Multimed Tools Appl (2019) 78:27045–27065
DOI 10.1007/s11042-017-4739-1

Large scale product search with spatial quantization and deep ranking

Shuhan Qi[1] · Zawlin Kyaw[2] · Xuan Wang[1] · Zoe L. Jiang[1] · Jian Guan[1]

Received: 16 January 2017 / Revised: 4 April 2017 / Accepted: 19 April 2017 /
Published online: 1 May 2017

Abstract Product image search aims to retrieve similar product images based on a query image. While deep learning based features work well in retrieving images of the same category (e.g. "searching for T-shirts from all the clothing images"), they perform poorly when retrieving variants of images within the same category (e.g. "searching for uniform of Chelsea football club from all T-shirts image"), since it requires fine-grained matching on image details. In this paper, we present a spatial quantization approach that utilizes spatial pyramid pooling (SPP) and vector of locally aggregated descriptors (VLAD) to extract more discriminative features for instance-aware product search. By using the proposed spatial quantization, spatial information is encoded into the image feature to improve the fine grained product image search. We also present an triplet learning to rank method to finetune the deep learning model on product image search task. Finally, the experiments conducted on a large scale real world dataset provided by Alibaba large-scale image search challenge (ALISC) demonstrate the effectiveness of our method.

✉ Xuan Wang
wangxuan@cs.hitsz.edu.cn

Shuhan Qi
shuhanqi@gmail.com

Zawlin Kyaw
kzl.zawlin@gmail.com

Zoe L. Jiang
zoeljiang@gmail.com

Jian Guan
j.guan@cs.hitsz.edu.cn

[1] Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China

[2] School of Computing, National University of Singapore, Singapore, Singapore

Keywords Product image retrieval · Deep learning · Spatial quantization · Triplet metric learning · Salient region detection

1 Introduction

Widespread proliferation of smart phones coupled with fast growing acceptance for online shopping[1] has prompted most online retailers[2] to include "search by image" feature. This allows a user to find a product easier when it is difficult to describe the product textually such as when the user is looking for a specific design style and pattern. In general, there are two key differences between product image search and traditional Content Based Image Retrieval (CBIR) [33]. First, product image can be taken under unprofessional conditions such as various lighting conditions, viewpoints and cluttered background (such as the top and middle rows of Fig. 1). These consumer level images significantly increase the visual diversity of the database. Second, the intra-category variance within each product category is very small. In other words, the product images in the same category is visually very similar to each other, (such as the bottom row of Fig. 1) and the task requires to return the products exactly the same with the query instance from such small intra-variance categories. In this way, making a instance aware product image search is a challenge task. These characteristics requires a robust and discriminative image representation.

Early work in image retrieval research, which mostly based on hand designed features such as SIFT [21], and bag-of-words based model [40], works well for retrieving near-duplicate images and highly-textured objects, but not for our task. Further, the product images are always weakly supervised, which means the product only sizes a small region in a image without annotation. However, common weakly supervised trained methods [55, 57, 59] are not suitable for our task, as most of them are designed for category level image retrieval.

Deep learning techniques, which attempt to model high-level abstractions in data by employing deep architectures composed of multiple nonlinear transformations [35, 45], have been very successful in CBIR task [2, 38, 46].

The deep learning techniques demonstrate great robustness in handling the consumer level image problem, and work well in image retrieval problem for categories (e.g. searching for T-shirt from the clothing images) [6]. However, addressing the instance-aware product image search (e.g. searching for uniform of Chelsea football club from all T-shirt images) by deep learning techniques is still a big challenge. As the small intra-category variance of product images, many images with different style may belong to the same category with similar appearance. The instance-aware product image search requires fine-grained matching on image details. Since the deep learning models abstract global image appearance into much higher level features, it is difficult for deep learning techniques to discriminate small local details in the images. Therefore, the instance-aware product image search needs deep learning based features with better discriminative ability.

In this paper, we propose a spatial quantization approach to enhance image representation and improve instance-aware product image search. Unlike most of the deep learning based image ranking models that use the fully connection layer outputs as image representation, we utilize the outputs of convolution layer. By using the spatial pyramid pooling (SPP) and

[1] http://www2.alizila.com/taobao-report-mobile-shopping-catching-quickly-china/.

[2] Amazon, Yahoo, Taobao.

