), then the azimuth of co-circular geometry β will satisfy:

$$\beta = \begin{cases} 2\gamma - \alpha + \pi, & 2\gamma - < 0 \\ 2\gamma - \alpha, & 0 \le 2\gamma - \alpha < \pi \\ 2\gamma - \alpha - \pi, & \pi \le 2\gamma - \alpha \end{cases} \tag{15}$$

γ is the orientation of the connection of center and the ambient component (0≤γ <π).

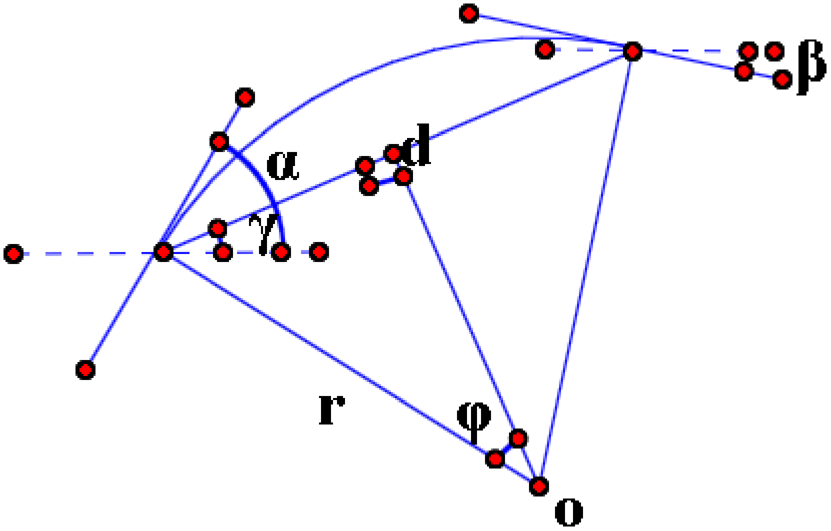


Fig. 3. Concyclic Geometric Diagrams

The curvature is an important factor to determine the detectability of the natural contours, and concyclic curvature k is calculated as follows:

$$k = \frac{1}{\gamma} = \frac{2}{d}\sin(\theta) = \begin{cases} \frac{2}{d}\sin\left|\frac{\beta-\alpha}{2}\right|, & 0 \le 2\gamma - \alpha < \pi \\ \frac{2}{d}\sin\left|\frac{\beta-\alpha}{2}\right|, & 2\gamma - \alpha < 0\, or 2\gamma - \alpha > \pi \end{cases} \tag{16}$$

α, β are the optimal orientation of the reference point of the center-periphery receptive field. You can get the weighting function according to the curvature and distance decay function, which is calculated as follows (D is a normalization constant):

$$W_c(x, y, \alpha; x', y', \beta) = \exp\left(-\frac{k^2}{2{\sigma_c}^2}\right) \tag{17}$$

$$W_d(d) = \frac{1}{D}\exp\left(-\frac{d^2}{2{\sigma_d}^2}\right) \tag{18}$$

The expression of spatial enhancement is as follows:

$$F_{\mathrm{air}}(x,y,\alpha) = \sum_{d\in R} \sum_{\beta} W_c(x,y,\alpha;x',y',\beta)\, W_d(d) \times \mathrm{E}_{\sigma}(x',y',\beta) \tag{19}$$

Where $\mathrm{E}_{\sigma}(x,y,\beta)$ is the best improved Gabor energy response in the direction of β .

5 Integrated Model

Environment suppression can suppress texture contour noise, spatial enhancement can combine discrete significant contour and the property model can highlight significant contours. Then use the property model to integrate these mechanisms, and achieve the goal of extracting significant goal contour ultimately. Comprehensive formula is as follows:

$$F_{\mathrm{n+1}}(x,y,\alpha_i) = \mathrm{F}(x,y,\alpha_i) + \eta(n)\big(F_{\mathrm{air}}(x,y,\alpha_i) - c \cdot F_{\mathrm{env}}(x,y)\big) \tag{20}$$

$$\mathrm{F(x,y)} = \overset{MAX}{i}\big(F_{\mathrm{n+1}}(x,y,\alpha_i)\big) \tag{21}$$

$$F_{\mathrm{com}}(x,\ y) = \sum_{x,y\in n} \mathrm{E(x,y)} \tag{22}$$

Where $\eta(n)$ the iteration parameter and c is enhanced suppression coefficient determining the degree of the inhibition and enhanced in the model.

$\mathrm{i} = 1,2,3,\dots,\mathrm{k}$ represents the number of directions of the improved Gabor energy function. $F_{\mathrm{com}}(x,\ y)$ is the comprehensive effects.

6 Global Operator

We use the non-maximal suppression and hysteresis thresholding that Canny used in his classic paper [28] to obtain the results of the binary processing. The specific algorithm process is as follows:

(1) Calculate the gradient
Firstly, convolute the artwork with Gaussian function, then smooth and filter the image, finally perform the finite difference operation on the filtered image. In Canny operator, the expression is given below.

$$\mathcal{F}\sigma(\mathrm{x,y}) = (\mathcal{F} * \mathcal{G}\sigma)(\mathrm{x,y}) \tag{23}$$

$$\mathcal{G}\sigma(\mathrm{x,y}) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\mathrm{x}^2+\mathrm{y}^2}{2\sigma^2}\right) \tag{24}$$

$$\nabla\mathcal{F}\sigma(\mathrm{x,y}) = (\frac{\partial\mathcal{F}\sigma(\mathrm{x,y})}{\partial x},\frac{\partial\mathcal{F}\sigma(\mathrm{x,y})}{\partial y}) \tag{25}$$

However, the study concluded that using the above expression to calculate is ill-posed. Therefore, use the following expression to calculate the gradient.

$$\mathcal{F}\sigma(\mathrm{x},\mathrm{y}) = (\mathcal{F} * \mathrm{h})(\mathrm{x},\mathrm{y}) \tag{26}$$

$$\mathrm{h}(\mathrm{x}, \mathrm{y}) = \frac{1}{2\pi\sigma_1^2} e\,\mathrm{xp}\left(-\frac{(x^2+y^2)}{2\sigma_1^2}\right) - \frac{1}{2\pi\sigma_2^2} e\,\mathrm{xp}\left(-\frac{(x^2+y^2)}{2\sigma_2^2}\right) \tag{27}$$

$$\nabla\mathcal{F}\sigma(\mathrm{x},\mathrm{y}) = (\mathcal{F} * \nabla\mathrm{h})(x,y) \tag{28}$$

Wherein, ∇h (x, y) is the first derivative of the Gaussian function. Then have further to calculate the gradient along the X, Y directions and the expression is shown below.

$$\nabla x\mathcal{F}\sigma(\mathrm{x},\mathrm{y}) = \left(\mathcal{F} * \nabla\frac{\partial\mathrm{h}}{\partial x}\right)(\mathrm{x},\mathrm{y}) \tag{29}$$

$$\nabla y\mathcal{F}\sigma(\mathrm{x},\mathrm{y}) = \left(\mathcal{F} * \nabla\frac{\partial\mathrm{h}}{\partial y}\right)(\mathrm{x},\mathrm{y}) \tag{30}$$

Calculate the gradient and then get the magnitude and direction. The expression is shown below.

$$\mathrm{M}_\sigma(\mathrm{x},\mathrm{y}) = \sqrt{\nabla x\mathcal{F}\sigma(\mathrm{x},\mathrm{y})^2 + \nabla y\mathcal{F}\sigma(\mathrm{x},\mathrm{y})^2} \tag{31}$$

$$\Theta_\sigma(\mathrm{x},\mathrm{y}) = \tan^{-1}\frac{\nabla y\mathcal{F}\sigma(\mathrm{x},\mathrm{y})}{\nabla x\mathcal{F}\sigma(\mathrm{x},\mathrm{y})} \tag{32}$$

(2) Combine effects

After the first step of the calculation, using the non-maxima suppression along the direction of gradient to locate the contour pixels of the image can obtain the contour image. In this step, we need to suppress background texture and noise of contour images and enhance the spatial.

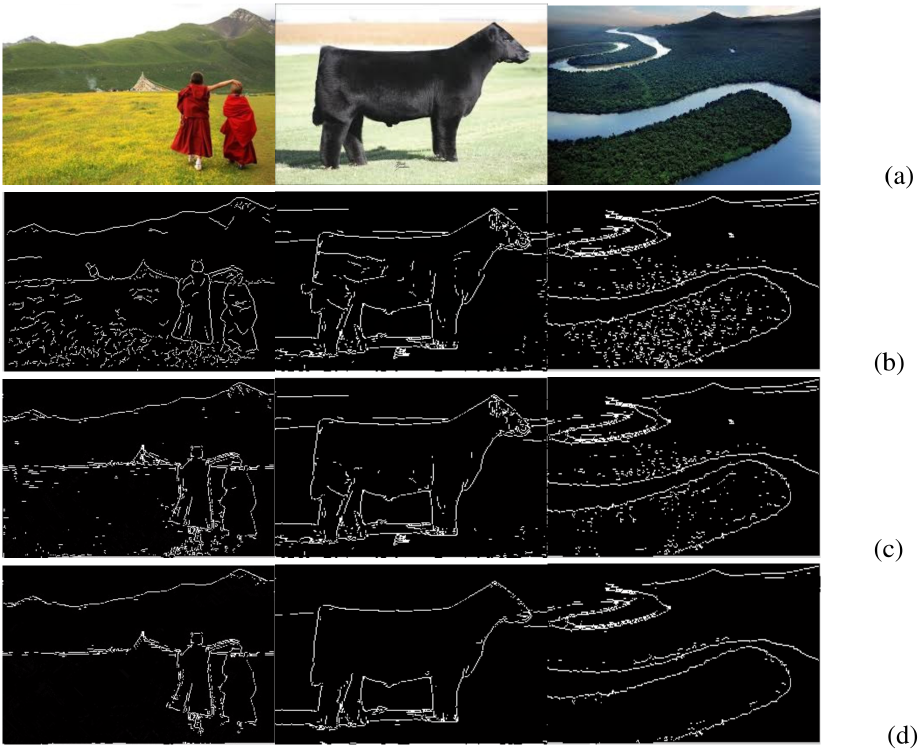
Finally, get the contour extraction operator and its expression is shown below.

$$C_\sigma^f(x, y) = H(\mathrm{M}_\sigma(\mathrm{x},\mathrm{y}) - \alpha F_{\mathrm{com}}(x, y)) \tag{33}$$

H (*) is operating operator involving the refinement of contours, hysteresis thresholding and contour connections. The specific operation can refer canny operator [28].

7 Results and Analysis

In order to test the effect of contour extraction, select a number of images from the library Pinterest. Furthermore, do an experiment and compare it with other algorithms. Results are as follows:



(a) Original Image (b) Canny Contour Extraction
(c) Traditional Detection Algorithm (d) the Proposed Algorithm

Fig. 4. Image Contour extraction Result

According to the performance criteria of the contour extraction:

$$P = \frac{card(E)}{card(E)+card(E_{fP})+card(E_{fN})} \tag{34}$$

Wherein card (E) represents the number of members in the set E; E, E_{fP}, E_{fN} represent the correct contour, false contour and omissions contour respectively. Performance testing index of Figure 3 is shown in the following table (frequency of bandwidth B_f = 1.5, the number of samples k=12, orientation bandwidth B_θ =-r / 6, variance of DoG σ=3, variance of weighting function=-r / 6):

Table 1.

Algorithm	Performance P	Performance P	Performance P
	Group1	Group2	Group3
Canny	0.20	0.15	0.15
Represent	0.30	0.30	0.27
The paper	0.32	0.33	0.31

Through the above performance comparison, this algorithm is significantly better than the canny algorithm and better than the representative detection algorithm.

8 Conclusion

In computer vision, the image contour extraction is a necessary link. Selecting the appropriate image contour extraction method is undoubtedly very important. According to traditional contour extraction methods' problems, the paper improves the Gabor function model and proposes the idea of using the Gaussian function Laplace transform as the mathematical description of receptive fields. In addition, the paper learns from the Gabor function's suppression effect on the environment and spatial enhancement of the integrated mechanism in order to establish the visual fusion model. Experiments show that this method has a continuous contour extraction, high precision, single-pixel width and other characteristics. When as compared with classical edge detection, contour extraction method given herein is of strong anti-interference, good stability and can meet on computer vision measurement requirements.

References

1. Held, R., Shattuck, S.R.: Color and edge-sensitive channels in the human visual system: Tuning for orientateon. Science **174**, 314–316 (1971)
2. Koivisto, M., Mantyla, T., Silvanto, J.: The role of early visual cortex (V1/V2) in conscious and unconscious visual perception. Neuroimage **51**, 828–834 (2010)
3. Larsson, J., Heeger, D.J., Landy, M.S.: Orientation selectivity of motion-boundary responses in human visual cortex. Journal of Neurophysiology **104**, 2940–2950 (2010)
4. Montaser-Kouhsari, L., Landy, M.S., Heeger, D.J., Larsson, J.: Orientation-selective adaptation to illusory contours in human visual cortex. Journal of Neuroscience **27**, 2186–2195 (2007)
5. Marr, D., Hildreth, E.C.: Theory of edge detection. Proceedings of the Royal Society **207**, 187–217 (1980)
6. Field, D.J., Hayes, A., Hess, R.F.: Contour integration by the human visual system: evidence for a local 'association field'. Vision Res. **33**(2), 173–193 (1993)
7. Fischer, S., Dresp, B., Kopp, C.: A Neural Network Model for Long-Range Contour Diffusion by Visual Cortex (2000). doi:10.1007/3-540-45482-9_33
8. Ulinski, P.S.: Neural Mechanisms Underlying the Analysis of Moving Visual Stimuli (1999). doi:10.1007/978-1-4615-4903-1_6
9. Glantz, R.M., Barnes, W.J.P.: Visual Systems: Neural Mechanisms and Visual Behavior (2002). doi:10.1007/978-3-642-56092-7_12
10. Walles, H., Robins, A., Knott, A.: A Neural Network Model of Visual Attention and Group Classification, and its Performance in a Visual Search Task (2013). doi:10.1007/978-3-319-03680-9_11
11. Wu, Q., McGinnity, T.M., Maguire, L., Valderrama-Gonzalez, G.D., Dempster, P.: Colour Image Segmentation Based on a Spiking Neural Network Model Inspired by the Visual System (2010). doi:10.1007/978-3-642-14922-1_7
12. Bartolucci, M., Smith, A.T.: Attentional modulation in visual cortex is modified during perceptual learning. Neuropsychological **49**, 3898–3907 (2011)

13. Braddock, O., Atkinson, J.: Development of human visual function. Vision Res. **51**, 1588–1609 (2011)
14. Grossberg, S., Mingolla, E., Ross, W.: Visual brain and visual perception: How does the cortex do perceptual grouping? Trends Neurosci. **20**, 106–111 (1997)
15. Li, Z.: A neural model of contour integration in the primary visual cortex. Neural Neural Comput. **10**, 903–940 (1998)
16. Knierim, J., van Essen, C.: Neuronal responses to static texture patterns in area VI of the alertmacaque monkeys. J. Neurophysiol **67**(4), 961–980 (1992). Comput. **10**, 903–940
17. Grigorescu, C., Petkov, N., Westerberg, M.A.: Contour extraction based on non-classical receptive field inhibition. IEEE Trans. Image Process. **12**, 729–739 (2003)
18. Grislier, W.S., Perry, J.S., Super, B.J., et al.: Edge co-occurrence in natural images predicts contours grouping performance. Vision Res. **41**, 711–724 (2001)
19. Fukushima, K.: Visual physiology and bionics. Science Press 58–203 (1980)
20. Wang, Y., Qi, X.: Biophysics newspaper, No. 2, 585 (1980)
21. Hubel, D.G., Wiesel, T.N.: J. Pjysiol (London) **160**, 106–154 (1962)
22. Cooper, G.F., Robson, J.G.: IEE-NPL Conference on Pattern Recognition. IEE Conf. Publ. (London) **42**, 134–143 (1968)
23. You, L., Xie, L., Dong, T.: Visual receptive fields of neurons Mathematical model and spatial frequency response characteristics. Biophysics Newspaper **7**(No 2) (1991)
24. Allman, J., Miezin, F., McGuinness, E.: Direction and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). Perception **14**, 105–126 (1985)
25. Adesnik, H., Bruns, W., Taniguchi, H., Huang, Z.J., Scanziani, M.: A neural circuit for spatial summation in visual cortex (2012). doi:10.1038/nature11526
26. Wilmore, B.D.B., Bestrode, H., Tolhurst, D.J.: Contrast normalization contributes to a biologically-plausible model of receptive-field development in primary visual cortex (V1). Vision Res. **54**, 49–60 (2012)
27. Nelson: Orientation-selective in habitation from beyond the classic visual receptive field. Brain Res. **139**, 359–365 (1978)
28. Canny, J.F.: A computational approach to edge detection. IEEE Trans. Pattern Anal. **8**, 679–698 (1986)

A Sparse Pyramid Pooling Strategy

Lu Wang, Shengrong Gong[✉], Chunping Liu, Yi Ji, and Mengye Song

School of Computer Science &Technology, Soochow University, Suzhou 215006, China
shrgong@suda.edu.cn

Abstract. In this paper, we introduce a more principled pooling strategy for the Convolutional Restricted Boltzmann Machine. In order to solve the information loss problem of pooling operation and inspired by the idea of spatial pyramid, we replace the probabilistic max-pooling with our sparse pyramid pooling, which produces outputs of different sizes for different pyramid levels. And then we use sparse coding method to aggregate the multi-level feature maps. The experimental results on KTH action dataset and Maryland dynamic scenes dataset show that the sparse pyramid pooling achieves superior performance to the conventional probabilistic max-pooling. In addition, our pooling strategy can effectively improve the performance of deep neural network on video classification.

Keywords: Probabilistic max-pooling · Spatial pyramid pooling · Sparse coding · Deep neural network

1 Introduction

Deep learning emerges as a new area of Machine Learning research, and has achieved significant success in many artificial intelligence applications, such as speech recognition [1] and image processing [2]. In recent years, as more and more high-tech companies invest their resources for the development of deep learning, new architectures or algorithms may appear every few weeks. As a branch of machine learning, Deep learning refers to these feature learning methods that use unsupervised or/and supervised strategies to learn abstract feature representations in each layer of deep architectures, with the layers forming a hierarchy from low-level to high-level features [3,4,5]. In order to good learn feature representations of data, deep neural network focuses on end-to-end feature learning based on raw inputs regardless of label information in training and can compactly represent complex functions with the number of hidden units that is polynomial in the number of inputs.

One important branch in the field of deep neural network, convolutional neural network (CNN) uses a combination of supervised and unsupervised method to learn multiple stages of invariant features. Each stage of the CNN includes convolution layer and pooling/subsampling layer. In the convolution step, the same feature is applied to different locations for the stationary property of natural images. In other words, the convolutional layer extracts the common patterns in local regions of the inputs. In the pooling step, responses over nearby locations are summarized to make the representation invariance to small spatial shifts and geometric distortions.

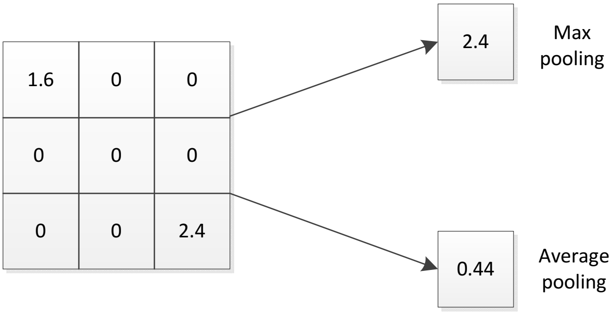
H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 406–416, 2015.
DOI: 10.1007/978-3-662-48570-5_39

The Convolutional Restricted Boltzmann machine (CRBM) is very similar to the stage of conventional convolutional network in terms of its structure. CRBM can be trained in an unsupervised way similar to that for the Restricted Boltzmann machine. In this work, a novel type of pooling strategy, called sparse pyramid pooling, is proposed for CRBM in order to increase scale invariance and reduce the risk of over-fitting. Inspired by the idea of spatial pyramid, we introduce the sparse pyramid pooling using different subsampling ratios for the pooling layer which partitions the feature maps from the convolutional layer into divisions from finer to coarser levels. The finer level extracts local precise details and the coarser level extracts global structures. Then sparse coding is used to aggregate these features in different scales. Finally the proposed method is demonstrated in the application of video classification, such as KTH action dataset and Maryland dynamic scenes dataset.

The remainder of the paper is organized as fellows. Section 2 introduces the conventional pooling strategy, including max pooling, average pooling and the probabilistic max-pooling. Section 3 gives the details of our sparse pyramid pooling strategy, including spatial pyramid pooling and sparse coding. Section 4 reports the experimental results on the KTH action dataset and the Maryland dynamic scenes dataset. Finally, section 5 concludes the paper and gives some prospective of this work.

2 Conventional Pooling Strategy

Currently, there are two conventional pooling strategies, including max pooling and average pooling. The max pooling strategy selects the maximum element in the pooling region, and the average pooling strategy takes the average of the element in the pooling region, as shown in Figure 1. The max pooling ignores the other elements in the pooling region that makes it easy to overfit for the training set and cannot generalize to the testing set well. The average pooling considers all elements in the pooling region including the low magnitudes, which reduces the contrast of the feature maps after pooling [6]. The max pooling is robust to background variability to some degree while average pooling is robust to intrinsic foreground variability [7].



1.6	0	0
0	0	0
0	0	2.4

Fig. 1. Max Pooling and Average Pooling in the 3×3 region

These conventional pooling strategies can only be used for feed-forward neural network, but CRBM is a generative model which supports both top-down and bottom-up inference. Therefore, Lee et al. [8] designed a probabilistic max-pooling for CRBM. The pooling unit is on if and only if one of the hidden units in the pooling region is on. As shown in Figure 2, the CRBM consists of three layers: a visible layer, a hidden layer and a pooling layer. Both the hidden layer and the pooling layer have K groups of units. For each group $k \in \{1,2,\cdots,K\}$, the units of pooling layer summarize $C \times C$ region in the hidden layer, here C is a small integer. The hidden layer is partitioned into $C \times C$ blocks, and each block is connected to one binary unit in the pooling layer. Different $C \times C$ blocks are defined in the hidden layer as B_α. The pooling region B_α in the hidden layer and the pooling unit p_α in the pooling layer follow these constraints: at most one unit of B_α in the hidden layer is on, and only if a unit of B_α is on, the unit p_α in the pooling layer is on. This pooling strategy is similar to max pooling, but the response unit in pooling layer is selected by probabilistic inference. It makes it possible to use top-down and bottom-up inference for the model. Assume that the bottom-up information from visible layer is expressed as $I_i\left(i=1,2,\cdots,C \times C\right)$, and the top-down information from another hidden layer is expressed as T, X_j is the unit in the pooling region of hidden layer, Y is the pooling unit, then the conditional probability is given by:

$$
\begin{aligned}
P(\mathrm{X}_j = 1 | V, H') &= \frac{\exp(\mathrm{T} + I_j)}{1 + \sum_i \exp(\mathrm{T} + I_i)} \\
P(\mathrm{Y} = 1 | V, H') &= \frac{\sum_i \exp(\mathrm{T} + I_i)}{1 + \sum_i \exp(\mathrm{T} + I_i)}
\end{aligned}
\tag{1}
$$

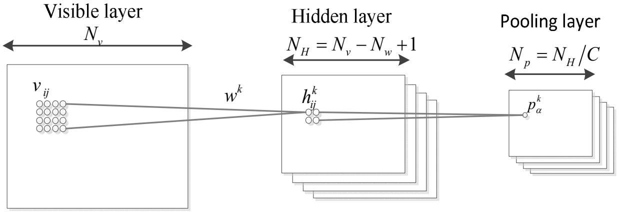


Fig. 2. Convolutional Restricted Boltzmann Machine, here filter number is 4 and pooling ratio is 2.

3 Sparse Pyramid Pooling

Pooling strategy is to aggregate statistics of the convolutional feature maps at various locations. The outputs of the conventional pooling strategies are deterministic by a fixed pooling ratio, and the pooling outputs have a problem of information loss. Inspired by the idea of spatial pyramid, we expand the outputs of the pooling layer for multi-level pyramid outputs and use sparse coding method to aggregate these outputs in different scales. The multi-level pooling provides different size pooling regions for the convolution layer and has been shown to be robust to object deformations [9, 10]. Sparse coding has similar properties to biological neurons and could reduce the dimension of multi-level pyramid outputs. Figure 3 shows a three-level sparse pyramid pooling layer, the sizes for different levels are $16\times16, 8\times8, 4\times4$ respectively. For each level of the pyramid, we use the probabilistic max-pooling strategy to summarize the convolution outputs.

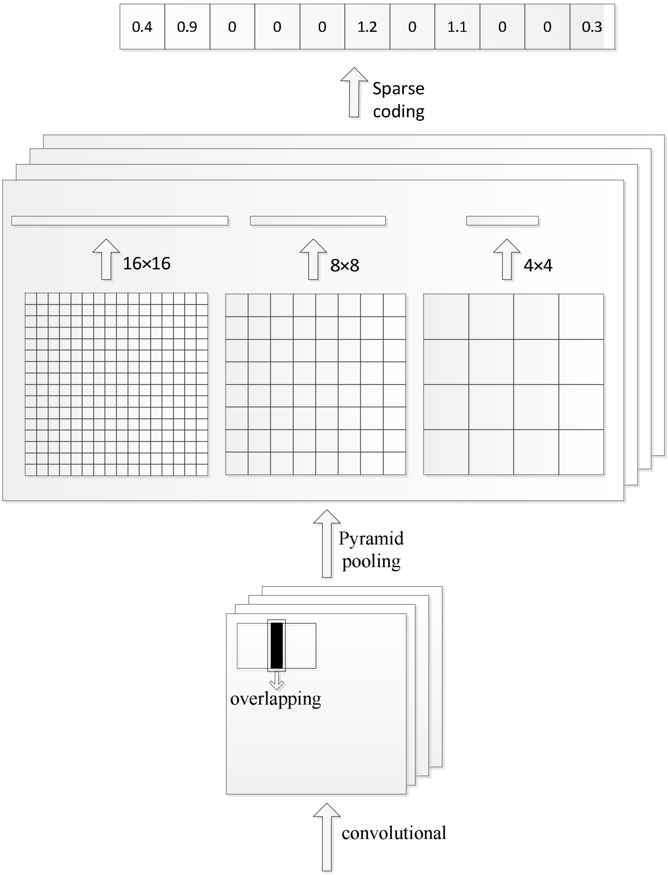


Fig. 3. Sparse pyramid pooling layer

3.1 Spatial Pyramid Pooling

The idea of pyramid mainly comes from Spatial Pyramid Matching (SPM) model [9], which is an extension of the Bag-of-Words (BoW) model. In order to solve the lack of location information in the BoW model, SPM partitions the image into divisions from finer to coarser levels, and aggregates features in different scales for image matching. Similarly, Spatial pyramid pooling uses different subsampling ratios for the pooling layer which partitions the feature maps from the convolutional layer into divisions from finer to coarser levels. The finer level extracts local precise details and the coarser level extracts global structures.

Here, the pooling strategy for each level of the pyramid is still the probabilistic max-pooling. The parameters in pooling layer can be pre-computed according to the size of inputs. Consider the size of feature maps after convolution layer is $a \times a$. The size of one level in the pyramid is $n \times n$, then the pooling ratio implemented is $\lceil a/n \rceil$ and the stride is $\lfloor a/n \rfloor$ with $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denoting ceiling and floor operations. We find that the adjacent pooling regions may be overlapping when the formula $\lfloor a/n \rfloor < \lceil a/n \rceil$ workable. More specifically, the pooling strategy can be viewed as a grid of pooling units stepped by s pixels, which summarizes a neighborhood of size $z \times z$ centered at the location of the pooling unit. If we set $s = z$, the pooling strategy is just the conventional pooling strategy. If we set $s < z$, the pooling strategy is overlapping pooling. In [2], the overlapping pooling strategy has been proved that it can improve the performance of deep neural network and slightly reduce the risk of overfitting. In this paper, we use the overlapping strategy in the pyramid pooling.

3.2 Sparse Coding

Sparse Coding provides a class of unsupervised methods to learn a set of sparse features for representing the data efficiently. It makes the feature representation more simple and explicit. The goal of sparse coding is to find a linear combination W , also called dictionary learning, to represent the sparse vector h of the input vectors x [11]. The basis vectors W should be overcomplete so that it can capture structures and patterns inherent in the input data. The overcomplete basis vectors mean the number of bases is greater than the input dimension. In order to favor the sparse coefficients, sparse coding will add a constraint, such as L_1 regularization, which makes only few values of the sparse vector far from zero. Sparse coding provides a method for learning sets of overcomplete basis vectors automatically to represent data efficiently. The cost function is defined as follows:

$$\begin{aligned}&\text{minimize}_{W,h} \sum_i \left\| x^{(i)} - Wh^{(i)} \right\|_2^2 \quad + \lambda \left\| h \right\|_1 \\ &\text{subject to} \quad \sum_j W_{i,j}^2 \le C\end{aligned} \tag{2}$$

Here, the penalty function is just L_1 regularization. We also constrain the basis vectors W to be less than some constraint C to prevent the sparsity penalty arbitrarily small.

The optimization problem with W and h is convex, but not convex in both simultaneously. We can perform two separate optimizations by holding the other fixed, the first optimizing over sparse vectors for training inputs and the second optimizing over basis vectors across training inputs. Despite the sparse coding model can find succinct and efficient representations for inputs, which has a limitation that an extra optimization must be performed to obtain the sparse vectors of a new data even after the basis vectors have been learnt. This means that sparse coding is computationally expensive and how to speed up is still a problem to be solved. In this paper, we exploit the efficient sparse coding algorithms [12] to speed up the calculation of basis vectors and sparse vectors, which propose the feature-sign search algorithm to solve the optimization problem of sparse vectors and develop the Lagrange dual for learning basis vectors. This method makes the encoding of big data possible.

4 Experiments

In this paper, we train a two layer time-space deep belief network (TS-DBN) which firstly use CRBM learn the temporal information and then learn the spatio information for video classification. We improve the pooling strategy of spatial CRBM with our sparse pyramid pooling. This strategy has been evaluated on the KTH action dataset and the Maryland dynamic scenes dataset.

The KTH action dataset consists of 6 types actions (boxing, handclapping, handwaving, jogging, running and walking), performed by 25 people in 4 different backgrounds. These videos capture variations in scale, illumination and action execution. We down sample the videos to a spatial resolution of 80×80 pixels each, and preserve the video length of 100. Videos from 9 subjects (subjects 2, 3, 5, 6, 7, 8, 9, 10 and 22) were chosen for test set, and the remaining 16 subjects were divided evenly into training and validation sets. We used a nonlinear SVM with a RBF kernel, and the parameters of the kernel were set using 5-fold cross-validation on the combined training and validation set. The Maryland dynamic scenes dataset contains 13 dynamic scene classes (avalanche, boiling water, chaotic traffic, forest fire, fountain, iceberg collapse, landslide, smooth traffic, tornado, volcanic eruption, waterfall, waves and whirlpool). These videos capture large variations in illumination, frame rate, viewpoint, image scale and various degrees of camera-induced motion. We down sample the videos to a spatial resolution of 160×120 pixels each, and preserve the video length of 190. The dynamic scene classification accuracy is averaged over the results of 10-fold cross-validation.

4.1 The Performance of Pooling Strategy

Table 1 compares the classification performance for three pooling strategies on KTH and Maryland dataset. The sparse pyramid pooling strategy increases the accuracy of classification by 2.4% on KTH dataset and 1.6% on Maryland dataset compared to the conventional probabilistic max-pooling strategy using the same architecture. In the KTH dataset, our two layer TS-DBN model with multi-scale inputs and sparse pyramid pooling strategy is superior to the three layer ST-DBN model with probabilistic max-pooling. In contrast to improve the depth of the network, our pooling strategy can also improve the performance of neural network.

Table 1. Classification performance for various pooling strategies on KTH and Maryland dataset

Model	Accuracy (%)	
	KTH	Maryland
Two layer ST-DBN + probabilistic max-pooling	84.6	-
Three layer ST-DBN + probabilistic max-pooling	86.6	-
Multi-scale inputs + two layer TS-DBN + probabilistic max-pooling	87.0	43.8
Multi-scale inputs + two layer TS-DBN + pyramid pooling	87.5	44.6
Multi-scale inputs + two layer TS-DBN + sparse pyramid pooling	89.4	45.4

4.2 Video Classification Results

Table 2 shows the detailed classification accuracy results on KTH action dataset. Compared with the hand-crafted features such as pLSA [13], ESURF [14] and LTP [15], our two layer TS-DBN model can achieve comparable accuracy. And our model increases the accuracy by 2.8% as compared to the conventional three layer ST-DBN model [16]. Although our average classification accuracy is slightly lower than the 3D CNN model [17], the results on some action types (Boxing, Jogging and Walking) are superior.

Table 2. Average classification accuracy results on the KTH action dataset.

Action classes	pLSA [13]	ESURF [14]	LTP [15]	ST-DB [16]	3D CNN [17]	Ours
Box	98	77.8	98	-	90	100
Clap	86	86.1	95	-	94	94
Wave	93	100	96	-	97	89
Jog	53	77.8	76	-	84	86
Run	88	72.2	86	-	79	86
Walk	82	91.7	90	-	97	81
Avg.(%)	83.3	84.26	90.1	86.6	90.2	89.4

Table 3. Average classification accuracy results on the Maryland dynamic scene dataset.

Scene classes	GIST [18]	HOF [19]	Chaos [20]	SOE [21]	Chaos+ GIST+ Color [21]	Ours
a.va	10	0	30	10	40	10
b.w	60	40	30	60	40	40
c.t	70	20	50	80	70	60
f.f	10	0	30	40	40	20
fnt	30	10	20	10	70	40
i.c	10	10	10	20	50	50
Ls	20	20	10	50	50	40
s.t	40	30	20	60	50	30
torn	40	0	60	60	90	80
v.e	30	0	70	10	50	70
Wf	50	20	30	10	10	30
Wv	80	40	80	80	90	80
Wp	40	30	30	40	40	40
Avg.(%)	38	17	36	41	52	45.4

As shown in table 3, our classification result on Maryland dataset outperforms the recently spatial temporal filters (SOE) proposed in [21], and many other features such as GIST [18], HOF [19] and Chaos [20]. The best performance is obtained by fusing the chaotic invariants, GIST and Color. Compared to our model, we only use the luminance information of the input videos and learn the spatial temporal features in a completely automated way. Most important of all, our model is able to perform other tasks just like many other deep neural networks.

4.3 Visualizations

Fig. 4 shows the three-level pyramid feature map activations for six actions of one person. The feature map activations become more blurred with the size being smaller, which is not suitable for video classification. But sparse coding could fine aggregate these outputs. When the size of feature maps is appropriate, the features learned from our model can distinguish different categories well.

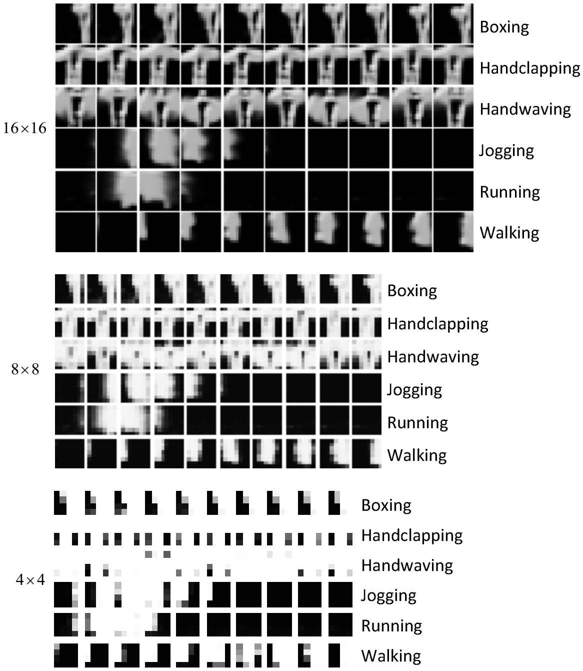


Fig. 4. Visualizations from feature map activations of the three-level pyramid pooling layer for six actions of one person

5 Conclusions

In this paper, we introduce a new sparse pyramid pooling strategy for CRBM. This method combines space pyramid with the probabilistic max-pooling. Sparse coding method is then used to aggregate the feature maps in different levels of the pyramid. Comparing to the conventional probabilistic max-pooling, our pooling strategy improves the performance of deep neural network on video classification. To extend this usage of this idea, further research will focus on the improvement of our pooling strategy on other deep neural networks.

Acknowledgements. This work is supported by National Natural Science Foundation of China (NSFC Grant No. 61170124, 61272258, 6130129), and a prospective joint research projects from joint innovation and research foundation of Jiangsu Province (BY2014059-14).

References

1. Dahl, G.E., Yu, D., Deng, L., et al.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing **20**(1), 30–42 (2012)
2. Krizhevsky, A., Sutskever, I., Geoff, H.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1106–1114 (2012)
3. Bengio, Y.: Learning deep architectures for AI. Foundations and trends® in Machine Learning **2**(1), 1–127 (2009)
4. Bengio, Y., Courville, A., Vincent, P.: Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013)
5. Bengio, Y.: Deep learning of representations: looking forward. In: Dediu, A.-H., Martín-Vide, C., Mitkov, R., Truthe, B. (eds.) SLSP 2013. LNCS, vol. 7978, pp. 1–37. Springer, Heidelberg (2013)
6. Yu, D., Wang, H., Chen, P., Wei, Z.: Mixed pooling for convolutional neural networks. In: Miao, D., Pedrycz, W., Slezak, D., Peters, G., Hu, Q., Wang, R. (eds.) RSKT 2014. LNCS, vol. 8818, pp. 364–375. Springer, Heidelberg (2014)
7. Boureau, Y.L., Bach, F., LeCun, Y., et al.: Learning mid-level features for recognition. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2559–2566. IEEE (2010)
8. Lee, H., Grosse, R., Ranganath, R., et al.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 609–616. ACM (2009)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE (2006)
10. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part III. LNCS, vol. 8691, pp. 346–361. Springer, Heidelberg (2014)
11. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell 1$ minimization. Proceedings of the National Academy of Sciences **100**(5), 2197–2202 (2003)
12. Lee, H., Battle, A., Raina, R., et al.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems, pp. 801–808 (2006)
13. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**(3), 299–318 (2008)
14. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
15. Yeffet, L., Wolf, L.: Local Trinary Patterns for Human Action Recognition. IEEE International Conference on Computer Vision **30**(2), 492–497 (2009)

16. Chen, B., Ting, J.-A., Marlin, B., et al.: Deep learning of invariant spatio-temporal features from video. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2010)
17. Ji, S., Xu, W., Yang, M., et al.: 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 221–231 (2013)
18. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision **42**(3), 145–175 (2001)
19. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2929–2936 (2009)
20. Shroff, N., Turaga, P., Chellappa, R.: Moving vistas: exploiting motion for describing scenes. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1911–1918 (2010)
21. Derpanis, K.G., Lecce, M., Daniilidis, K., et al.: Dynamic scene understanding: the role of orientation features in space and time in scene classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1306–1313. IEEE (2012)

A New Research on Contrast Sensitivity Function Based on Three-Dimensional Space

Jiachen Yang[1], Yun Liu[1,2](✉), Wei Wei[3], Qinggang Meng[4],
Zhiqun Gao[1], and Yancong Lin[1]

[1] School of Electronic Information Engineering, Tianjin University,
Tianjin 300072, China
{yangjiachen,yunliu}@tju.edu.cn, {gaozhiqunzz,linyc551}@163.com

[2] School of Optometry, University of California, Berkeley, Berkeley, CA, USA
yunliusally@berkeley.edu

[3] School of Computer Science and Engineering, Xi'an University of Technology,
Xi'an 710048, Shanxi, China
weiwei@xaut.edu.cn

[4] Department of Computer Science, School of Science at Loughborough University,
Loughborough LE11 3TU, UK
q.meng@lboro.ac.uk

Abstract. In this paper, we try to extend human eyes' contrast sensitivities characteristics (CSF) to three-dimensional space, but the experimental results show that the traditional characteristics of CSF has limitations in three-dimensional space. In order to investigate the characteristics of human eyes' CSF in three-dimensional space, the traditional CSF test method is developed to measure the corresponding values of CSF in different inclined planes in three-dimensional space. Human visual contrast sensitivity characteristics with different inclined angles θ are analyzed, and the mathematical expression of $\theta - CSF$ is built up based on the experimental results. The proposed $\theta - CSF$ model of three-dimensional space in this paper can well reflects human visual contrast sensitivity characteristics in 3D space and has significant effect on three-dimensional image processing.