Fig. 1 Illustrative examples of product image. Product images with various viewpoints (*top*); cluttering background (*middle*); product images belong to the same category with similar appearance but different instance (*bottom*)

vector of locally aggregated descriptors (VLAD), feature maps are encoded into a global descriptor while preserving spatial information. SPP enables more discriminative features by integrating global and local information. It also allows the features to be more resistant to deformation. VLAD further improves SPP features through quantization to reduce the influence of noise. Moreover, based on the observation that product is almost always the salient region of image, an efficient saliency based product region localization method is employed to reduce the impact of clutter background. The final distance between the query and the testing image is determined by combining the Euclidean distances of different types of feature linearly. A preliminary version of this work is reported in a conference paper [25], which only utilized some classification task pretrained deep learning models for spatial quantization. However, due to the intrinsic difference between image classification and similar image ranking tasks, a good pretrained deep learning model for image classification may not be optimal enough for distinguishing product image similarity. In this paper, we proposed an triplet learning to rank method to further improve the accuracy of product search. Specifically, we finetune the pretrained deep learning models by utilizing an triplet ranking loss function, and the final performance achieves an MAP@20 of 37.64 on ALISC dataset with the 3567 query images and 3 million evaluation images.

The main contributions of this paper can be summarized as follows:

1. We investigate the product image search problem and propose a practical framework for large scale product image search. In this framework, multiple deep learning models with triple loss are utilized to enhance the robustness and discriminability of image representation. Moreover, SPP and VLAD are utilized to encode the spatial information into the image representation.
2. We also propose an efficient saliency based product region localization algorithm to reduce the impact of clutter background in the images.

3. We conduct experiments on the Alibaba large scale image search dataset consisting of 4984 queries and 3.1 million testing images collected from E-commerce platform TAOBAO. The evaluation validates the proposed approach is able to achieve remarkable search performance for product image search.

2 Related work

2.1 Deep learning for image retrieval

Early work in image retrieval research is done mostly under the bag-of-words model adopted from text retrieval. Hand designed features such as SIFT [21] are extracted from the query image and histogram of visual words is constructed based on a predefined dictionary. This histogram is treated as a feature vector matched against database images. The performance of these models is largely bottleneck-ed by the representation power of hand crafted features.

Deep learning models have been very successful at classification tasks [10, 12, 35].

Indeed, the features learned from classification perform well as a strong baseline in image retrieval tasks [26]. Deep learning features have been utilized for CBIR in [2], which use a transfer learning deep neural works to learn the high-level image representation of queries.

In [38], by conducting a series of evaluations on various deep convolutional neural networks with application of CBIR, the authors suggested that the retrieval performance could be boosted significantly by deep learning. In [42], a convolutional network has been trained to output binary similarity decision from a pair of input images. However, it suffers from having to perform expensive forward pass when the similarity needs to be measured. An alternative approach is to extract the features in advance and calculate Euclidean distance between features as a measure of similarity. For product search in reality situation, query images are generally more uncontrolled and the system needs to provide the results fast enough in order to meet usability needs. In [33], Shen et al. computes a weighted mask to locate and split query object before extraction. Lin et al. [11] learns binary hash code with deep learning to reduce retrieval time cost.

Learning a distance metric that directly optimizes the distance in the embedding space has been well studied [3, 7]. A popular variant of this approach, Siamese network, which maps input patterns into a target space that the L1 norm distance in the target space approximates the semantic distance in the input space by minimizing a discriminative similarity loss function, has been applied to face verification [4, 36]. However, the Siamese network is sensitive to calibration in the sense that the notion of similarity vs dissimilarity requires context [8]. Triplet network improves the Siamese network by introducing the triplet which contains an anchor, an positive image and an negative image instead of a pair to induce the relative pairwise similarity ordering [39]. Zhang et al. [61] proposed an triplet-based hashing learning method for image retrieval, which incorporates a hashing related regularization term into the triplet metric learning to preserve the adjacency relation in Hamming distance.

2.2 Multimedia content analysis

Recently, many graph-based models are applied in multimedia and computer vision. They can be used as geometric image descriptors [57, 58] to enhance image categorization. These methods can also be used as image high-order potential descriptors of superpixels [52, 53].