Keywords: Human visual system · Contrast sensitivity function · Spatial frequency · Three-dimensional space

1 Introduction

With the development of science and technology, information processing technology is becoming more mature [1]-[4]. In order to really reflect what people seen in the natural scene, human visual system (HVS) is often incorporated into the technology of image processing [5],[6]. The most important one among human visual characteristics in HVS is the contrast sensitivity characteristic [7],[8] which has been widely used in the area of image processing in plane space [9],[10]. Chen *et al.* [11] proposed a perceptual quality evaluation method for

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 417–428, 2015.
DOI: 10.1007/978-3-662-48570-5_40

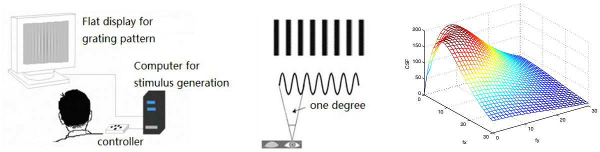
image fusion based on CSF which is focus on the night vision application; Tao *et al.* [12] developed a novel reduced-reference image quality assessment scheme by incorporating CSF and the objective assessment results, which can well reflect the visual quality of images. Some effective quality assessment methods also effectively built by incorporating CSF characteristics and wavelet transforms of image, such as Gao [13] and Li [14], which can well reflect human visual perception. Besides, Zhang [15], Wu [16] and M*ü*ller [17] *et al.* applied CSF characteristic to the technology of image processing and made great achievements. Urvoy [18] and Tsai [19] *et al.* proposed an effective perceptual watermarking technique based on CSF and achieved good robustness against the common operations.

The characteristic of human eyes' traditional CSF (in this paper CSF proposed by the predecessors called traditional CSF) is built based on the grating test system, in which the monitor paralleled to the viewer's face in two-dimensional (2D) space without considering other inclined planes that not paralleled to human face. The traditional CSF is just aim at the plane display technologies, such as 2DTV and 2D movie. In fact, what people see in the real world are not all displayed on the plane paralleled to human face. With the development of three-dimensional image processing technology and display technology [20],[21], more and more 3D displays are widely used, such as 3DTV, holographic display and so on. Whether traditional CSF suits for three-dimensional space has not been verified, and limited its application in three-dimensional space. It is very necessary to study human visual contrast sensitivity characteristics in 3D space.

The rest of this paper is organized as follows. In Section 2, it describes the theory of the traditional CSF. Section 3 the extension of traditional CSF in three-dimensional space is illustrated. And the comparison between experiment results and the results of the extension of traditional CSF is showed in Section 4. In Section 5, the model of CSF in three-dimensional space is built. Finally, conclusion and future work are given in Section 6.

2 An Overview of Traditional CSF

Human eyes have different visual characteristics in different frequency bands, i.e., we are unable to recognize a stimuli pattern if its frequency of visual stimuli is too high. For example, given an image consisting of horizontal black and white stripes, we will perceive it as a gray image if stripes are very thin; otherwise, we can distinguish these stripes. Based on the visual characteristics, traditional contrast sensitivity function (CSF) has been proposed which measures how sensitive we are to the various spatial frequencies of visual stimuli. Now the value of traditional CSF can be measured by grating contrast sensitivity test system (shows in Fig.1.(a)) which has been used to study the physiology of the visual system [22] for some time and increasingly used to study ophthalmology [23]. Unlike the Snellen letter acuity test [24], which establishes visual acuity in terms of the smallest recognizable object presented at 100% contrast, the grating test allows the specification of an observer's sensitivity to larger targets of lower contrast, and sensitivity is defined as the reciprocal of the contrast threshold. Mannos and


Flat display for grating pattern
Computer for stimulus generation
controller
one degree
CSF
200
150
100
50
0
fx
fy
10
20
30

(a) (b) (c)

Fig. 1. (a)Traditional CSF test system: the observers view the gratings monocular with the fellow eye occluded (b) The concept of spatial frequency:two circles per degree. (c)2D CSF characteristics surface

Sakrison [25], after conducting a series of psychophysical experiments on human subjects, found that CSF can be modeled by the function in Eq.(1).

$$A(f) = (0.0499 + 0.2964f)e^{-(0.114f)^{1.1}} \tag{1}$$

where f is the spatial frequency which means the number of cycles per degree subtended at the eye (shows in Fig.1.(b)), with unit of cycles/degree. Eq.(1) reveals that the values of traditional CSF are related with the circles of grating in human eyes (that is f).

Generally in image processing area, human visual system has the same contrast sensitivity in all directions of plane space, and the 2D version [26] can be easily obtained by replacing f with radial frequency $\sqrt{f_x^2 + f_y^2}$

$$A = (0.0499 + 0.2964\sqrt{f_x^2 + f_y^2})e^{-(0.114\sqrt{f_x^2+f_y^2})^{1.1}} \tag{2}$$

where f_x and f_y are the horizontal and vertical frequencies respectively, and they make no sense for frequencies above 30 cycles/degree. Fig. 1(c) shows the 2D CSF characteristics surface.

3 The Extension of Traditional CSF in Three-Dimensional Space

Traditional CSF is a nonlinear function of spatial frequency which is built based on the plane of two-dimensional space (such as plane 1 shows in Fig.2(a)). Here we try to extend it to three-dimensional space including many inclined planes such as plane 2 and plane 3. Because of the existence of inclined angles (θ_1 and θ_2 show in Fig.2(a)), the value of spatial frequency in the inclined plane will be changed. Based on the theory of traditional CSF, the values of CSF in different inclined planes can be get by applying the spatial frequency of each inclined plane to Eq.(1). To verify the practicability of traditional CSF in 3D space, subjective experiments, shown in Fig.2(b) (to get the inclined plane by rotating the test monitor), are conducted to study the relationship between calculated

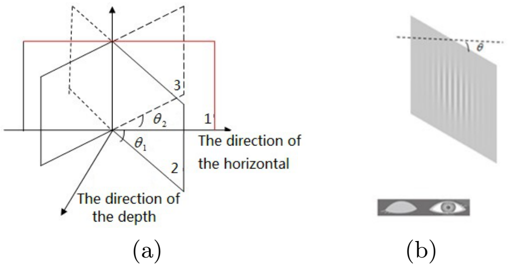


Fig. 2. (a) Geometric simulation figure of three-dimensional image (b) CSF test system of different inclined plane.

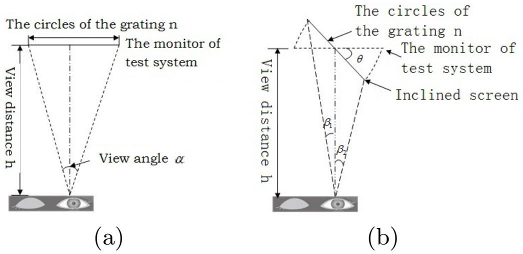

The circles of the grating n
The monitor of test system
View distance h
View angle α
(a)
The circles of the grating n
The monitor of test system
θ
Inclined screen
View distance h
β_1
β_2
(b)

Fig. 3. (a) Geometric simulation figure of test system of traditional CSF (b) Geometric simulation figure of test system of CSF in the inclined plane

CSF values based on Eq.(1) and the experimental data on different inclined planes, and investigate human eyes' characteristics of CSF in three-dimensional space.

According to the test theory of traditional CSF, the monitor of test system is set to paralleled to viewers' face, the observers viewed the gratings monocular with the fellow eye occluded and the view angle is α (shows in Fig.3(a)). Define the circles of grating in human eyes as n, and the viewing distance is h (the value is far greater than n). After rotating the monitor clockwise to form the inclined angle θ (shows in Fig.3(b)), the circles of grating is still n, because what the observer see is the same monitor and the circles of gratings are not change, while the view angle is changed to $\beta_1 + \beta_2$, so the spatial frequency f is changed. Based on the concept of the spatial frequency [27] shown in Eq.(3).

$$f(\frac{cycles}{degree}) = \frac{f_i}{\arcsin \frac{1}{\sqrt{[1+d^2(H^2+V^2)]}}} \tag{3}$$

where f_i is the image frequency (obtained from Fourier transform), the normalizing factor f_n is the number of pixels within 1 degree at the viewing distance of d times the diagonal image size, H is the horizontal image size and V is the vertical image size in pixels. Based on the geometrical relationships, we have:

$$f_u(\frac{cycles}{degree}) = \frac{n}{\alpha} \tag{4}$$

$$f_\theta(\frac{cycles}{degree}) = \frac{n}{\beta} \tag{5}$$

where $\beta = \beta_1 + \beta_2$. f_u, horizontal spatial frequency, is the spatial frequency of the plane with inclined angle 0°; f_θ, inclined spatial frequency, is the spatial frequency of the inclined plane with inclined angle θ; Because h is far greater than n, so the viewing angle is very small. According to the mathematical theory, when an angle is too small, the value of tangent of the angle and the value of angle are approximately equal. So:

$$\alpha = 2arctan\frac{n/2}{h} \approx 2 \times \frac{n/2}{h} = \frac{n}{h} \tag{6}$$

$$\begin{aligned}\beta &= arctan\frac{\frac{n}{2}cos\theta}{h+\frac{n}{2}sin\theta} + arctan\frac{\frac{n}{2}cos\theta}{h-\frac{n}{2}sin\theta} \\ &\approx \frac{\frac{n}{2}cos\theta}{h+\frac{n}{2}sin\theta} + \frac{\frac{n}{2}cos\theta}{h-\frac{n}{2}sin\theta}\end{aligned} \tag{7}$$

From the above analysis, the horizontal spatial frequency f_u and the inclined spatial frequency f_θ are obtained:

$$f_u = \frac{n}{\alpha} \approx h \tag{8}$$

$$f_\theta = \frac{n}{\beta} \approx \frac{n}{hncos\theta/(h^2 - \frac{n^2}{4}sin^2\theta)} \tag{9}$$

Because h is far greater than n, so :

$$h^2 - \frac{n^2 sin^2\theta}{4} \approx h^2 \tag{10}$$

$$f_\theta = \frac{n}{\beta} \approx \frac{h}{cos\theta} \tag{11}$$

Then the relationship between the horizontal spatial frequency f_u and the inclined spatial frequency f_θ is given by :

$$f_\theta = f_u/cos\theta \tag{12}$$

Besides, based on the geometric symmetry, it can be easily get the conclusion: when the monitor of the test system rotated counterclockwise (plane 3 in Fig.2(a)), the corresponding spatial frequency in these inclined directions and the horizontal spatial frequency are all meet the mathematical expression of Eq.(12). And together with the expression of traditional CSF and the replacement of f in Eq.(1) by $f_u/cos\theta$, the CSF's expressions of each inclined plane is expressed by:

$$A(f_\theta) = (0.0499 + 0.2964(f_u/cos\theta))e^{-(0.114(f_u/cos\theta))^{1.1}} \tag{13}$$

To verify the correctness of Eq.(13), the rest work in the paper will take the clockwise rotation direction as an example, and use the grating test system to test the values of contrast sensitivities in different inclined planes when observers view the monitor monocular with the fellow eye occluded.

4 The Comparison between Experiment Result and the Result of Extension of Traditional CSF

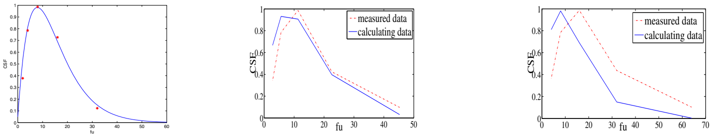
This work firstly tests the value of human eyes' traditional CSF values to verify the reliability of the experimental data and then rotated the monitor of the test system to form different inclined angles. The values of CSF in different inclined planes is measured and verified whether the above $A(f_\theta)$ meet human visual characteristics.

The gratings of the test system are electronically generated on the screen of a monochrome television monitor (Melford Electronics DU1/20, 625 lines, 50Hz, 2:1 interlaced, P4 phosphor). The experiments are performed under photopic conditions with the mean luminance of the gratings constant at $100cd/m^2$. The television screen is subtended $180 \times 15°$. To minimize fading at low spatial frequencies and the generation of after-images the gratings are continuously reversed at the rate of $1cycles/s$ and observers are instructed to fixate a small spot in the center of the screen. The observers view the gratings monocular with the fellow eye occluded. At each spatial frequency the contrast of the grating is reduced by means of a potentiometer until the observer indicate that the grating is just disappeared, then record the contrast value as the threshold of this spatial frequency.

We select 18 observers, age from 20 to 40 years, to conduct this experiment. Before the test, all subjects receive a complete clinical evaluation to ensure that their visual fields, color vision, and intraocular pressure within the normal range for the Clinical Branch of the National Eye Institute. All subjects receive a fundus evaluation by a staff ophthalmologist to ensure that there is no detectable ocular pathology affecting either retinal or preretinal (e.g., cataract) levels within the eye. In addition, each subject receive a complete refractive evaluation before testing. All subjects have best-corrected visual acuities of 20/20 or better in both eyes. Before the grating tests officially start, the observers take a period of time to adopt to the experiment process. Every observer conducts this experiments

Table 1. CSF test values of observers

f_u	2	4	8	16	32
CSF_0	0.3786	0.7849	0.9862	0.7270	0.1233
$CSF_{\frac{\pi}{4}}$	0.3434	0.7849	0.9862	0.4227	0.0971
$CSF_{\frac{\pi}{3}}$	0.3663	0.7849	0.9862	0.4371	0.1028



(a) (b) (c)

Fig. 4. (a) Comparison results between test result with the inclined angle 0° and traditional CSF curve. (b)The fitting curve with the inclined angle of $\frac{\pi}{4}$ (c)The fitting curve with the inclined angle of $\frac{\pi}{3}$

twice and the mean value is calculated as the final data. Then human eyes' contrast sensitivities test value is defined as the mean of 18 observers' independent decisions. The values of observers' CSF in the plane with inclined angle 0° is tested at the beginning of the project and the range of the spatial frequency of the test system is from $2circles/degree$ to $32circles/degree$. The result is shown in Table 1, the first row.

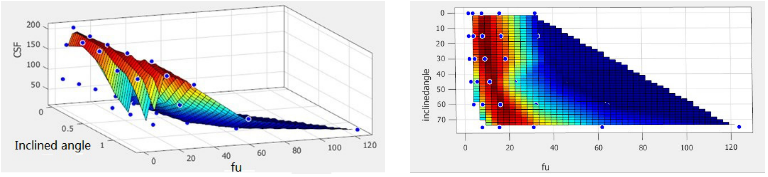
The test data is normalized within 1 and frequencies above $60cycles/degree$ is meaningless in reality. Compared with the traditional CSF curve, the test results, as shown in Fig.4(a), are accurate. Then two inclined angles, $\frac{\pi}{4}$ and $\frac{\pi}{3}$ are set to carry out the experiment, the test data shown in Table 1, the second and third rows. With the function of $A(f_\theta)$, the CSF values of the inclined planes with the inclined angles $\frac{\pi}{4}$ and $\frac{\pi}{3}$ are calculated. The test results and the calculated values are shown in Fig.4(b) and (c). Fig.4(b) and (c) indicate a great difference between experiment results and calculated values, which means that traditional CSF has limitations in three-dimensional space, and cannot be expanded to three-dimensional space directly.

5 The Proposal of CSF in Three-Dimensional Space

In order to build human eyes' CSF in different inclined planes, it is necessary to obtain data from more inclined angles and values of observers' contrast sensitivities, so $\frac{\pi}{12}$,$\frac{\pi}{6}$ and $\frac{5\pi}{12}$ are set to test. The results shown in Table 2, and fitted shown in Fig.5 (a) and (b) with the amplitude normalized within [0,200]. Using geometry relationship to analyze the test data, it reveals that the curve trends of different inclined angles are the same and all of them have the nature of band pass filter.

Table 2. Human eyes' test values of CSF with different inclined angles

f_u	2	4	8	16	32
$CSF_{\frac{\pi}{12}}$	0.4227	0.9185	0.9862	0.7849	0.1378
$CSF_{\frac{\pi}{6}}$	0.4634	0.9158	0.9862	0.7849	0.1233
$CSF_{\frac{5\pi}{12}}$	0.3434	0.9158	0.7849	0.3260	0.0986



(a) (b)

Fig. 5. (a)The fitting surface of the experiment data. (b) The vertical view of fitting surface.

Specifically, contrast sensitivities of human eyes are different in varied inclined planes; all CSF characteristic curves experience the same trend at different stereo-angles and all have the nature of band pass filter, but the position of the peaks are gradually moved back with the increased inclined angle; besides, the descending velocity of the curve at the high spatial frequency gradually decreased with the increased inclined angle; what's more, the CSF curves in each inclined plane are all like the traditional CSF curve, so it is possible for us to build the θ-CSF's expression of each inclined plane $A(f_\theta, \theta)$ based on the traditional CSF's expression, as follows:

$$A(f_\theta, \theta) = (a + bf_\theta)e^{-(cf_\theta)^d} \tag{14}$$

where f_θ is spatial frequency of the inclined plane with inclined angle θ.

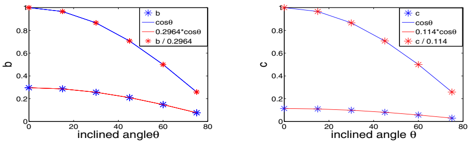
Based on the experimental results, we apply MATLAB fitting tool to obtain the specific value of each parameter, shown in Table 3. Each equation is well fitted with fitting coefficient $R - square > 0.9$, while the values of RMSE are less than 0.4. Table 3 indicates that among all parameters only b and c are changed, and other coefficients are the same as those in traditional CSF under different inclined angle θ. In order to get a general $\theta - CSF$ model, the relationship between b and θ, and c and θ are fitted, shown in Fig.6, which indicate a cosine relationship between b (c) and the inclined angle θ. The good linear correlation with the Pearson correlation coefficients $r > 0.9$ indicates the goodness of the fit. Then the expression of $\theta - CSF$ based on the inclined angles θ in three-dimensional space is defined as Eq.(15):

$$A(f, \theta) = (0.0499 + 0.2964f \times cos\theta)e^{-(0.114f \times cos\theta)^{1.1}} \tag{15}$$

Here the inclined angles are in the unit of radians.

Table 3. Parameters of each inclined plane

$A(f_\theta, \theta)$	a	b	c	d
$A(f_0, 0)$	0.0499	0.2964	0.114	1.1
$A(f_{\frac{\pi}{12}}, \frac{\pi}{12})$	0.0499	0.2863	0.1101	1.1
$A(f_{\frac{\pi}{6}}, \frac{\pi}{6})$	0.0499	0.2567	0.0987	1.1
$A(f_{\frac{\pi}{4}}, \frac{\pi}{4})$	0.0499	0.2096	0.0806	1.1
$A(f_{\frac{\pi}{3}}, \frac{\pi}{3})$	0.0499	0.1482	0.0570	1.1
$A(f_{\frac{5\pi}{12}}, \frac{5\pi}{12})$	0.0499	0.0767	0.0295	1.1



(a) (b)

Fig. 6. (a)The fitted model of b. (b) The fitted model of c.

We quantifies the goodness of the fits with a Q measure given in the figures. Q is a χ^2 distribution function, and Q of 0.1 suggests an acceptable model fit [28]. Each sub-figure provides a very good fit for $Q > 0.1$, which indicates that this model is able to fit our psychophysical data. The surface of the $\theta - CSF$ is illustrated in Fig.8(a), it reveals that the data are agree with the above test values in Fig.5(a). Besides, the CSF curve with the inclined angle 0° are the same as the traditional CSF curve and the vertical view shown in Fig.8(b) coincide with that in Fig.5(b). These results indicate that the proposed $\theta - CSF$ here is in according with human visual characteristics.

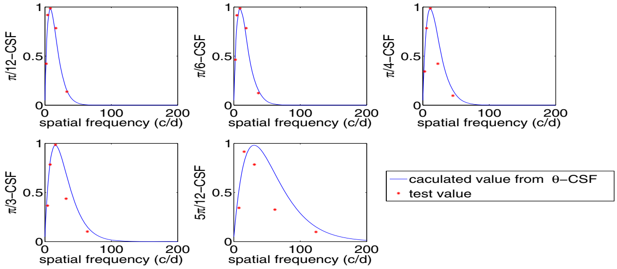
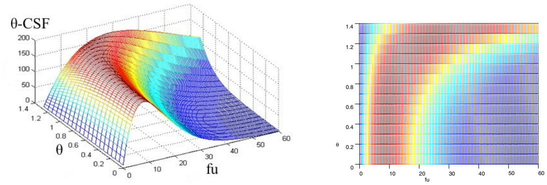


Fig. 7. Comparision results between test value and caculated value from proposed $\theta - CSF$



(a) (b)

Fig. 8. (a)$\theta - CSF$ characteristics surface in three-dimensional space. (b) The vertical view of $\theta - CSF$ characteristics surface.

6 Conclusion

This paper aims to apply the concept of the plane spatial frequency to analyze the relationship among different inclined plane, and to extend the traditional CSF to three-dimensional space. According to the experimental results and the geometric relationship between horizontal spatial frequency and the spatial frequency in the directions of inclined angles, $\theta - CSF$ characteristic surface of human eyes based on the inclined angle is built with a specific function expression. The proposed CSF characteristics in three-dimensional space is consistence with human visual characteristics.

Acknowledgments. This research is supported by the National Natural Science Foundation of China (No.61471260 and No.61271324), and Program for New Century Excellent Talents in University (NCET-12-0400).

References

1. Legg, P.A., Rosin, P.L., Marshall, D., Morgan, J.E.: Feature Neighbourhood Mutual Information for Multi-Modal Image Registration: An Application to Eye Fundus Imaging. Pattern Recognition **48**(6), 1937–1946 (2015)
2. Martinez, F., Carrasco, A., Salas, J., di Baja, G.S.: Pattern Recognition Application in Computer Vision and Image Analysis. Pattern Recognition **48**(4), 1025–1026 (2015)
3. Smith, S., Williams, I.: A Statistical Method for Improved 3D Surface Detection. IEEE Signal Processing Letters **22**(8), 1045–1049 (2015)
4. Wei, W., Qi, Y.: Information Potential Fields Navigation in Wireless Ad-Hoc Sensor Networks. Sensors **11**(5), 4794–4807 (2011)
5. Peng, R.B., Varshney, P.K.: A Human Visual System-Driven Image Segmentation Algorithm. Journal of Visual Communication and Image Representation **26**, 66–79 (2015)
6. Chang, H.W., Zhang, Q.W., Wu, Q.G., Gan, Y.: Perceptual Image Quality Assessment by Independent feature detector. Neurocomputing **151**(10), 1142–1152 (2015)

7. Rosén, R., Lundström, L., Venkataraman, A.P., et al.: Quick Contrast Sensitivity Measurements in the Periphery. Journal of Vision **14**(8) (2014)
8. Liu, R., Zhou, J.W., et al.: Immature Visual Neural System in Children Reflected by Contrast Sensitivity with Adaptive Optics Correction. Scientific Reports **4**(4687), April 2014
9. Wei, Z.Y., Ngan, K.N.: Spatio-Temporal Just Noticeable Distortion Profile for Grey Scale Image/Video in DCT Domain. IEEE Transactions on Circuits and Systems for Video Technology **19**(3), 337–346 (2009)
10. Brandã, T., Queluz, M.P.: No-Reference Quality Assessment of H.264/AVC Encoded Video. IEEE Transactions on Circuits and Systems for Video Technology **20**(11), 1437–1447 (2010)
11. Chen, Y., Blum, R.S.: A New Automated Quality Assessment Algorithm for Image Fusion. Image and Vision Computing **27**(2), September 2009
12. Tao, D., Li, X., Lu, W., Gao, X.: Reduced-Reference IQA in Contourlet Domain. IEEE Transaction on Systems Man and Cybernetics-Part B: Cybernatics **39**(6), December 2009
13. Gao, X., Lu, W., Tao, D., Li, X.: Image Quality Assessment Based on Multiscale Geometric Analysis. IEEE Transactions on Image Processing **18**(7), July 2009
14. Li, S., Zhang, F., Ma, L., Ngan, K.N.: Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments. IEEE Transactions on Multimedia **13**(5), October 2011
15. Zhang, F., Ma, L., Li, S., Ngan, K.N.: Practical Image Quality Metric Applied to Image Coding. IEEE Transactions on Multimedia **13**(4), August 2011
16. Wu, G.-L., Wu, T.-H., Chien, S.-Y.: Algorithm and Architecture Design of Perception Engine for Video Coding Applications. IEEE Transactions on Multimedia **13**(6), December 2011
17. Müller, K., Merkle, P., Wiegand, T.: 3-D Video Representation Using Depth Maps. Proceedings of the IEEE **99**(4), 643–656 (2011)
18. Urvoy, M., Goudia, D., Autrusseau, F.: Perceptual DFT Watermarking With Improved Detection and Robustness to Geometrical Distortions. IEEE Transactions on Information Forensics and Security **9**(7), 1108–1119 (2014)
19. Tsai, M.J., Liu, J., Yin, J.S., Yuadi, I.: A Visible Wavelet Watermarking Technique based on Exploiting the Contrast Sensitivity Function and Noise Reduction of Human Vision System. Multimedia Tools and Applicatios **72**(2), 1311–1340 (2014)
20. Boisvert, J., Drouin, M.A., Jodoin, P.M.: High-Speed Transition Patterns for Video Projection, 3D Reconstruction, and Copyright Protection. Pattern Recognition **48**(3), 720–731 (2015)
21. Elaiwat, S., Bennamoun, M., Boussaid, F., et al.: A Curve Let-based Approach for Textured 3D Face Recognition. Pattern Recognition **48**(4), 1235–1246 (2015)
22. Schade, S.R.: Optical and Photoelectric Analog of the Eye. Journal of the Optical Society of America **46**(9), 721–738 (1956)
23. Bodis-Wollner, I., Diamond, S.P.: The Measurement of Spatial Contrast Sensitivity in cases of Blurred Vision Associated with Cerebral Lesions. Journal of Neurology **99**, 695–710 (1976)
24. Mocan, M.C., Najera-Covarrubias, M., Wright, K.W.: Comparison of Visual Acuity Levels in Pediatric Patients with Amblyopia using Wright Figures((c)), Allen Optotypes, and Snellen Letters. Journal of Aapos **9**, 48–52 (2005)
25. Mannos, J.L., Sakrison, D.J.: The Effects of a Visual Fidelity Criterion on the Encoding of Images. IEEE Transactions on Information Theory **IT-20**(4), July 1974

26. Zeng, W., Daly, S., Lei, S.: An Overview of the Visual Optimization Tools in JPEG 2000. Signal Processing: Image Communication **17**(1), 85–104 (2002)
27. Gaddipatti, A., Machiraju, R., Yagel, R.: Steering Image Generation with Wavelet Based Perceptual Metric. Computer Graphics Forum **16**(3), C241–C251 (1997)
28. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C: The Art of Scientific Computing. Cambridge Univ. Press (1992)

An Algorithm Combined with Color Differential Models for License-Plate Location

Xiangdong Zhang[1], Peiyi Shen[2](✉), Juan Song[2], Liang Zhang[2], Weibin Gong[3], Wei Wei[4], Yuanmei Tian[1], and Guangming Zhu[2]

[1] School of Telecommunications Engineering, Xidian University, Xi'an 710071, People's Republic of China
[2] School of Software, Xidian University, Xi'an 710071, People's Republic of China
pyshen@xidian.edu.cn
[3] School of Information Engineering, Chang'an University, Xi'an 710062, People's Republic of China
[4] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China

Abstract. Vehicle license plate recognition technology is one of the core technologies of intelligent transportation systems. The first and most important step in the entire license plate recognition system is positioning the license plate. The positioning accuracy will directly influence the subsequent segmentation and recognition accuracy. This paper presents a new adaboost algorithm combined with color differential model. First, we introduce the process of calculation of the color differential model. Second, we give a full distribution about the adaboost algorithm combined with color differential model. At last, we analyze the results of the algorithm based on RGB color model and give a comparison between the adaboost algorithm combined with the new feature and other license plate location algorithms. This novel adaboost algorithm overcomes the problems of license plate location algorithms based on color information, such as the sensitivity to light and the difficulty to locate license plates in complex background and so on. The experimental results show that the novel adaboost algorithm combined with color differential model is timeliness and robustness. At night, the precision rate of the novel adaboost algorithm can attain above 95.0%.

Keywords: License plate location · Color differential model · Adaboost algorithm

1 Introduction

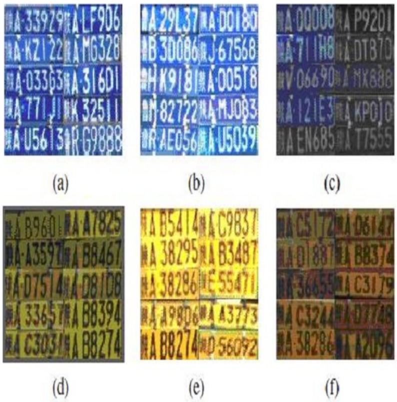
In recent years, theoretical and applied research of Intelligent Transportation (Intelligent Transportation Systems, referred to as ITS) set off a boom in various countries. License plate recognition technology as the key step to achieve intelligent traffic management has been developed rapidly. License plate recognition technology mainly contains three processes: license plate location, character segmentation and character recognition, among which license plate location is

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 429–438, 2015.
DOI: 10.1007/978-3-662-48570-5_41

the first and the most crucial step. With the growing attention to license plate location and the continuous deepening research, different new license plate location algorithms are constantly proposed. These algorithms can be classified into the following categories: license plate positioning algorithm based on texture information, license plate positioning algorithm based on HSI color space, the algorithm of license plates location based on mathematical morphology, license plate positioning algorithm based on genetic algorithm, the algorithm of license plates location based on neural network and etc. For the inadequacies of previous algorithms, a novel adaboost algorithm combined with color differential model for locating license plates is proposed [1,2]. First, we obtain the feature thresholds of RGB color model and introduce the LP positioning algorithm based on RGB color model. In this paper, feature thresholds are obtained under a series of different conditions by using statistical histogram training method [3,4]. Then we combine the original adaboost algorithm [5] with the new feature. The novel adaboost algorithm not only overcomes the problems of license plate location algorithms based on color information, such as the sensitivity to light and the difficulty to locate license plates in complex background and so on, but also is timeliness and robustness.

1.1 Calculation of Characteristic Value and Binarization of Color Images

Various image processing methods for color image processing [6] can be divided into three categories: (1) color transformation, that is, processing pixels of each color plane based strictly on their values, such as conversion of the RGB color model to the HIS color model; (2) spatially processing of the color plane, that is, doing spatial filtering in each color space; (3) process techniques which can deal with the three components of color image concurrently, such as the algorithm of region segmentation. The algorithm based on RGB color model is an application of the third method [11,12]. Our goal is to locate one or a few regions of specific color in an RGB image, which is the license plate region(s).Blue, yellow, white and black are four colors of license plate in China. The plates of black characters with yellow background are generally for busses and trucks; the plates of black characters with white background for military or police vehicles; special vehicles generally with the plates of white characters with black background; the rest of the vehicles usually with the plates of white characters with blue background. Blue background license plates are more common in China. For white characters with blue background license plates, when the light conditions are good and the license plate is relatively clear, R, G, B values of pixels are about at 150,30, 60. When following the light, each pixel value of R, G, B components will increase. When backlighting, each pixel of the R, G, B components values will reduce. The results for the different conditions are shown in Fig. 1.



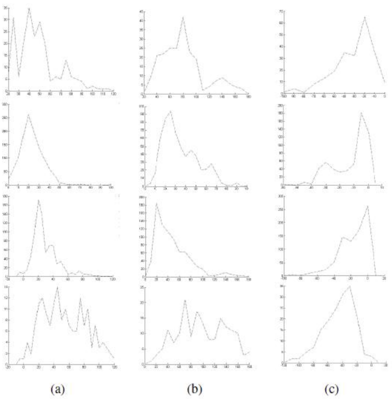


Fig. 1. (a)clear blue plates,(b)license plates following the light,(c)license plates with backlighting,(d)clear yellow background plates,(e)license plates following the light,(f)license plates with backlighting

Fig. 2. Column(a) statistical curve graph ofr_g,column(b) statistical curve graph of b_g,column(c) statistical curve graph of b_r

2 Algorithm Principles

Determine the Threshold. A statistical histogram method is used to determine the threshold of the same background color in different situations. The detail process is shown as follows: (1) Select three sample sets of different conditions respectively: bright license plate set(over exposed), dark license plate 3 set, and mixed set(normal exposed). The size of license plates in each sample set is 70 × 20. The amounts of license plates in dark license plate set, bright license plate set and the mixed set are 589, 230, 811 respectively. (2) We obtain characteristic values between the three color components of each pixel in each sample plate, which is the characteristic value of this algorithm.

$$r\_g_{i,j} = R_{i,j} - G_{i,j}, b\_g_{i,j} = B_{i,j} - G_{i,j}, b\_r_{i,j} = B_{i,j} - G_{i,j} \qquad (1)$$

where $R_{i,j}, G_{i,j}, B_{i,j}$ are R, G, B values of pixel (i,j) in the plate sample respectively and have a range is from 0 to 255.$r\_g_{i,j}, b\_g_{i,j}, b\_r_{i,j}$ are the characteristic values between three components of each pixel. (3) We calculate the sum of each characteristic value of each pixel in each plate sample using the following equations:

$$sum\_r\_g = \sum_{i=0}^{69}\sum_{j=0}^{19} r\_g_{i,j}, sum\_b\_g = \sum_{i=0}^{69}\sum_{j=0}^{19} b\_g_{i,j}, sum\_b\_r = \sum_{i=0}^{69}\sum_{j=0}^{19} b\_r_{i,j} \qquad (2)$$

where $sum\_r\_g, sum\_b\_g, sum\_b\_r$ are sums of characteristic values of all pixels in a plate sample.

(4) We calculate the average characteristic value of each sample plate using the following equations:

$$r\_g = \frac{sum\_r\_g}{70\times 20}, b\_g = \frac{sum\_b\_g}{70\times 20}, b\_r = \frac{sum\_b\_r}{70\times 20} \qquad (3)$$

Where r_g, b_g and b_r are mean characteristic values of all the pixels in one plate sample.
(5) The statistical curve graph of r_g, b_g and b_r can be got next, as shown in Fig. 2: the first row is the statistical curve graph of bright license plate set,the second row is the statistical curve graph of dark license plate set,the third row is the statistical curve graph of mixed set,the fourth row is the statistical curve graph of license plate set at night.Abscissa shows characteristic values between each two of the three components in RGB space, ordinate corresponds to the frequency of different characteristic values. Comparing images containing license plate(s) during the day and night, we found that there is less background interference in the picture at night, and the characteristic values between R, G, B components of each pixel at night are different from that during the day time. So we obtained the statistical curve graph of the sample set, in which include 168 pictures of license plates with the size 70×20 at night, as shown in the fourth line of Fig. 2.
(6) In the curve graph obtained from different situations, we have six thresholds $r\_g\_h, r\_g\_l, b\_g\_h, b\_g\_l, b\_r\_h$ and b_r_l need to be determined. First, we scan the histogram for r_g from left to right. Once the frequency value of r_g is more than 3%, the threshold r_g_left is obtained. Then, scanning the histogram from right to left. Once the frequency value of r_g is more than 3%, the threshold r_g_right is obtained. Then we can yield the threshold r_g_h and r_g_l using formula (4).

$$\begin{cases} r\_g\_h = r\_g\_right \\ r\_g\_l = r\_g\_left \end{cases} \tag{4}$$

Similarly, the threshold b_g_h, b_g_l, b_r_h and b_r_l can also be obtained.

Determine Light Conditions. We identify light conditions according to color images firstly. Concrete steps are as follows: Step 1: calculating gray value of each pixel in the original color images according to the formula (5).

$$G(i,j) = (R_{i,j} + G_{i,j} + B_{i,j})/3 \tag{5}$$

where $R_{i,j}, G_{i,j}$ and $B_{i,j}$ are values of the three components of color image pixel, whose range is $[0 \sim 255]$. Step 2: Counting the total number(sum(i)) of each pixel value(i). Initially determine the light conditions according to the total number of pixel values within a certain range.

Binarization. Binary images has a very special meaning in the process of images. A binary image is a logic array of only 0 and 1 values. For the input color image X, calculate characteristic values $r\_g_{i,j}, b\_g_{i,j}$ and $b\_r_{i,j}$ of each pixel (i,j). According to the thresholds calculated above, perform binarization to color image X using formula (6) to obtain binary image Y, which is shown on the right of the first row in Fig. 5.

$$Y_{i,j} = \begin{cases} 1, if \begin{bmatrix} (r\_g\_l \le r\_g_{i,j} \le r\_g\_h) \\ \&(b\_g\_l \le b\_g_{i,j} \le b\_g\_h) \\ \&(b\_r\_l \le b\_r_{i,j} \le b\_r\_h) \end{bmatrix} \\ 0, otherwise \end{cases} \tag{6}$$

Where $r\_g_{i,j}, b\_g_{i,j}, b\_r_{i,j}$ are characteristic values of pixel (i,j) in the original color image.

2.1 Mathematical Morphological Operation and Final Location

Obtain Candidate Areas. Mathematical morphology is a nonlinear filtering method which can be used to noise suppression, feature extraction, edge detection, image segmentation and other image processing problems. Dilation and erosion operations are the basis for morphological image process. In practical applications of image processing, we use dilation and erosion in several combination forms. There are many relatively isolated and scattered points in binary images and these points are concentrated near the plate regions. So we can use mathematical morphology operations by the following steps: Step 1: Performing closing operation to the binary image two times. The structural elements of dilation and erosion used in the first closing operation are 1×10 matrix and 1 × 3 matrix. The operation removes a number of small protrusions and left a number of breaking marks and narrow connections. And then use the closing operation again. The structural elements of dilation and erosion used this time are 1 × 10 matrix and 1 × 18 matrix. After the operation, some high density areas form a closed and connected rectangle area, the surrounding noises of license plate areas can be removed, and the areas of license plates are closed as far as possible to form rectangular blocks. Step 2: Removing noise. Because the area of license plate would not be small, we regard some small areas as noise and remove them, that is, removing the connected regions whose area is less than 300 pixels. As shown in Fig. 3, considering the distribution of the white dots in binary images, we use the special operators of erosion and dilation to the binary images.

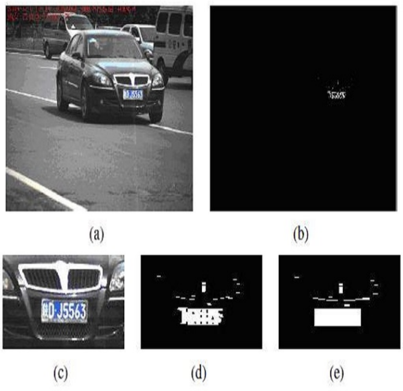


Fig. 3. (a) original color image,(b) binary image,(c) magnified portion of license plate in column(a),(d) image after morphological processing,(e) image after second morphological processing

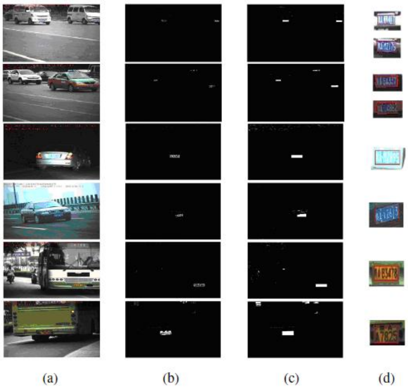


Fig. 4. Results of the experiment: column(a) original color images containing license plate,column(b) binary images,column(c) images in the process of morphology,column(d) enlarged images of the positioning results

Extract Plate Area. In China, there are many features of license plates which can be used to extract plate areas, mainly in the following categories: (1) Geometry features: the standard license plates are rectangular, and have uniform size, 450mm width and 150mm height. Each character on the plate is with 45mm width and 90mm high. Interval between the second and the third character is 34mm,which includes one dot whose diameter is 10mm. Interval between other neighboring characters is 12mm. (2) Texture features: the edge gray histogram of license plate includes two distinctly separated distribution center. Horizontal or vertical projection shows a continuous peak-valley-peak distribution. (3) Color Features: as mentioned above there are four standard types of license plates. These characteristics make a great contrast between backgrounds and characters of license plates.