Further, graph-based descriptors can be used as a general image structure descriptors to improve the results of fine-grained image categorization [48, 60]. Besides, many progress in related-technical such as image segmentation [50, 59], photo retargeting and cropping [49, 54, 57] may help to localize the product region in images, and some aesthetics based methods [51, 56] can be used to improve the quality of returning results. Some developments in cross-media retrieval [24, 43] and discrete image hashing [44], are also helpful to improve the accuracy of product image search.

With the increasing amount of the UGC, the requirement of searching specific content in UGC is ever growing [18, 22]. However, because of the complexity of UGC, deploying an accurate product search method in UGC is an challenge problem. Thus, some UGC oriented feature extraction and recognition methods have significant value of referential. Liu et al. [19] utilized Sift descriptors to model the visual contextual information for refinement of the video retrieval. Shah et al. [31, 32] leveraged contextual information such as mobile sensors to recommend matching soundtracks for user-generated videos on social media. The multimodal (both content and contextual) information access of such UGC benefits a number of diverse multimedia applications[27, 28]. For instance, it is advantageous in an effective semantics understanding such as event summarization [29] and sentics understanding [30] from large multimedia collections. In [13, 16] and [17], vision-based action recognition models are trained by multi-tasks learning to identify temporal patterns among actions and further utilize the identified patterns to represent activities for automated recognition. Moreover, multimodal information is exploited in determining multi-view structures such as tracking the human motion in video [5, 14] and predicting the water quality from ubiquitous sensor data [15, 20].

3 The proposed method

3.1 Architecture overview

As shown in Fig. 2, the procedure of the proposed framework contains four stages. First, given an image, an efficient saliency based object detection is utilized to localize the product region in the image. Second, the local descriptors in the product ROI (region of interest) of the image are extracted by multiple models. In this work, two deep learning models with spatial pyramid pooling (SPP), are employed to extract high level local descriptors, while sparse SIFT is employed as well to extract the low level local descriptors. Third, the extracted local descriptors with different modalities are encoded by vector of locally aggregated descriptors (VLAD) into global descriptors. Finally, Euclidean distances of the three extracted global descriptors between the testing image and the query are calculated, and the final distance is decided by optimal linear combination of the three distances.

3.2 Product region localization

In many E-commerce images, the product object occupies only a small area. Extracting features from the whole image will include cluttered background, which significantly reduces the image retrieval accuracy. According to our observation on the dataset, most of the product regions are the salient regions of E-commerce images. Inspired by the work in [47], we propose an efficient saliency based product region localization approach, as shown in Fig. 3. The generation of saliency map contains two steps: Fast Minimum Barrier Transform (fast MBD) and Image Boundary Contrast (IBC) Transform.

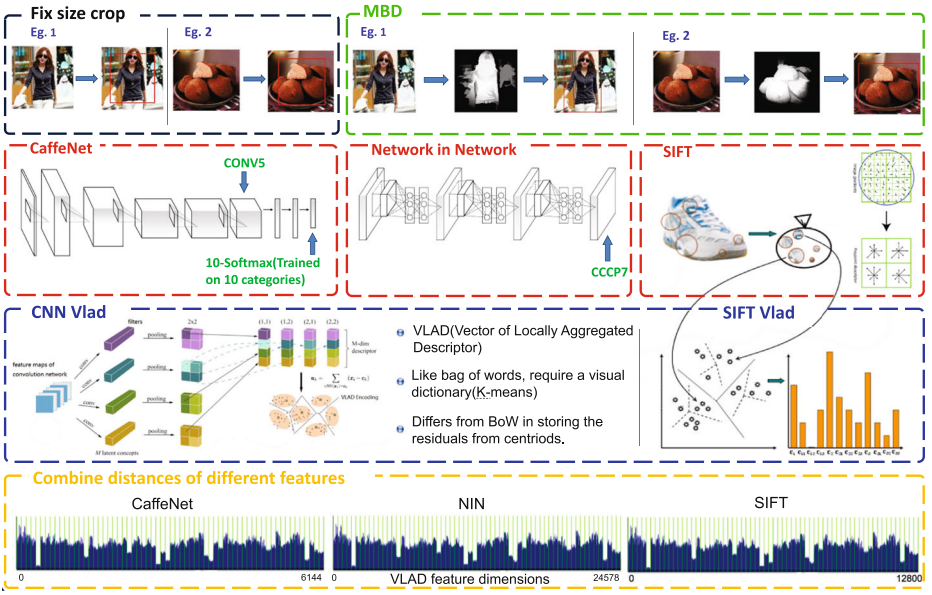


Fig. 2 The framework of the proposed method. The framework contains 4 stages: product localization, local descriptors extraction, VLAD encoding and distance calculation