2.2 Combined with Adaboost Algorithm

The basic idea of adaboost [5–9] is to get a strong classifier by selecting a combination of weak classifiers.The weak classifiers are weak because their accurate rates are not very high. We get the set of weak classifiers by training procedure which selects the best weak classifier with the lowest error rate after each round of training. The training procedures are shown as following: 1) The training data set $(x1, y1), ..., (xn, yn)$ must be prepared carefully, the images which often appear in the normal circumstance should be chosen.$x_i \in X$ stands for the sample and each $y_i \in Y = 0, 1, Y_i = 0$ means this sample is a positive one,$Y_i = 1$ with the opposite meaning, n is the total number of the samples in the set. 2) At beginning, the weight of every sample is initialized with the value 1/2m or 1/2n, m indicates the number of positive samples while n for negative ones, thus each positive sample will get the weight 1/2m, the negative sample will get a 1/2n as well. 3) The training will be done for t = 1, ..., T rounds, after each round, we will get a weak classifier. To start the process, the weights must be normalized as $\omega_{t,i} = \omega_{t,i} / \sum_{j=1}^{n} \omega_{t,i}$ respectively to make a $\omega_{t,j}$ probability distribution. Then for each feature(weak classifier), the error rate is evaluated with respected to ω_t.

$$\epsilon_j = \sum \omega_i \left| h_j(x_i, y_i) \right| \tag{7}$$

4) Find the weak classifier h_t with the least error rate. 5) Get the weights updated according to:

$$\omega_{(t+1),i} = \omega_{t,i} \beta_t^{1-e_i} \tag{8}$$

where $e_i = 0$ if x_i is classified correctly, otherwise $e_i = 1$,and $\beta_t = \epsilon_t / 1 - \epsilon_t$.It forces the weak classifiers to concentrate on the "harder" examples that are often misclassified. 6) Finally we get the strong classifier which combines the weak classifiers' respective votes in a weighted manner:

$$C(x) = \begin{cases} 1, \sum_{t=i}^{T} \alpha_t h_t \geq (1/2) \sum_{t=i}^{T} \alpha_t \\ \alpha_t = \log 1/\beta_t \\ 0, other \end{cases} \tag{9}$$

7) The strong classifier we have got above will be applied to the scanning of a whole image to find the subregions covers a license plate. First we can obtain several First Level Classifiers by using each character like color, gray scale and texture which are shown in previous section because adaboost algorithm can integrate them. And it will eventually achieve a Third Level Classifier. From the view point of classifier, we can treat R-G, B-G and B-R as a First Level Classifier respectively. They combine to form a classifier that involves only information of the color D-value, thus it can be classified as a Color D-value Second Level Classifier. Each type of information, such as information of sum of gray value, variance information and transition point information, is constituted by a number of specific features.

3 Experiment Simulation and Analysis

To verify the practicality and reliability of the new adaboost algorithm combined with color differential model, color images with different amount of license plates under different conditions have been tested by the algorithm based on RGB color model in Matlab7.0 and VC6.0. The size of each image is 1392×1040. The results are shown Fig. 4. From the results shown in the first line and second line in Fig. 4, we can see that the algorithm based on RGB color model can locate plates under strong or dim light conditions, indicating that the algorithm can overcome the light sensitive issues of typical location algorithms based on color information. It is also easy to find from the final position results that the positioning accuracy of the algorithm based on RGB color model is very high, and the algorithm can simultaneously locate multiple license plates in the same image. We can see from the third row, when positioning license plates in the images of the rear of cars at night, we can locate license plates accurately. But when positioning license plates in the images of the front of cars at night, because of the impact of light of headlamps, we can not locate license plates accurately.

4 Comparison with Other License Plate Location Algorithms

We have discussed the procedure of adaboost algorithm combined with color differential model, which has provided good robustness to adjust to various weather conditions and backgrounds. Now we do some comparisons between adaboost algorithm combined with color differential model and other algorithms.

4.1 Comparison with Color Algorithm

Though the license plate location algorithm based on color information can locate multi-plates in one image with an agreeable speed, it is poor robustness. However, because of the contribution of the other characteristics such as texture information, variance and information of hopping frequency, adaboost

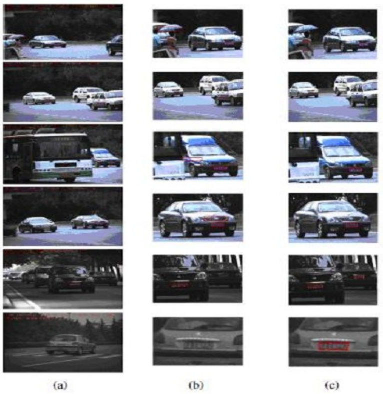


Fig. 5. Comparison results:column(a) images processed by color algorithm alone,column(b) magnified portion of images showed in column(a),column(c) images processed by adaboost algorithm combined with color differential feature

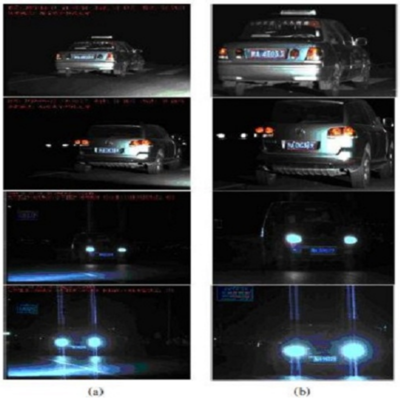


Fig. 6. Experiment results:column(a) original color images gained at night,column(b) images processed by adaboost algorithm combined with color differential model

algorithm combined with color differential model is able to overcome the most shortcomings of color algorithm and can locate accurately license plates . In both conditions, license plates in the color images will be detected more or less than the number of license plates directly using the RGB color model algorithm. By combining other characteristics like gray scale, texture, variance information and so on, license plates will be detected accurately by adaboost algorithm combined with color differential model.The experimental results of comparison are shown in Fig.5.

4.2 Experiment Results of Images Obtained at Night by the Novel Adaboost Algorithm

In order to further evaluate the performance of the proposed algorithm, the experiments using images obtained at night are made, as shown in Fig. 6. It can be easily indicated that the algorithms based on color information or texture hopping could not reach a high detection rate in a large image with high resolution as we used for the testing, especially the algorithm based on texture hopping. When the color information of license plates is always not clear, the adaboost algorithm combined with color differential model has a higher performance than the algorithms based on color information or texture hopping. Especially at night, the precision rate of the adaboost algorithm combined with color differential model is even higher than that of the algorithm based on RGB color model. However, when the illumination is strong enough, the algorithm based on RGB color model performs better than the adaboost algorithm combined with color differential model. Applying these four algorithms to 758 color images which containing different amount and different colors of license plates under different conditions, the results of positioning are shown in the Table 1. In the Table 1, We use precision rate defined in [10]. The data shown in the Table 1 indicates that adaboost algorithm combined with color differential model is superior than the original adaboost algorithm. Because color images contain a lot of interference texture information in the daytime, we also can see that the precision rate of the adaboost algorithm combined with color differential model is less than that of the algorithm based on RGB color model in the condition of strong and weak light in the Table 1.

Table 1. License plate location statistics

Conditions		*Strong light*	*Weak light*	*At night*
Total NO. of images		307	371	80
Adaboost algorithm combined with color feature	*NO. of images with accurate positioning*	281	326	76
	Precision rate	91.5%	87.9%	95.0%
Original adaboost algorithm	*NO. of images with accurate positioning*	250	290	68
	Precision rate	81.4%	78.2%	85.0%
Algorithm based on RGB color model	*NO. of images with accurate positioning*	294	348	52
	Precision rate	95.8%	93.8%	65.0%
Algorithm based on texture hopping	*NO. of images with accurate positioning*	244	284	66
	Precision rate	79.5%	76.5%	82.5%

5 Conclusion

This paper describes a color differential model and a novel adaboost algorithm. By combining the original adaboost algorithm with color differential model, we establish the novel adaboost algorithm. Basically, the adaboost algorithm combined with color differential model overcomes a lot of shortcomings of the traditional algorithms like the algorithm based on color information and the algorithm based on texture hopping. The robustness of the novel adaboost algorithm is stronger than that of other algorithms. However, the time-consuming of the processing in Matlab7.0 for each image(1392*1040 pixels) by the novel adaboost algorithm, which is less than 1s, is more than that of other algorithms. Results of experiment show that the novel adaboost algorithm is of high practical value.

Acknowledgments. We would like to thank the anonymous reviewers for their valuable comments.This work is supported by Beilin District 2012 High-tech Plan, Xi'an, China (No.GX1504) and supported by Xi'an science and technology project(CXY1440(6)) and Shannxi Scientific research(2014k07-11). This job is also supported by Scientic Research Program Funded by Shaanxi Provincial Education Department (Program No.2013JK1139) and Supported by China Postdoctoral Science Foundation (No.2013M542370)and supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China Grant (No. 20136118120010)and supported by Shaanxi Province scientic and technological project (No.2013K06-09)and supported by National Natural Science Foundation of China (No.61305109 and No.61401324)and supported by 863 Item (No.2013AA014601).And this project is also supported by NSFC Grant (Program No. 11301414 and No.61272283 and No.61272283 and No.61472318).

References

1. Li, Y., Gao, M.: A license plate location algorithm based on the edge and Color information. J. of Comput. Simulation **26**(8), 262–265 (2009)
2. Mahini, H., Kasaei, S., Dorri, F., Dorri, F.: An efficient features-based license plate localization method. In: Proc. of 18th Int. Conf. on Pattern Recognition, pp. 1101–1103 (2006)
3. Dong, Y., Ma, J.: Wavelet-based image texture classification using local energy histograms. In: LSP, vol. 18, pp. 247–250 (2011)
4. Wang, J., Xiao, G., Jiang, J.: An Image Retrieval Algorithm Based on the HSI Color Space Accumulative Histogram. Computer Engineering & Science **129**(14) (2007)
5. Zhang, X., Shen, P., Xiao, Y., Li, B.: License plate-location using adaboost algorithm. In: Proc. of the 2010 IEEE Int. Conf. on Inf. And Autom., pp. 20–23, June 2010
6. Gonzalez, R.C.: Digital Image Processing using MATLAB. Publishing House of Electronics Industry, Beijing (2005)
7. Zhang, X., Shen, P., Bai, J., Lei, J., Hu, Y., Xiao, Y., Li, B., Qi, D.: License plate location based on adaboost. In: 2010 IEEE International Conference on Information and Automation (ICIA), pp. 1705–1710, June 20–23, 2010
8. Sami ul Haq, Q., Tao, L.: Neural network based adaboosting approach for hyperspectral data classification. In: ICCSNT, vol. 1, pp. 241–245 (2011)
9. Tang, X., Ou, Z., Su, T., Zhao, P.: Cascade adaboost classifiers with stage features optimization for cellular phone embedded face detection system. In: Wang, L., Chen, K., S. Ong, Y. (eds.) ICNC 2005, Part III. LNCS, vol. 3612, pp. 688–697. Springer, Heidelberg (2005)
10. Gerace, I., Martinelli, F.: Restoration of recto-verso archival documents through a regularized nonlinear model. In: TPAMI, vol. 24, pp. 1399–1404 (2002)
11. Wei, W., Yong, Q.: Information Potential Fields Navigation in Wireless Ad-Hoc Sensor Networks. Sensors **11**, 4794–4807 (2011)
12. Wei, W., Yang, X., Zhou, B., Shen, P.: Holes Detection in Anisotropic Sensornets: Topological Methods. International Journal of Distributed

A New Image Watermarking Algorithm Using the Contourlet Transform and the Harris Detector

Dandan Zhu[1,2](✉) and Lizhi Lv[3]

[1] College of Information Science and Engineering, Northeastern University,
Shenyang 110819, China
zhudan0526@163.com

[2] Department of Computer Science, Tonghua Normal University, Tonghua 134000, China

[3] College of Computer Science and Technology,
Taiyuan University of Technology, Taiyuan 030024, China

Abstract. In this paper, we propose a new feature-based image watermarking scheme based on multiscale theory and the Contourlet transform (CT). We use the multiscale Harris detector to extract stable feature points from the host image. Next, according to feature scale theory, we determine the local feature regions (LFR) and scale the regions to a standard size. We then embed the digital watermark into the Contourlet low frequency area calculated using the pseudo-Zernike moment. The results of our experiments demonstrate that the algorithm results in an invisible watermark and is robust against conventional signal processing (median filtering, sharpening, noise adding, and JPEG compression), geometric attacks (rotation, translation, scaling, row or column removal, shearing, local geometric distortion) and combined attacks.

Keywords: Image watermarking · Geometric attacks · Contourlet transform · Pseudo-Zernike moments · Low sub-band1

1 Introduction

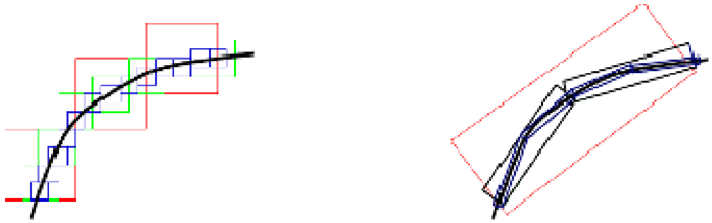
Digital watermarking is widely used in Internet multimedia intellectual property protection [1-2]. Great progress has been achieved in the image transform domain by applying the discrete cosine transform (DCT)[3-4] and discrete wavelet transform (DWT)[5-6] in digital watermarking algorithms. However, these theories do not adequately represent the anisotropy of signals. The Ridgelet [7], Curvelet [8], and Concourlet [9] transforms better address the anisotropy of signals and can effectively resist Random bending attack (RBA), Geometric attacks (i.e., Rotation, Scale and Translation, RST), and Shearing. Many methods have been proposed to evade geometric attacks. These methods can be roughly divided into three categories: (1) invariant transforms [12-13], which are robust against global geometrical distortion but not against shearing; (2) feature-based synchronization [14-15], whose watermarking capacity is very limited; and (3) template insertion [16], which cannot withstand any malicious attacks.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 439–447, 2015.
DOI: 10.1007/978-3-662-48570-5_42

To address the above problems, we present a new digital watermarking scheme, which includes two main features. First, an analysis of Contourlet transform characteristics indicates that the proposed scheme can adaptively embed a watermark into low frequency sub-bands with many textures. Second, we review feature-based synchronization methods and propose a novel watermark synchronization strategy. In experiments, we compare the performance of the proposed algorithm with other algorithms by applying various attacks. The results demonstrate the superior robustness of the proposed watermarking algorithm.

2 Contourlet Transform

The Contourlet transform is a geometrical transform that uses multiresolution analysis and multi-direction analysis to effectively show the contour and texture of an image. The basis functions of the Contourlet transform provide different scales and thus extend support for the aspect ratio, assuming that linear and face discontinuities are preferred. Compared with the wavelet transform, the Contourlet transform offers more and richer basis functions. It can use fewer coefficients to represent smooth edges and combine the discontinuity points with the same directionality into a discontinuity line or face. The main characteristics of the Contourlet are compared with those of DWT in Figure 1.



(a) DWT (b) Contourlet

Fig. 1. Comparison of basis functions

3 Scale-Adaptive Harris Detector

The Harris detector is based on the second moment matrix [17] which is an image descriptor reflecting the distribution of the local gradient directions of the image. The scale-adaptive second moment matrix is defined as:

$$\begin{aligned} & M(x,y,\delta_I,\delta_D) \\ & = \delta_D^2 \cdot G(\delta_I) * \begin{bmatrix} L_x^2(x,y,\delta_D) & L_x L_y(x,y,\delta_D) \\ L_x L_y(x,y,\delta_D) & L_y^2(x,y,\delta_D) \end{bmatrix} \end{aligned} \tag{1}$$

δ_I denotes the integration scale, δ_D denotes the differentiation scale, and L_a denotes the partial derivative in the a direction. The uniform Gaussian scale space representation L is defined as:

$$L(x, y, \delta_D) = G(x, y, \delta_D) * f \tag{2}$$

Here, G denotes the Gaussian function with zero mean and δ_D is the standard deviation. f denotes the image and $*$ indicates linear convolution.

Given δ_I and δ_D, the second moment matrix $M(x, y, \delta_I, \delta_D)$ can be used to compute the scale-adaptive Harris corner strength (SHCS) detector:

$$\begin{aligned} &R(x, y, \delta_I, \delta_D) \\ &\quad = Det(M(x, y, \delta_I, \delta_D)) - k \cdot Tr^2(M(x, y, \delta_I, \delta_D)) \end{aligned} \tag{3}$$

Here $Det(\bullet)$ denotes the determinant of the matrix and $Tr(\bullet)$ is the trace. At each scale-space level, the feature points are extracted according to the following rules:

$$\begin{aligned} &R(x, y, \delta_I, \delta_D) > R(\hat{x}, \hat{y}, \delta_I, \delta_D) \qquad \forall(\hat{x}, \hat{y}) \in A \\ &R(x, y, \delta_I, \delta_D) \geq t_u \end{aligned} \tag{4}$$

Here, A is the neighborhood of pixel (x, y) and t_u is the threshold.

Automatic Scale Selection and Scale-Invariant Feature Points

We use the Laplacian-of-Gaussians (LOG) operator to determine the characteristic scale. The LOG is defined as:

$$LOG(x, y, \delta_I) = \delta_I^2 \left| \frac{\partial^2 G(x, y, \delta_I)}{\partial x^2} + \frac{\partial^2 G(x, y, \delta_I)}{\partial y^2} \right| \tag{5}$$

$$LOG = \frac{1}{10}(6 \cdot LOG_Y + 2 \cdot LOG_{Cb} + 2 \cdot LOG_{Cr}) \tag{6}$$

The steps to extract feature points using the Harris-Laplace detector are as follows:

(i) A scale-space representation is built with the Harris function for the preselected scales $\delta_n = 1.4\xi^n$, where ξ denotes the scale factor between continuous levels. At each representation level, the SHCS is computed with $\delta_I = \delta_n$ and $\delta_D = s\delta_n$ (s is a constant). After that, the candidate points that are maximal in the 8-neighborhood with SHCS greater than $t_u = 1000$ are extracted.

(ii) An iterative algorithm is applied to compute each candidate point's scale and location. The scales of feature points are selected by the extrema over the LOG scale.

For an initial point p having scale δ_I, the iteration scheme can be described as the following:

1) For a point p_k, search its local extremum over the LOG scale, otherwise reject it. Limit the investigated scope of the scale to $\delta_I^{(k+1)} = t \cdot \delta_I^{(k)}$ with $t \in [0.7, \cdots, 1.4]$.

2) Search the spatial point p_{k+1} of a maximum of the SHCS nearest to p_k for $\delta_I^{(k+1)}$.

3) Return to Step 1 until $\delta_I^{(k+1)} = \delta_I^{(k)}$ or $p_{k+1} = p_k$.

Local Characteristic Regions

We consider the problem of geometric synchronization in our choice of LCR construction method. We select a circular area that is independent of image rotation and determine its size by the characteristic scale. Define the radius $\Re$ of the LCRs as:

$$\Re = \tau \cdot \lceil \delta \rceil \tag{7}$$

Here δ denotes the characteristic scale and τ denotes a positive integer to adjust the size of the LCRs. The robustness of the CBIR system increases for a small LCR while the capacity decreases. Thus, there is a tradeoff between these two factors. The theoretical range of $\Re$ is:

$$round(\delta) \le \Re \le \frac{\min(M, N)}{2} \tag{8}$$

We reserve the LCR with the highest SHCS and discard the others as the higher SHCS indicates a more robust feature point.

4 Watermark Embedding Algorithm

Watermark embedding can be regarded as an additive process of a strong background signal (the original image) and a weak signal (the watermark). The watermark is embedded into the significant features. The embedding scheme can be decomposed into the following seven steps:

1) Generate a random sequence $W = \{w_i, i = 1, \cdots, L\}$ using the secret key K_1, where the sequence values are $w_i \in \{0,1\}$ and L denotes the size of the sequence.

2) Apply the Harris-Laplace detector to the host image to obtain a feature point set $P = \{p_i, i = 1, \cdots, n\}$. Use these points for the reference centers of the local feature regions.

3) Use our proposed method to construct the local feature regions $O = \{o_k, k = 1, \cdots, m\}$.

4) Perform a zero-padding operation on each $o_k \in O$. Apply the normalization procedure to each block of size $2R_K \times 2R_K$ (R_K is the radius of the LFR) to map the disk (LFR) to the block.

5) Transform the $2R_K \times 2R_K$ image using the Contourlet transform, search the low frequency regions of the directional sub-band to be embedded, and calculate the pseudo-Zernike moment for these regions. The watermark signal is generated by using the quantification of the modulation of the pseudo-Zernike torque amplitude. The quantitative rule is defined as follows:

$$A'_{p_i q_i} = \left\lceil \frac{A_{p_i q_i} - d(w_i)}{\Delta} \right\rceil \Delta + d(w_i) \quad (i = 1, \ldots L) \tag{9}$$

6) Generate watermarked images. First, the pseudo-Zernike moment for the LFR image $f^*(x, y)$ reconstruction is modified by:

$$f^*(x, y) = f_O(x, y) - f_Z(x, y) \tag{10}$$

where $f_O(x, y)$ is the original LFR image and $f_Z(x, y)$ denotes the modified reconstruction image of the pseudo-Zernike moment:

$$f_Z(x, y) = \sum_{i=1}^{L} Z_{p_i q_i} V_{p_i q_i}(\cdot) + Z_{p_i, -q_i} V_{p_{i,}, -q_i}(\cdot) \tag{11}$$

Second, the pseudo-Zernike moment for the image $f_{Z'}(x, y)$ reconstruction is modified by:

$$f_{Z'}(x, y) = \sum_{i=1}^{L} Z'_{p_i q_i} V_{p_i q_i}(\cdot) + Z'_{p_i, -q_i} V_{p_{i,}, -q_i}(\cdot) \tag{12}$$

7) Reconstruct the CT to obtain the watermarked image. The zero-removal operation is used on each watermarked block to obtain the watermarked disk ok*. After that, ok* is substituted for ok. Repeat steps (4)–(7), until all LFRs are used. Then the watermarked image is obtained.

5 Watermark Detection Algorithm

Generally, watermark detection is the inverse process of watermark embedding. In this paper, the detection process can be divided into five steps:

1) The original watermark $W = \{w_i, i = 1, \cdots, L\}$ is extracted according to the key K_1.

2) The Harris-Laplace detector (see Section 2) is used to process the host image. We can extract many feature points ($\tilde{P} = \{\tilde{p}_i, i = 1, \cdots, n\}$) regarded as the reference centers of the local feature regions.

3) Our proposed method is used to construct a set of local feature regions $\tilde{O} = \{\tilde{o}_k, k = 1, \cdots, m\}$.

4) The zero-padding operation is performed on each $\tilde{o}_k \in \tilde{O}$ and the normalization procedure is used to process each block of size $2R_K \times 2R_K$ to map the disk (LFR) to the block of size $2R_K \times 2R_K$ (where R_K is the radius of the LFR).

5) The Contourlet is used to transform the $2R_K \times 2R_K$ image by extracting the low frequency area of the directional sub-band in the low frequency regions calculated using the pseudo-Zernike moment. The embedding process uses the following expression.

$$(A'_{p_i q_i})_{Q\,j} = \left[\frac{A'_{p_i q_i} - d(j)}{\Delta}\right]\Delta + d(j) \quad (j = 0,1) \tag{13}$$

Through the above expression, we can obtain the two sets of vectors $(A'_{p_i q_i})_0$ and $(A'_{p_i q_i})_1$, $i = 1, \ldots, L$.

The following definition of wi' gives the result of the detection:

$$w'_i = \arg\min_{j \in \{0,1\}} \left((A'_{p_i q_i})_{Q_j} - (A'_{p_i q_i})\right)^2 (i = 1, \ldots, L) \tag{14}$$

If wi′=0 then the watermarking detection fails; otherwise, if wi′=1 then the watermarking detection is successful. When at least two disks are determined to be watermarked, the final detection is labeled as a success; otherwise, it fails.

6 Simulation Results

We conduct experiments to verify the effectiveness of our proposed method and compare it with the method in [17]. The proposed watermarking scheme is tested on the standard 512 × 512 pixel test images Lena, Baboon, and Pepper. The watermark pattern is a 16-bit pseudo random binary sequence. The watermark is embedded in the perceptually textured region and is therefore less visible.The comparison of the detection results under common signal processing and de-synchronization attacks are shown in Table 1 and Table 2. Each result is expressed as the ratio between the number of correctly detected watermarked LCRs and the number of original embedded watermarked LCRs.

Table 1. The watermark detection results for common signal processing attacks (detection rates)

Attacks	Lena		Baboon		Pepper	
	Proposed scheme	Scheme in [17]	Proposed scheme	Scheme in [17]	Proposed scheme	Scheme in [17]
Median filter(3×3)	6/9	3/10	10/12	3/9	8/11	1/12
Sharpening (3×3)	4/9	8/10	8/12	4/9	7/11	4/12
Salt&Pepper noi (0.02)	5/9	2/10	7/12	5/9	5/11	2/12
Gaussian noise(3×3)	5/9	4/10	9/12	4/9	6/11	1/12
Sharpen-ing(3×3)+JPE G70	4/9	2/10	7/12	4/9	4/11	3/12
Median filter(3×3) +JPEG70	3/9	2/10	6/12	2/9	3/11	2/12

Table 2. The watermark detection results for de-synchronization signal processing attacks

Attacks		Lena		Baboon		Pepper	
		Proposed scheme	Scheme in [17]	Proposed scheme	Scheme in [17]	Proposed scheme	Scheme in [17]
Removed 1 row and 3 columns		7/9	5/10	9/12	6/9	5/8	5/12
Removed 5 rows and 10 columns		6/9	4/10	7/12	3/9	3/8	4/12
Centered crop-ping 5% off		5/9	3/10	6/12	3/9	4/8	3/12
Centered crop-ping 10% off		5/9	4/10	5/12	2/9	3/8	2/12
-x-shearing5%, -y-shearing 5%		4/9	2/10	4/12	2/9	3/8	2/12
Shearing 60%		3/9	1/10	3/12	3/9	3/8	2/12
Rotation 5		4/9	2/10	4/12	2/9	3/8	2/12
Rotation 10		3/9	1/10	3/12	2/9	2/8	1/12
Translation-x-10 and –y-10		7/9	4/10	6/12	4/9	5/8	4/12
Scaling	0.8	5/9	1/10	4/12	2/9	4/8	4/12
	1.4	3/9	2/10	5/12	3/9	3/8	3/12
Local random bending		4/9	3/10	6/12	4/9	3/8	2/12
Removed 5 rows and10 columns and Scaling 0.8		3/9	2/10	4/12	3/9	3/8	1/12
Removed 5 rows and10 columns and +JPEG70		4/9	3/10	5/12	3/9	5/8	1/12
Centered crop-ping 5%+ Rota-tion 10		4/9	2/10	4/12	2/9	3/8	2/12

7 Conclusion

In this paper, we propose a robust image watermarking scheme based on scale-space theory to resist common signal-processing and even de-synchronization attacks. The key characteristics of the proposed scheme are:

1) The CT provides multiresolution analysis and directional preservation, which provides robustness against translation and noise attacks.
2) Under various attacks, the feature points extracted by the Harris-Laplace detector are reliable, which facilitates re-synchronization between watermark embedding and detection.

In addition, our proposed method can extract the watermark without using the original image and has a low computational complexity rendering it applicable to many different situations.

References

1. Dragoi, I.C., Coltuc, D.: Local prediction based difference expansion reversible watermarking. IEEE Trans. Image Process. **23**, 1779–1790 (2014)
2. Bad, P., Furon, T.: A new measure of watermarking security: the effective key length. IEEE Trans. on Infor. Forensics and Security **43**, 1306–1317 (2013)
3. Harish, N.J., Kumar, B.B.S., Kusagur, A.: Hybrid robust watermarking techniques based on DWT, DCT, and SVD. Int. J. Adv. Electron. Eng. **2**(5), 137–143 (2013)
4. Singh, S.P., Rawat, P.: A robust watermarking appoach using DCT-DWT. Int. J. Emerg. Technol. Adv. Eng., 300–305 (2012)
5. Awasthi, M., Lodhi, H.: Robust image watermarking based on discrete wavelet transform, discrete cosine tranform & singular value decomposition. Adv. Electron. Eng., 971–976 (2014)
6. Jayalakshmi, M., Merchant, S.N., Desai, U.B.: Significant pixel watermarking using human visual system model in wavelet domain. In: Kalra, P.K., Peleg, S. (eds.) ICVGIP 2006. LNCS, vol. 4338, pp. 206–215. Springer, Heidelberg (2006)
7. Jiao, L.C., Tan, S., Liu, F.: Ridgelet theory: From ridgelet transform to curvelet. Chinese J. Engrg. Math. **22**, 761–773 (2005)
8. Starck, J.L., Candes, E., Donoho, D.: The Curvelet Transformfor Image Denoising. IEEE Trans. Image Processing **11**, 670–684 (2002)
9. Do, M.N., Vetterli, M.: The Contourlet transform: an efficient directional multiresolution image representation. IEEE Transactions on Image Processing **14**(12), 2091–2106 (2005)
10. da Cunha, L., Zhou, J., Do, M.N.: The Nonsubsampled Contourlet Transform: Theory, Design, and Applications, 3089–3101 (2004)
11. Cunha, A.L., Zhou, J., Do, M.N.: Nonsubsampled contourilet transform: filter design and applications in denoising. In: IEEE International Conference on Image Processing, vol. 1, pp. 749–752 (2005)
12. Zheng, D., Zhao, J., Saddik, A.E.: RST-invariant digital image watermarking based on log-polar mapping and phase correlation. IEEE Trans. Circuits Syst. Vid. Tech. **13**, 753–765 (2003)
13. Zheng, D., Liu, Y., Zhao, J., Saddik, A.E.: A survey of RST invariant image watermarking algorithms. ACM Comput. Surv. **39**(2), 1–91 (2007)

14. Hao, Y., Bao, G., Zhang, H.: Secure public digital watermarking detection scheme. In: Congress on Image and Signal Processing, CISP 2008, vol. 5, pp. 725–729 (2008)
15. Bas, P., Chassery, J.M., Macp, B.: Geometrically invariant image watermarking based on statistical features in the low-frequency domain. IEEE Tranc. Circ. Sys. Video Technol. **18**(6), 777–90 (2008)
16. Chen, J., Yao, H., Gao, W., Liu, S.: A robust watermarking method based on wavelet and Zernike transform. In: Proceedings of the 2004 International Symposium on Circuits and Systems, ISCAS 2004, vol. 2, pp. 173–176 (2004)
17. Yuan, X.C., Pun, C.M.: Geometrically invariant image watermarking based on feature extraction and Zernike transform. Int. J. Secur. Appl. **6**(2), 217–222 (2012)

Multi-object Visual Tracking Algorithm Based on Grey Relational Analysis and Generalized Linear Assignment

Yanzhao Su(✉), Aihua Li, Zhigao Cui, Hao Fang, and Tao Wang

502 Faculty, Xi'an Institute of High Technology, Xi'an, Shan Xi, China
syzlhh@163.com, {aihuali,zhigaocui,haofang}@126.com,
taowang@hotmail.com

Abstract. In the view of multi-object tracking in video sequences affected by the issues of similar objects and occlusion in objects, etc., a hierarchy fusion visual tracking algorithm based on gray relational analysis were proposed in this paper. In the algorithm, object trajectory was associated step by step and the video sequences was processed by adding time windows. First, tracklets were provided by a conservative association of the detections. Then, in every time window, combined with the improved grey degree of incidence and moving information, the similarity of two trajectory was calculated. In the end, the optimal association of the tracklets was achieved according to the generalized linear assignment. By comparison with typical algorithms, experimental results show that the algorithm is applicable to multi object tracking in the scenes without reliable appearance characteristic provided with higher tracking accuracy, and adapt to the effect of object occlusion, similar appearance, camera motion and so on.

Keywords: Grey relational analysis · Generalized linear assignment · Multi-object · Visual tracking

1 Introduction

Multi-object visual tracking[1-2] is important for many computer vision applications including intelligent control, human-computer interaction, virtual reality, etc. Compared with single object tracking, multi-object tracking faced more challenges (e.g. unknown number of objects, appearance changes and mutual occlusion)[3].With the development of object detection technology (such as background modeling, pedestrian detection, etc.), most current approaches to multi-target tracking are based on tracking by detection. A complete motion trajectory is formed by linking the detections in every frame with motion and appearance features of objects. Zhang. et al. [4] formulate multi-object tracking as the minimum cost flow problem in networks. The algorithm could reduce the number of tracklets significantly and maintain the integrity of the object trajectory with the global objective function. To calculate the similarity of object detections, both the appearance and motion characteristics were adopted. The metric methods are relatively simple, such as the Euclidean distance, Bhattacharya distance, etc. As the appearance characteristics of the object in the actual mon-

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 448–457, 2015.
DOI: 10.1007/978-3-662-48570-5_43

itoring scene were not stable and discriminative, many scholars use motion features to achieve mul-ti-object tracking. Wen et al. [5] measure the motion feature similarity of objects by adopting the forward and backward prediction information with the assumption of uniform motion. However, the accuracy of tracking results relied on the movement pattern of the object. Dicle et al. [6] construct a Hankel matrix with the current tracklets and represent the dynamics-based similarity the matrix rank. To calculate the rank, an improved Hankel total least squares (IHTLS) algorithm was proposed. The algorithm could reduce the influence of data noise and predict the missing information between two tracklets. However, the real-time ability need to be further improved due to the mass computation of the rank minimization estimation.

In this paper, we proposed a multi-object tracking algorithm based on the combination of grey relational analysis and generalized linear assignment. In our method, just the motion feature is adopted to associate the tracklets, and the motion feature similarity was measured by grey relational analysis without assumption on the motion mode of objects and the scenes. Then, the data association was optimized with the generalized linear assignment. Our approach can track the similar appearance objects with high accuracy and few time consumption.

2 Moving Track Association Based on Generalized Linear Assignment

As the interference of object detection and scene factors, tracklets formed by conservative association are discontinuous in time and space. When one tracklet is associated with another, a variety of factors need to be considered such as similarity, time, etc. Therefore, multi-object tracking is usually formulated as an optimization problem, and the above factors are regarded as constraints or parameters of the optimization process.

Assume that N tracklets exist in the scene during a period of time, let $T = \{T_1, T_2, ...T_N\}$ represent the tracklets set, X_{ij} represent the association relationship between tracklet T_i and T_j. X_{ij} means that the two tracklets belong to the same object, otherwise means the opposite. C_{ij} is defined to represent the degree of similarity between two tracklets. Then, the optimization object function of multi-object tracking can be derived as follows:

$$\begin{aligned} &\arg\max_X C_{ij}X_{ij} \ \ or \ \arg\min_X C_{ij}X_{ij} \\ &st.\ X_{ij} \in (0,1) \end{aligned} \tag{1}$$

Using maximization or minimization relies on the value of . However, the above function couldn't be optimized directly and some constraints should be added in. For example, one tracklet could be only associated with one successor or predecessor. These constraints were adopted in [4], and two virtual nodes were produced to simulate the emergence and disappearance of object trajectory. However, as the randomness of object movement, it is difficult to accurately estimate the probability of objects emergence and disappearance without prior knowledge. Therefore, the constraint condition needs to relax that the object track cannot match the trajectory, and the optimization object function was revised as follows:

$$\begin{aligned}&\arg\max_{X} C_{ij} X_{ij}\\ &st.\ X_{ij} \in \{0,1\};\\ &\sum_{i=1}^{N} X_{ij} \le 1; \sum_{j=1}^{N} X_{ij} \le 1;\end{aligned} \tag{2}$$

The above formula is a special form of the generalized linear assignment problem. Although such a relaxation does not need to estimate the object emergence and disappearance, it is an “NP-hard” problem completely and can only be solved by an approximate feasible solution. A “soft partition” algorithm called deterministic annealing was proposed in [7]. An optimal estimation solution could obtain by the method.

In fact, the core of tracking by data association is calculating the similarity parameters. Its accuracy would greatly influence the tracking performance. And the constraints could only affect the solving process and the optimal degree of solution. If the similarity parameters are error, it will not be able to get the right association result with the optimization algorithm. Therefore, we propose to use the gray correlation analysis to measure the motion feature similarity of the tracklets.

3 Multi-object Tracking Based on Grey Relational Analysis

3.1 Grey Relational Analysis

As an important part of the grey theory, grey relational analysis is widely used in image engineering, decision analysis, etc. The essence is to find the complicated relationships among various factors of the system through the geometric similarity between data sequence curves [8][9]. The degree of grey incidence is a specific index of the gray correlation analysis. Its value indicates the degree of data correlation, higher for more correlation, whereas lower for few correlation. The original degree of grey incidence proposed by Deng [10], and the absolute degree of incidence proposed by Liu [11] are commonly used in practical problems. Define two behavior sequences: $X_i = (x_i(1), x_i(2), ..., x_i(n))$, $X_j = (x_j(1), x_j(2), ..., x_j(n))$. Let the sequences $X_i^0 = (x_i^0(1), x_i^0(2), ..., x_i^0(n))$, $X_j^0 = (x_j^0(1), x_j^0(2), ..., x_j^0(n))$ represent another two sequences, and the elements of which are generated by subtracting the start point of X_i and X_j, e.g. $x_i^0(k) = x_i(k) - x_i(1)$. Denote the absolute degree of incidence between X_i and X_j as ε_{ij}, which can be calculated as follows:

$$\varepsilon_{ij} = \frac{1 + |s_i| + |s_j|}{1 + |s_i| + |s_j| + |s_i - s_j|} \tag{3}$$

where $s_i - s_j = \int_1^n (X_i^0 - X_j^0)dt$ is the integral of the difference on two sequences, $s_i = \int_1^n (X_i - x_i(1))dt$ and $s_j = \int_1^n (X_j - x_j(1))dt$ are the integral of the difference in the sequences each other. As the absolute degree of incidence is symmetrical and unrelated to the order of X_i and X_j, we use this method to measure the tracklets similarity in this paper.

3.2 Tracklets Similarity Based on Grey Relational Analysis

Reliable tracklets of the objects could be acquired by conservative association method (such as bipartite graph method [12]). And these tracklets need to be associ-ated again for a complete trajectory of the object.

3.2.1 Tracklets Similarity Based on the Absolute Degree of Incidence

Define $T_i = \{T_i^k, k = 1...n\}$ as a tracklet, where $T_i^k = (x_i^k, y_i^k, w_i^k, h_i^k)$ contains the center coordinates, width and height information of object i, n is the length of the tracklet. So the similarity φ_{ij} between T$_\mathrm{i}$ and T$_\mathrm{j}$ could be calculated as follows:

$$\varphi_{ij} = \frac{\varepsilon_{ij}^x + \varepsilon_{ij}^y}{2} \tag{4}$$

where ε_{ij}^x and ε_{ij}^y represent the absolute degree of incidence in the x , y direction between the center coordinates of the two tracks respectively. And this method is called grey relational analysis (GRA). The specific calculation steps are as follows:

(1) Suppose that the end time t_i^e of T_i is less than the start time t_j^s of T_j, and the track length is n_i and n_j respectively.

(2) Extract center coordinates data of a tracklet with length n_m . For the tracklet T_i, the data is extracted backward, which is started from the time t_i^e . And for the tracklet T_j, the data is extracted forward, which is started from the time t_j^s, The parameter n_m can be calculated as $n_m = \min(\min(n_i, n_j), K)$, where K is a constant, and usually set to 5.

(3) Let X_i , Y_i represent the behavior sequences taken from tracklet T_i, and X_j , Y_j as the behavior sequences taken from tracklet T_j. Then calculate the grey degree of incidence ε_{ij}^x and ε_{ij}^y according to equation (3).

3.2.2 Revised Track Similarity Based on Corrected Degree of Incidence

Using grey relational analysis to measure the similarity mainly rely on the geometrical characteristics of two tracklets only. And there is no need to make any assumption about the movement of objects.