3.2.1 Fast minimum barrier transform

Given an image A, the fast MBD transform aims to minimize the barrier distance in the image. To update the barrier distance of each pixel in the image, the fast MBD scans the whole image in forward (left to right, up to down) and backward (right to left, down to up) iteratively. In each scan, the barrier distance L of pixel x is updated as

$$L(\mathrm{x}) = \begin{cases} L(x) \\ \eta_y(x), \end{cases} \tag{1}$$

where $\eta_y(x)$ is the MBD cost and defined as:

$$\eta_y(x) = \max(P(y), A(x)) - \min(Q(y), A(x)). \tag{2}$$

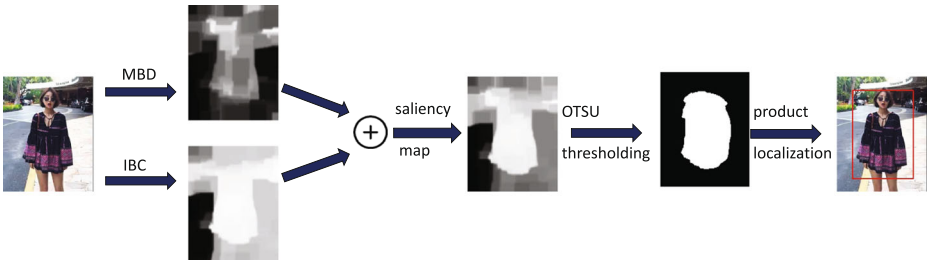


Fig. 3 The illustration of product localization in the image. In the product localization, two transforms MBD and IBC is made to form an enhanced saliency map, then the Otsu thresholding is preformed to generate bounding box of product

Note that y is an adjacent neighbor of current pixel x, and the P and Q are two auxiliary maps that keep track of the highest and lowest pixel values on the path for each pixel. Since in each iteration the barrier distance L is non-negative and non-increasing, the iterative update will converge in the end. The Fast MBD is an efficient saliency map transformation approach since the complexity is linear to the number of image pixels.

3.2.2 Image boundary contrast transform

The fast MBD transform is not stable in situation that the product region touches the image boundary. An Image Boundary Contrast (IBC) map [47] is generated to complement the fast MBD.

Assuming that the background regions are likely to possess similar appearance to the image boundary regions, the IBC map highlights regions with a high contrast against the boundary regions. The IBC takes four boundary regions (upper, lower, left and right) into consideration. For each region $t = \{1, 2, 3, 4\}$, the mean color of each color channel $\mathbf{m_t} = [m_l, m_a, m_b]$ and color covariance matrix $W_t \in R^{3\times3}$ are calculated. Then four intermediate IBC maps $v_t = [v_t^{ij}] \in R^{3\times3}$ are computed based on the Mahalanobis distance:

$$v_t^{ij} = \sqrt{(x_t^{ij} - \mathbf{m_t}) W_t^{-1} (x_t^{ij} - \mathbf{m_t})^T}. \tag{3}$$

The final IBC map $\mathrm{V} = [v^{ij}]$ is determined by:

$$v^{ij} = \left(\sum_{t=1}^{4} v_t^{ij} \right) - \max_t v_t^{ij}. \tag{4}$$

By considering the appearance of four boundary regions in the image, the IBC map will be more robust when one of the boundary regions is mostly occupied by product objects.

A linear combination of the MBD map U and IBC map V is made to form an enhanced saliency map $S = U + V$. Then we preform the Otsu thresholding [23] to generate a binary mask, and then use the size and aspect ratio as heuristics to select the final bounding box.

3.3 Feature representation

The accuracy of product image retrieval depends highly on quality of feature representation. Recent research efforts have shown that combining multiple features extracted by different models may yield a good image representation [37]. In this work, we use two deep learning based features (CaffeNet and Network in Network) and a hand crafted feature (SIFT) to make a better set of image features. It is worth noting that for Network in Network and SIFT, we use saliency based product location to detect the product region in the images, while for CaffeNet, the fixed size center crop is taken. We use the fixed size crop for CaffeNet because the CaffeNet also takes the response of general classification in our solution (to be elaborated in Section 3.4). Fixed size crop maintains better global information which is important for general classification.