However, the disadvantage of this method is not able to represent a negative correlation [13]. Two tracklets with the same geometrical characteristics, may have the opposite direction and belong to different objects. Furthermore, the contribution of degree of incidence in and direction may be different. In equation (4), it is simply combined with equal weight. Then, an error may yield on the similarity measure. In order to solve the above issue, the moving direction and speed change of objects were taken to revise equation (4), which termed as weighted grey relational analysis(WGRA). The specific process is as follows:

(1) Define $v_x^i, v_y^i, v_x^j, v_x^j$ as the speed in x, y direction of tracklets T_i and T_j which can be derived as follows ：

$$\begin{cases} v_x^i = x_i^{n_i} - x_i^{n_i - 1} \\ v_y^i = y_i^{n_i} - y_i^{n_i - 1} \\ v_x^j = x_j^2 - x_j^1 \\ v_y^j = y_j^2 - y_j^1 \end{cases} \tag{5}$$

(2) Define θ as the angle between motion the directions of object and calculate its cosine as below.

$$\cos(\theta) = \frac{v_x^i v_x^j + v_y^i v_y^j}{\sqrt{v_x^i v_x^i + v_y^i v_y^i}\sqrt{v_x^j v_x^j + v_y^j v_y^j}} \tag{6}$$

When the difference in the motion direction of two objects is insignificant, the possibility of them from the same object is larger, and vice versa is two objects. Then, this corrected degree of incidence between two tracks can be calculated as follows:

$$\begin{cases} \overline{\varepsilon_{ij}} = \mathrm{sgn}[\cos(\theta) < \lambda] * \varepsilon_{ij} \\ \mathrm{sgn}[\cos(\theta) < \lambda] = \begin{cases} 1 & ,\cos(\theta) < \lambda \\ -1 & ,else \end{cases} \end{cases} \tag{7}$$

(3) Using the speed difference of the traclets to weight the degree of incidence in x and y direction. Then the revised similarity coefficient $\overline{\varphi_{ij}}$ could be calculated as follows:

$$\cos(\theta) = \frac{v_x^i v_x^j + v_y^i v_y^j}{\sqrt{v_x^i v_x^i + v_y^i v_y^i}\sqrt{v_x^j v_x^j + v_y^j v_y^j}} \tag{8}$$

The corrected grey degree of incidence can represent the positive and negative correlation relationship of two tracklets based on the difference of motion direction . The similarity coefficient weighted with speed can provide better discriminate

Compared to the algorithm based on Hankel matrix, measuring the similarity of tracklets based on grey relational analysis has the advantages of more fast and high accuracy.

3.3 Cascade Optimization

After the similarity of tracklets is required, set $C_{ij} = \varphi_{ij}$ to optimize the GLAP. In order to speed up the calculation process, the multi-object tracking was conducted on the cascade optimization method similar with that in [6]. The whole process is as follows: firstly, divide the video sequence into a series of equally spaced clips and associate the object tracklets in each clip; secondly, according to a certain offset, slide a constant width in time and associate tracklets again; thirdly, in double time window clips, associate the exist tracklets again. This method can tolerate various intervals on the motion of objects effectively, so that the trajectory is more and more complete.

4 Experimental Results

4.1 Evaluation Criteria and Experiment Condition

4.1.1 Evaluation Criteria

Many evaluation criteria could be used to judge the performance of multi-object tracking. The author of [14] puts forward the most-used performance-evaluation index. We used four indexes: Multi-Object Tracking Accuracy (MOTA), False Negative (FN), False Positive (FP) and Miss Match (MM), which could be calculated as follows:

$$\begin{cases} \text{MOTA=1-}\dfrac{\sum_t (fn_t + fp_t + mm_t)}{\sum_t gt_t} \\ FN = \sum\limits_t fn_t \\ FP = \sum\limits_t fp_t \\ MM = \sum\limits_t mm_t \end{cases} \tag{9}$$

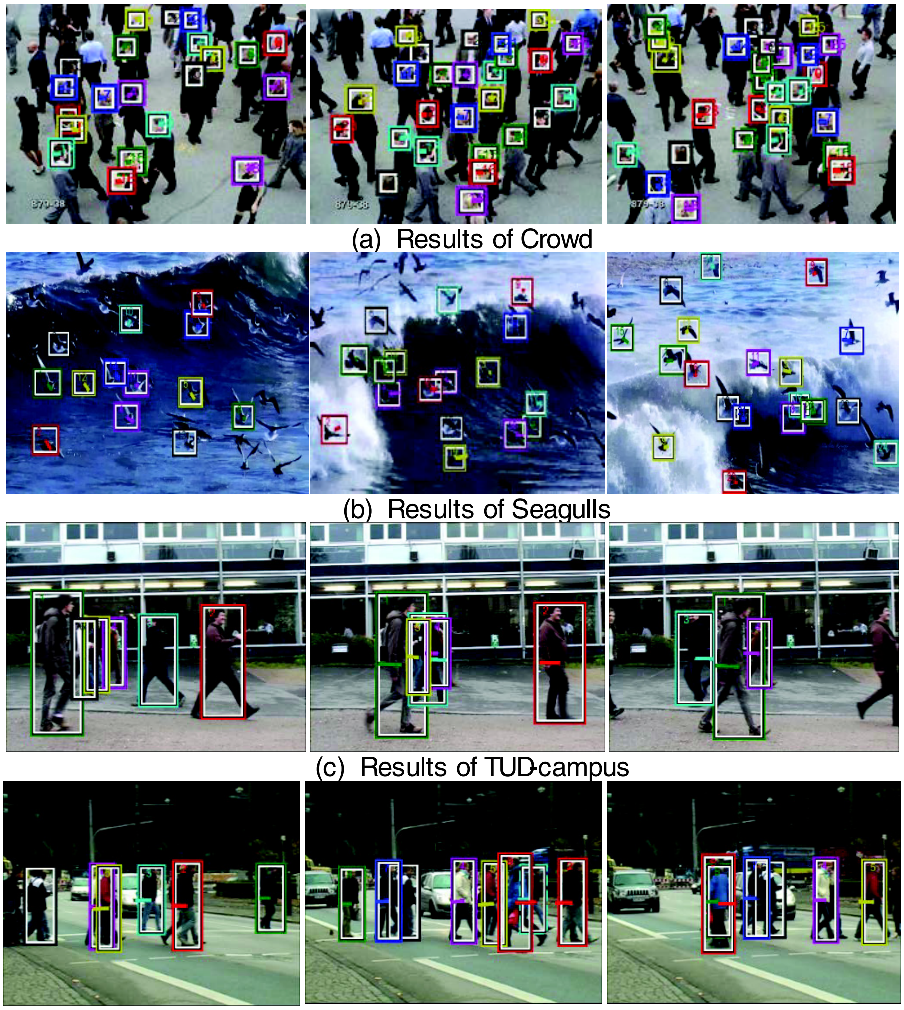
Where fn_t, fp_t , mm_t , gt_t represent the number of incorrectly-correlated tarcklets, the remaining number of positive tarcklets, the number of correlation-changed tracklets and the number of referenced tracklets respectively in frame t. And MOTA combined FN, FP and MM, is a relatively comprehensive reflection on the accuracy of multi-object tracking algorithms. Besides, we also use average processing time per frame (TF) to measure the speed of an algorithm.

4.1.2 Experiment Condition

Our algorithm was developed under Matlab. The tested dataset and parameters are the same with [6]. The parameter λ was always set between (-1,0), such as -0.9 can get good results. Our dataset SMOT has eight videos and targets have similar features in these videos. All the detection results of moving targets were manually annotated. And we compare the results with [6] by the proposed two algorithm, which denote as GRA and WGRA for convenience.

4.2 Experiments Results

The results of our algorithm are shown in Fig. 1. From the results, we can see that utilizing the motion feature, similar appearance objects can be tracked accurately, and the algorithm has the more strong robustness to occlusions.

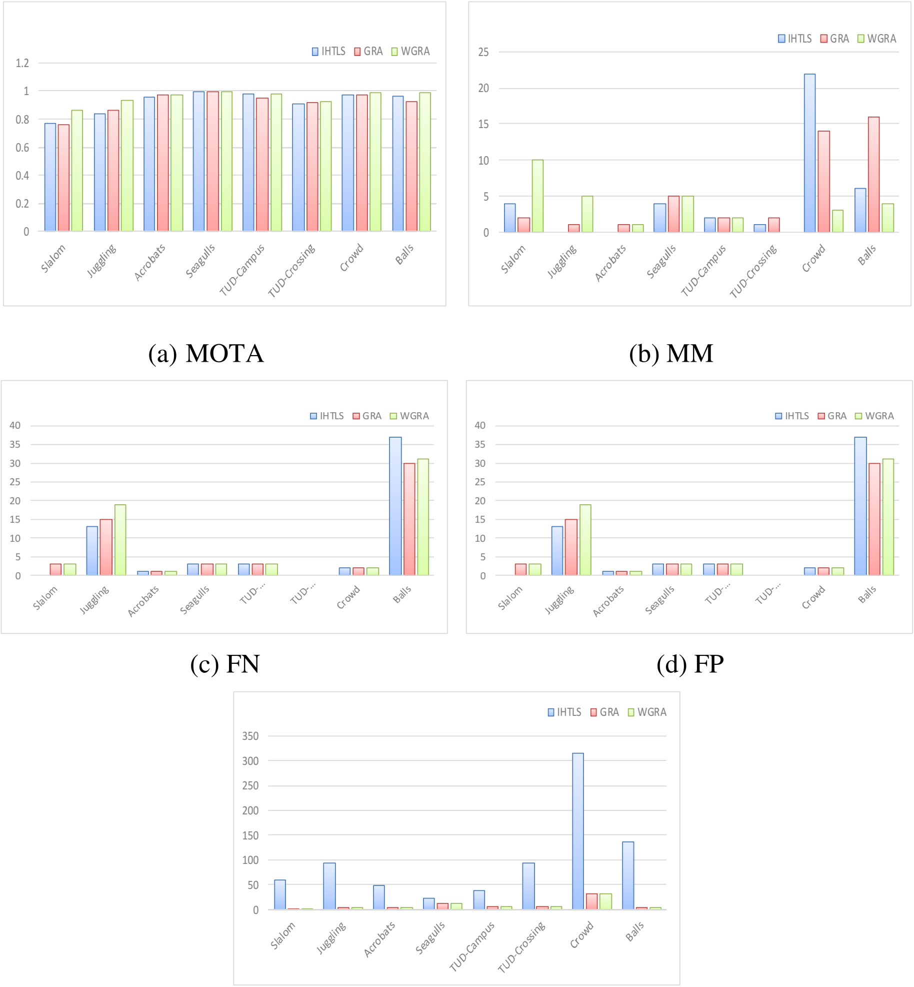

(a) Results of Crowd

(b) Results of Seagulls

(c) Results of TUD-campus

(d) Results of TUD-crossing

Fig. 1. The tracking results of proposed algorithm



(a) MOTA (b) MM

(c) FN (d) FP

(b) TF

Fig. 2. The metrics comparison of three algorithms

Fig. 2 shows a quantitative comparison of three algorithms. From the four histograms, we can see that WGRA performs best on MOTA, FP and MM while GRA is the optimal algorithm on FN. Furthermore, the average TF of WGRA is about 8 ms, 12 times shorter than that of IHTLS.

We also found that WGRA had excellent performance in distinguishing alternating movement (such as in TUD-crossing) because the corrected degree of incidence with the different direction of movement. However, for a single object performing back and forth movement (such as in Juggling), WGRA would handle it as two separate objects, but GRA could have better performance. So the calculation method of grey correlation influenced the tracking results largely.

5 Conclusion and Future Work

Under the influence of mutual occlusion, similar objects and other factors, the appearance features are not stable and discriminative. In this paper, we use motion features only to track similar appearance objects. The data association problem was resolved by generalized linear assignment optimization. To measure the similarity of tracklets, grey relational analysis is adopting to calculate the similarity on motion feature. Furthermore, we associate the discrete tracklets in a hierarchical processing way. Our method could deal with occlusion and camera motion effectively without appearance features, and is applicable to multi-object tracking in complicated scenes. Comparison experimental results with IHTLS show that our algorithm has obvious advantages. The average tracking accuracy rate reached 95% and the processing time per frame was just 8 ms. Our method can be applied in the offline video analysis. In the future, we will study new calculation method of grey degree of incidence and apply it to online tracking.

References

1. Nawaz, T., Poiesi, F., Cavallaro, A.: Measures of effective video tracking. IEEE Transactions on Image Processing **23**(1), 376–388 (2014)
2. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Computing Surveys (CSUR) **38**(4), 1–46 (2006)
3. Wang, X.: Intelligent multi-camera video surveillance: A review. Pattern Recognition Letters **34**(1), 3–19 (2013)
4. Zhang, L., Li, Y., Nevatia, R.: In: 2008 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, p. 1, June 23–28, 2008
5. Wen, L., Li, W., Yan, J., et al.: Multiple target tracking based on undirected hierarchical relation hypergraph. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, pp. 1282–1289. IEEE Computer Society (2014)
6. Dicle, C., Campsm O.I., Sznaier, M.: The way they move: tracking multiple targets with similar appearance. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), New Jersey, pp. 2304–2311. IEEE (2013)
7. Gold, S., Rangarajan, A.: Softmax to softassign: Neural network algorithms for combinatorial optimization. Journal of Artificial Neural Networks **2**(4), 381–399 (1996)
8. Jiang, B.C., Tasi, S.L., Wang, C.C.: Machine vision-based gray relational theory applied to IC marking inspection. IEEE Transactions on Semiconductor Manufacturing **15**(4), 531–539 (2002)
9. Lin, Y., Pang, J., Li, Y.: A New Grey Relational Fusion Algorithm Based on Approximate Antropy. Journal of Computational Information Systems **9**(20), 8045–8052 (2013)
10. Julong, D.: Introduction to grey system theory. The Journal of Grey System **1**(1), 1–24 (1989)
11. Naiming, X., Sifeng, L.: Research on evaluations of several grey relational models adapt to grey relational axioms. Journal of Systems Engineering and Electronics **20**(2), 304–309 (2009)

12. Cox, L.J., Hingorani, S.L.: An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. PAMI **18**(2), 138–150 (1996)
13. Ma, Y.: Study on the second grey relational grade and its properties. Kybernetes **39**(8), 1330–1335 (2010)
14. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP Journal on Image and Video Processing **2008** (2008)

Codebook Cardinality Spectrum of Distributed Arithmetic Codes for Nonuniform Binary Sources

Yayun Liu and Yong Fang(✉)

College of Information Engineering, Northwest A&F University,
Yangling 712100, Shaanxi, China
{yayunliu526,yfang79}@gmail.com

Abstract. The codebook cardinality spectrum (CCS) of distributed arithmetic codes (DAC) provides an effective tool for analyzing the decoding complexity of the full-search DAC decoder. However, in our previous work, the study on CCS is limited to equiprobable binary sources. In this paper, by defining the CCS for nonuniform binary sources, we try to provide a more general theoretical analysis on this complex problem. At first, a general explicit form of the initial CCS is obtained by utilizing the Fourier transform. Then, we get a recursive equation for deducing the CCS forwardly. For the convenience of computation, the paper also provides a numerical method for calculating the CCS. Finally, we define the expansion factor for nonuniform binary sources to measure the complexity of the full-search DAC decoder.

Keywords: Codebook cardinality spectrum (CCS) · Distributed arithmetic codes (DAC) · Nonuniform binary sources

1 Introduction

Distributed source coding (DSC) considers a situation in which two (or more) statistically dependent data sources must be encoded by separate encoders. Conventionally, channel code can be used to implement DSC, *e.g.* turbo codes [1] and low-density parity-check (LDPC) codes [2], *etc.* By contrast, DSC is implemented by entropy coding, such as distributed arithmetic coding (DAC). DAC can be seen as an extension of the classic AC by allowing overlapped intervals [3–5]. There are also some variants of DAC, *e.g.* time-shared DAC (TSDAC) [6], and rate-compatible DAC [7], *etc.* TSDAC is proposed to deal with the symmetric DSC problem. The rate-compatible DAC is proposed to realize rate-incremental DSC.

Though distributed arithmetic coding (DAC) is an effective implementation of Slepian-Wolf coding, its performance and decoding complexity have not been analyzed thoroughly. In [8], we provide a tool named codebook cardinality spectrum (CCS) and make some progresses in the analysis of the complexity of full-search DAC decoder. However, it only applies to binary sources with

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 458–467, 2015.
DOI: 10.1007/978-3-662-48570-5_44

equiprobable symbols. In this paper, we provide a more general tool by defining the CCS for nonuniform binary sources.

The contributions of this paper are as follows. Firstly, we define $g_i(u)$ as the stage-i CCS and formulate the initial CCS $g_0(u)$ by a functional equation. The Fourier transform is exploited to obtain the general explicit form of $g_0(u)$. Then the paper formulates the CCS evolution by an equation through which $g_{i+1}(u)$ can be deduced from $g_i(u)$, and proves that $g_i(u)$ will converge to the uniform distribution asymptotically as i approaches infinity. In addition, a numerical method is proposed to compute the CCS, whose convergency is proved. Secondly, the paper answers the complexity problem of the full-search binary distributed arithmetic coding (BDAC) decoder by introducing the expansion factor. The stage-i expansion factor γ_i is defined as the ratio of the number of stage-$(i+1)$ nodes to that of stage-i nodes in the decoding trees created by the full search DAC decoder. It is shown that γ_i can be related to $g_i(u)$ by a functional equation.

The rest of this paper is arranged as follows. Section 2 gives the definition of the initial CCS and extends it to a more general explicit form. In this section, we also give the evolution of the CCS and define the expansion factor, and relate expansion factor to the CCS. Section 3 provides a numerical method to calculate the CCS. Section 4 presents some representative experimental results. At last, we give some conclusions in Section 5.

2 CCS of DAC for Nonuniform Binary Sources

Let $X \sim p_X(x)$ for $x \in \mathbb{B} \overset{\Delta}{=} \{0, 1\}$ be the infinite-length binary sources at the encoder with *bias probability* $p_X(1) = p$, and $p_X(0) = 1 - p$. At the BDAC encoder, the symbols 0 and 1 are iteratively mapped onto overlapped subintervals $[0, (1-p)^\alpha)$ and $[(1-p^\alpha), 1)$ respectively, where $0 < \alpha \leq 1$. The parameter α controls the length of overlapped part, we call it *overlapping factor*. The resulting rate is $R = -p \log_2 p^\alpha - (1-p)\log_2(1-p)^\alpha = \alpha H(X)$. We consider p as the extrinsic metric computation factor at the decoder.

We define $g_i(u)$ as the stage-i CCS, *i.e.* this probability density function (PDF) of random variables u at all stage-i nodes in all incomplete binary trees. This function must satisfy constraints

$$\int_0^1 g_i(u)du = 1 \tag{1}$$

and

$$\begin{cases} g_i(u) \geq 0, & 0 \leq u < 1 \\ g_i(u) = 0, & u < 0 \text{ or } u \geq 1. \end{cases} \tag{2}$$

For the infinite-length nuniform binary sources, the initial CCS is constrained by

$$2qf(u) = f(\frac{u}{q}) + f(\frac{u-(1-q)}{q}). \tag{3}$$

We firstly give the initial CCS for infinite-length nonuniform binary sources. For the stage-0 CCS, $f(u)$ is used instead of $g_0(u)$ below for simplicity.

2.1 Initial CCS of Infinite-Length Nonuniform Binary Sources

In this subsection, we give the definition of CCS of Classic AC. Moreover, the initial CCS is defined and we extend it to a more general explicit form.

Theorem 1. ***Initial CCS*** *The initial CCS of infinite-length nonuniform binary sources is constrained by*

$$qf(u) = (1-p)f(\frac{u}{q}) + pf(\frac{u-(1-q)}{q}). \tag{4}$$

Proof. The source X is divided into two subsets X_0 and X_1. X_0 and X_1 are the subset of all infinite-length binary sequences beginning with symbol 0 and symbol 1 respectively. We define $f_0(u)$ ($f_1(u)$, resp.) to be the initial CCS of X_0 (X_1 , resp.). Considering the bias probability p, we can obtain

$$\begin{aligned} f(u) &= p_X(0)f_0(u) + p_X(1)f_1(u) \\ &= (1-p)f_0(u) + pf_1(u). \end{aligned} \tag{5}$$

We discard the first symbol 0 of each binary sequence in X_0, and a new subset $\hat{X}_0$ is emerged. As both $\hat{X}_0$ and X are the sets of all infinite-length binary sequences with the same bias probability, they are equivalent to each other and the initial CCS of $\hat{X}_0$ and X must have the same shape. The only difference between the two sets is that $\hat{X}_0$ is mapped onto interval $[0, q)$, so $f_0(u)$ is defined over $0 \leq u < q$. Then we can get

$$f_0(u) = \frac{f(u/q)}{\int_0^q f(u/q)du} = \frac{f(u/q)}{q}, \quad 0 \leq u < q. \tag{6}$$

Similarly, we can know

$$f_1(u) = \frac{f((u-(1-q))/q)}{q}, \quad (1-q) \leq u < 1. \tag{7}$$

On substituting (6) and (7) into (5), we obtain (4).

Then we define the CCS of Classic AC and analyze the explicit form of the initial CCS with nonuniform binary sources.

Corollary 1. ***CCS of Classic AC*** *The CCS of Classic AC is* $f(u) = G_1(u - \frac{1}{2})$*, where gate function* $G_T(u)$ *is defined as*

$$G_T(u) = \begin{cases} 1/T, & -\frac{T}{2} \leq u < \frac{T}{2} \\ 0, & u < -\frac{T}{2} \ or\ u \geq \frac{T}{2} \end{cases} \tag{8}$$

Proof. From (2) and (4), we can obtain

$$qf(u) = (1-p)f\left(\frac{u}{q}\right), \quad 0 \leq u < (1-q). \tag{9}$$

The above equation can be rewritten as

$$f(u) = \frac{q}{(1-p)} f(qu), \quad 0 \le u \le \frac{1-q}{q}. \tag{10}$$

When (10) comes to Classic AC, *i.e.* $q = \frac{1}{2}$, we get $f(u) = \frac{1}{2(1-p)} f(u/2), 0 \le u < 1$. Recursively, $f(u) = \lim_{k\to\infty} \frac{1}{(2(1-p))^k} f(u/2^k), 0 \le u < 1$.

Theorem 2. ***General Explicit Form of Initial CCS*** *Inspired by the previous works in [8], we solve this problem with the help of Fourier transform. We firstly shift $f(u)$ left by $\frac{1}{2}$ to get $\hat{f}(u) = f(u + \frac{1}{2})$. The general explicit form of initial CCS is*

$$\hat{f}(u) = \mathscr{F}^{-1} \left\{ \prod_{k=0}^{\infty} \left(\cos\left(\frac{\omega(1-q)q^k}{2} \right) + i(1-2p) \sin\left(\frac{\omega(1-q)q^k}{2} \right) \right) \right\}. \tag{11}$$

Proof. From (4), we obtain

$$q\hat{f}(u) = (1-p)\hat{f}\left(\frac{u + \frac{1-q}{2}}{q} \right) + p\hat{f}\left(\frac{u - \frac{1-q}{2}}{q} \right).$$

Fourier transform is applied on both side of the above equation, and

$$q\mathscr{F}\{\hat{f}(u)\} = (1-p)\mathscr{F}\left\{ \hat{f}\left(\frac{u + \frac{1-q}{2}}{q} \right) \right\} + p\mathscr{F}\left\{ \hat{f}\left(\frac{u - \frac{1-q}{2}}{q} \right) \right\}.$$

Let $\hat{F}(\omega) \triangleq \mathscr{F}\{\hat{f}(u)\}$, then

$$q\hat{F}(\omega) = (1-p)q\hat{F}(q\omega)e^{i\omega(1-q)/2} + pq\hat{F}(q\omega)e^{-i\omega(1-q)/2},$$

where i denotes the imaginary unit. Thus

$$\begin{aligned}\hat{F}(\omega) =& \hat{F}(q\omega) \left(\cos\left(\frac{\omega(1-q)}{2} \right) + i(1-2p) \sin\left(\frac{\omega(1-q)}{2} \right) \right) \\ =& \hat{F}(\omega q^k) \prod_{k'=0}^{k-1} \left(\cos\left(\frac{\omega(1-q)q^{k'}}{2} \right) + i(1-2p) \sin\left(\frac{\omega(1-q)q^{k'}}{2} \right) \right) \\ =& \hat{F}(0) \prod_{k=0}^{\infty} \left(\cos\left(\frac{\omega(1-q)q^k}{2} \right) + i(1-2p) \sin\left(\frac{\omega(1-q)q^k}{2} \right) \right).\end{aligned}$$

Remind $\hat{F}(0) = \int_{-\infty}^{\infty} \hat{f}(u)du = 1$, hence

$$\hat{F}(\omega) = \prod_{k=0}^{\infty} \left(\cos\left(\frac{\omega(1-q)q^k}{2} \right) + i(1-2p) \sin\left(\frac{\omega(1-q)q^k}{2} \right) \right).$$

The general explicit form of initial CCS is the inverse Fourier transform of $\hat{F}(\omega)$.

If $p = 0.5$, $(1-2p)\sin(\frac{\omega(1-q)q^k}{2}) \equiv 0$ and (11) is simplified to the general explicit form of initial CCS for equally-likely symbols in [1]. As resulting rate R goes to infinitesimal, q approaches 1. If $q = 1$, then $\cos(\frac{\omega(1-q)q^k}{2}) \equiv 1$ and $\sin(\frac{\omega(1-q)q^k}{2}) \equiv 0$, so $\hat{F}(\omega) \equiv 1$. Thus $\hat{f}(u) = \mathscr{F}^{-1}\{1\} = \delta(u)$.

2.2 Evolution of CCS and Ultimate CCS

Theorem 3. ***Evolution of CCS*** *The stage-$(i+1)$ CCS $g_{i+1}(u)$ can be deduced from stage-i CCS $g_i(u)$ by*

$$g_{i+1}(u) = \frac{q(g_i(qu) + g_i(qu + (1-q)))}{1 + \int_{1-q}^{q} g_i(u)du}. \tag{12}$$

Proof. We divide all stage-$(i+1)$ nodes in DAC decoding tree into two sets. The set of 0-branch nodes (1-branch nodes, resp.) contains all nodes coming from their parents though 0-branches (1-branches, resp.), and the CCS of this set is denoted by $g_{i+1}^0(u)$ ($g_{i+1}^1(u)$, resp.). Considering bias probability p, we obtain

$$g_{i+1}(u) = g_{i+1}^0(u) + g_{i+1}^1(u). \tag{13}$$

For $g_{i+1}^0(u)$, the variables u falls into interval $[0, q)$ at the parents of 0-branch nodes. After renormalization, interval $[0, q)$ at stage-i nodes are mapped onto interval $[0, 1)$ at stage-$(i+1)$ nodes. Correspondingly, $g_i(u)(\ 0 \le u < q)$ at stage-i nodes will be mapped onto $g_i(qu)(\ 0 \le u < 1)$ at stage-$(i+1)$ nodes. We get

$$g_{i+1}^0(u) = g_i(qu). \tag{14}$$

With the similar analysis, we get

$$g_{i+1}^1(u) = g_i(qu + (1-q)). \tag{15}$$

To normalize the stage-$(i+1)$ CCS, $g_{i+1}(u)$ will be divided by a normalization coefficient defined as

$$\begin{aligned}\int_0^1 g_i(qu)du + \int_0^1 g_i(qu + (1-q))du &= \frac{\int_0^q g_i(u)du + \int_{1-q}^1 g_i(u)du}{q} \\ &= \frac{1 + \int_{1-q}^q g_i(u)du}{q}.\end{aligned} \tag{16}$$

Substituting (14), (15) and (16) into (13), we get (12).

Corollary 2. ***Ultimate CCS*** *The evolution function of CCS is just the same as its in [8], through the same proof, we can obtain that the $g_i(u)$ will still converge to the shifted unit gate function $G_1(u - \frac{1}{2})$.*

2.3 Expansion Factor

This subsection will answer the complexity problem of the full search BDAC decoder by introducing the *expansion factor*, a terminology describing the speed that branches increase during the decoding process.

Definition 1. ***Expansion Factor*** *We define the stage-i expansion factor γ_i as the ratio of the number of stage-$(i+1)$ nodes to that of stage-i nodes in the decoding trees created by the full search DAC decoder.*

Corollary 3. ***Relation of Expansion Factor to CCS*** *The stage-i expansion factor γ_i can be related to the stage-i CCS $g_i(u)$ by $\gamma_i = 1 + \int_{1-q}^{q} g_i(u)du$.*

Proof. At stage-i nodes, two branches are created if u falls into $[(1-q), q)$; otherwise only one branch is created. Hence

$$\begin{aligned}\gamma_i &= \int_0^{1-q} g_i(u)du + 2\int_{1-q}^{q} g_i(u)du + \int_q^1 g_i(u)du \\ &= 1 + \int_{1-q}^{q} g_i(u)du.\end{aligned}$$

Corollary 4. ***Ultimate Expansion Factor*** *As i approaches infinity, expansion factor γ_i will converge to $2q$.*

Proof. $\lim_{i\to\infty} \gamma_i = 1 + \int_{1-q}^{q} h(u)du = 2q$.

3 Numerical Calculation of CCS

Usually, It is very complex to calculate the explicit form of initial CCS $f(u)$ directly by 11 as it contains the product of infinite terms. Hence, we propose a numerical method to calculate $f(u)$, whose convergency is proved. And then a numerical method is put forward to mimic the evolution of CCS. Before introducing the algorithms, let us define: $round(x) \triangleq \lfloor x + 0.5 \rfloor$.

3.1 Numerical Calculation of Initial CCS

1. **Discretization**. The interval $[0, 1)$ is divided into N equal segments. If N is large enough, then $f(u)$ can be approximated by $f(n/N)$, where $n \in \{0, \cdots, N-1\}$. For simplicity, $f(n/N)$ is abbreviated to $f(n)$ below.
2. **Initialization**. Let $f^{(t)}(n)$ be the approximate of $f(n)$ after t iterations. Initially, $f^{(0)}(n)$ can be arbitrarily set, provided that $\sum_{n=0}^{N-1} f^{(0)}(n) = N$.
3. **Iteration**. Let $L = round(N(1-q))$ and $H = round(Nq)$.
   - $0 \le n < L$: This corresponds to interval $[0, (1-q))$, thus
     $$f^{(t)}(n) = \frac{(1-p)f^{(t-1)}(round(\frac{n}{q}))}{q}.$$
   - $L \le n < H$: This corresponds to interval $[(1-q), q)$, thus
     $$f^{(t)}(n) = \frac{(1-p)f^{(t-1)}(round(\frac{n}{q})) + pf^{(t-1)}(round(\frac{n-L}{q}))}{q}.$$
   - $H \le n < N$: This corresponds to interval $[q, 1)$, thus
     $$f^{(t)}(n) = \frac{pf^{(t-1)}(round(\frac{n-L}{q}))}{q}.$$

4. **Normalization**. Recall $\int_0^1 f(u)du = 1$. Hence, $f^{(t)}(n)$ should be normalized by

$$f^{(t)}(n) = \frac{Nf^{(t)}(n)}{\sum_{n=0}^{N-1} f^{(t)}(n)}.$$

5. **Termination**. The mean squared error (MSE) between two successive iterations is used to terminate the iteration. Let $\varDelta$ be a small quantity. The iteration is terminated if

$$\text{MSE}^{(t)} = \frac{1}{N}\sum_{n=0}^{N-1}(f^{(t)}(n) - f^{(t-1)}(n))^2 < \varDelta.$$

3.2 Numerical Estimation of the CCS

1. **Discretization**. We divide the interval $[0, 1)$ into N equal segments. If N is large enough, then $g_i(u)$ can be approximated by $g_i(n/N)$, where $n \in \{0, \cdots, N-1\}$. For simplicity, $g_i(n/N)$ is abbreviated to $g_i(n)$ below.
2. **Initialization**. We firstly set $g_0(n) = f^{(t)}(n)$, where $f^{(t)}(n)$ is the numerical result of $f(u)$.
3. **Iteration**. $g_i(n)$ can be deduced from $g_{i-1}(n)$ recursively by

$$g_i(n) = g_{i-1}(round(nq)) + g_{i-1}(round(nq + N(1-q))).$$

4. **Normalization**. As $\int_0^1 g_i(u)du = 1$, $g_i(n)$ must be normalized by

$$g_i(n) = \frac{Ng_i(n)}{\sum_{n=0}^{N-1} g_i(n)}.$$

5. **Expansion Factor**. Let $L = round(N(1-q))$ and $H = round(Nq)$. The stage-i expansion factor can be approximated by

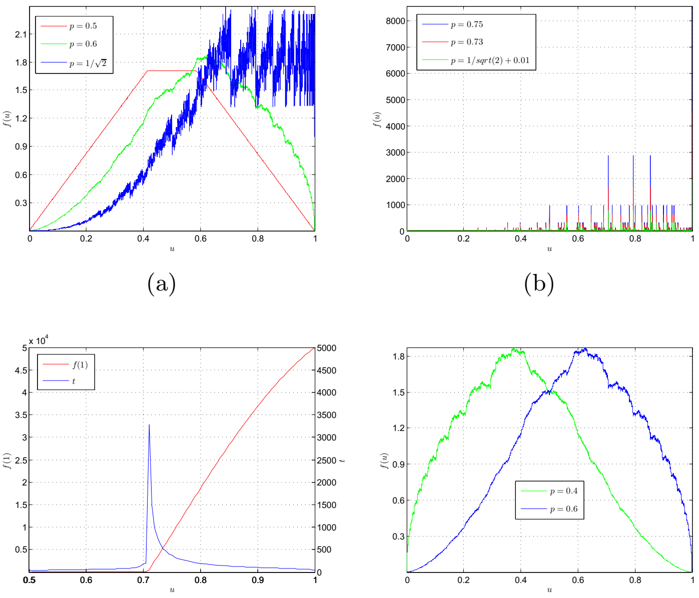
$$\gamma_i = 1 + \frac{\sum_{n=L}^{H-1} g_i(n)}{N}. \qquad (17)$$

4 Examples

Some experimental results will be provided below for verifying our theoretical analyses. Fig. 1 and Fig. 2 show how the shape of $f(u)$ changes with respect to p. Fig. 3(a) shows the convergence of ultimate CCS. Experimental results about the expansion factor are showed by Fig. 3(b).

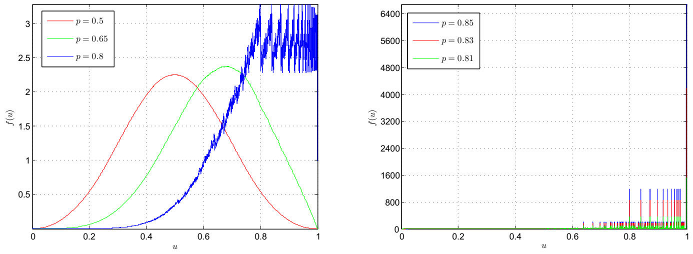
Fig. 1(a)-(d) give some examples of $f(u)$, which are obtained by the numerical algorithm with $N = 10^5$, $q = \frac{1}{\sqrt{2}}$, and $f^{(0)}(n) \equiv 1$, $0 \leq n < N$. To achieve no perceptible difference between $f^{(t)}(n)$ and $f(u)$, we set the threshold of successive MSE to 10^{-10} in these four simulations.

Fig. 1(a) shows three examples of convergent $f^{(t)}(n)$ by the numerical algorithm with respect to p. Before iteration termination, 38, 57, and 105 iterations



(a) (b) (c) (d)

Fig. 1. Examples of numerical approximates to $f(u)$ when $q = \frac{1}{\sqrt{2}}$. (a) Three examples for $p \in [0.5, \frac{1}{\sqrt{2}}]$. (b) Three examples for $p > \frac{1}{\sqrt{2}}$. (c) The value of $f(1)$ and iteration times t for $p \in [0.5, 1]$. (d) Shape of $f(n)$ for $p = 0.4$ and $p = 0.6$.



(a) (b)

Fig. 2. Examples of numerical approximates to $f(u)$ when $q = 0.8$. (a) Three examples for $p \in [0.5, 0.8]$. (b) Three examples for $p > 0.8$.

are run for $p = 0.5$, $p = 0.6$, and $p = \frac{1}{\sqrt{2}}$, respectively. With the increase of p, the shap of $f(u)$ changes greatly.

Fig. 1(b) includes three examples for $p > \frac{1}{\sqrt{2}}$. It can be seen that though p differs by only 0.01, the shape of $f(u)$ is very different for $p = \frac{1}{\sqrt{2}}$ and $p =$

$\frac{1}{\sqrt{2}}+0.01$. 1222, 617, and 366 iterations are run for $p = \frac{1}{\sqrt{2}}+0.01$, $p = 0.73$, and $p = 0.75$, respectively. As p becomes bigger, there is a obvious Dirac delta spike at $u = 1$.

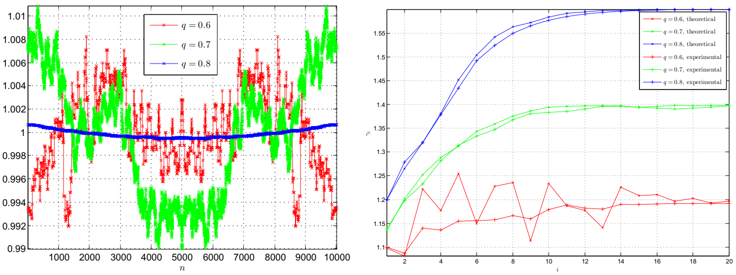
Fig. 1(c) shows how the value of $f(1)$ and iteration times t change for $p \in [0.5, 1]$. As we set the increase step size for p to 0.005, the first point belongs to interval $(\frac{1}{\sqrt{2}}, 1]$ is 0.71. It can be observed from Fig. 1(c) that the value of $f(1)$ increasing when $p > 0.71$ and the maximal value 5×10^4 can be reached when $p = 1$. When $p = 0.71$, the maximal iteration times 3276 is needed to cease the iteration.

Fig. 1(d) includes two examples for $p = 0.4$ and $p = 0.6$. Similar results can be found to verify that $f(u)$ shows symmetric shapes around axis $u = 0.5$ when p changes to $(1 - p)$.

Fig. 2(a) and (b) give some examples when $q = 0.8$. The shape change of $f(u)$ when $q = 0.8$ is similar with that when $q = \frac{1}{\sqrt{2}}$, but the watershed which affects the shape of $f(u)$ for a fixed q changes from $p = \frac{1}{\sqrt{2}}$ to $p = 0.8$. The shape of $f(u)$ is very different for $p = 0.8$ and $p = 0.81$. Due to the symmetry as shown in Fig. 1(d), we assume that there are two watersheds $p = q$ and $p = (1 - q)$ when q is fixed.

Fig. 3(a) shows the ultimate CCS converges to the shifted unit gate function $G_1(u - \frac{1}{2})$ after 20 iterations.

Fig. 3(b) shows the theoretical and experimental results of expansion factor. The theoretical results are computed by (17). As for experimental results, we encode nonuniform binary sources by a 31-bit DAC encoder. The stage-i expansion factor γ_i equals the ratio of the number of stage-$(i + 1)$ nodes to that of stage-i nodes. From Fig. 3(b), we can find that the experimental results agree well with the theoretical results.



(a) (b)

Fig. 3. (a) The CCS after 20 iterations. The interval $[0, 1)$ is divided into 10000 equal segments. (b) Theoretical and experimental results of expansion factor for $q = 0.6$, 0.7, and 0.8.

5 Conclusion

This paper defines the CCS for nonuniform binary sources. The general explicit form of the initial CCS is deduced by using Fourier transform. Furthermore, we define the expansion factor and link it with the CCS to measure the complexity of the full-search DAC decoder. To calculate the CCS and the expansion factor, a numerical method is proposed, whose convergency is proved. With the help of CCS, we can analyze the complexity of the full-search DAC decoder. Moreover, some representative examples are given to verify our theoretical analyses.

References

1. Garcia-Frias, J., Zhao, Y.: Compress of correlated binary sources using turbo codes. IEEE Commun. Lett. **5**(10), 417–419 (2001)
2. Liveris, A., Xiong, Z., Georghiades, C.: Compression of binary sources with side information at the decoder using LDPC codes. IEEE Commun. Lett. **6**(10), 440–442 (2002)
3. Grangetto, M., Magli, E., Olmo, G.: Distributed arithmetic coding. IEEE Commun. Lett. **11**(11), 883–885 (2007)
4. Grangetto, M., Magli, E., Olmo, G.: Distributed arithmetic coding for the Slepian-Wolf problem. IEEE Trans. Signal Process. **57**(6), 2245–2257 (2009)
5. Artigas, X., Malinowski, S., Guillemot, C., Torres, L.: Overlapped quasi-arithmetic codes for distributed video coding. In: Proc 2007 IEEE ICIP, vol. 2, pp. 9–12
6. Grangetto, M., Magli, E., Olmo, G.: Symmetric distributed arithmetic coding of correlated sources. In: Proc. 2007 IEEE MMSP, pp. 111–114
7. Grangetto, M., Magli, E., Tron, R., Olmo, G.: Rate-compatible distributed arithmetic coding. IEEE Commun. Lett. **12**(8), 575–577 (2008)
8. Fang, Y.: DAC spectrum of binary sources with equally-likely symbols. IEEE Trans. Commun. **61**(4), 1584–1594 (2013)

Author Index