VLAD has been proposed for image retrieval to encode image local descriptors such as SIFT and HOG. Inspired by the work in [41], in order to diversifiy the outputs with aggregation on multiple spatial locations at deeper stage of network, we accumulate the response of specific location of CNN convolution filter into high level local descriptors, and employ VLAD to encode the high level local descriptors as well as spatial information into a more discriminative image representation.

3.3.1 High level local descriptors

In this work, the high level local descriptors are extracted by two deep learning models respectively.

The first model is the CaffeNet model from CAFFE [9]. The model is pre-trained on ImageNet ILSVRC-12. The CaffeNet consists of 8 learned layers, which include 5 convolutional layers and 3 fully-connected layers, follows with some ReLU activation functions and max pooling layers. In this paper, the outputs of the last convolution layer with shape size $14 \times 14 \times 256$ are extracted.

The second model is the NIN-Imagenet model [12] from CAFFE. The NIN builds micro neural network with more complex structures to abstract the data within the receptive field. By using the micro neural network, people can use much less parameter to build a more complex deep learning model. The NIN model contains 12 learning convolutional layers, as well as several ReLU activation functions and pooling layers. The outputs of cccp7 layer with shape size of $6 \times 6 \times 1024$ are extracted.

The feature maps which are extracted from the convolution layer of CNN contain spatial information. The standard way of converting feature maps into feature vector, which flattens the feature maps into a vector, will leads to high dimensional feature and heavy computational cost. The convolutional filters can be regarded as the generalized linear classifiers of some underlying patterns, in this way, each convolutional filter corresponds to a latent concept. The latent concept descriptor, which consists of the activations of specific position in feature maps, is able to represent the responses from convolutional filters for a corresponding spatial location. In this work, the latent concept descriptor is adopted as the high level local descriptor. Specifically, as shown in Fig. 4, for the feature maps in shape $a \times a \times M$, where a is the size of each feature map and M is the number of convolution filters in the convolution layer (for CaffeNet $a = 14$ and $M = 256$, while for NIN $a = 6$ and $M = 1024$), the spatial pyramid pooling (SPP) layer is utilized to pool each feature map into three size: 1×1, 2×2, 3×3. Three groups of feature Maps with $1 \times 1 \times M$, $2 \times 2 \times M$, $3 \times 3 \times M$ are formed. Then 13 latent concept descriptors with length M are formed by accumulating the value of specific position of M feature maps in the same group.

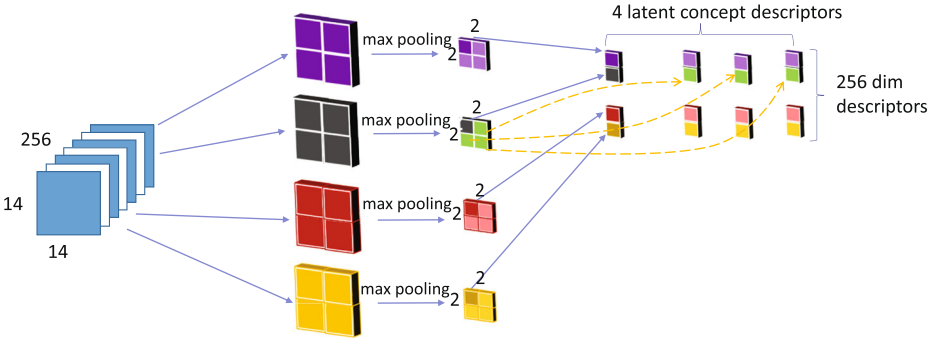


Fig. 4 The illustration of converting feature maps into latent concept descriptors. The feature maps are pooled into different size (2×2 in this example) by SPP layer, latent concept descriptors are formed by accumulating the value of specific position in different feature maps

3.3.2 Learning to rank

Most of the pretrained CNN models are trained to solve the category level classification problems, for example, in [10, 35], all the images belong to the same category are considered similar. However, such CNN models are insufficient to handle the product search task. As discussed in the Section 1, the product search requires a model can distinguish the small intra-category differences between two images in the same category. In the way, the classification task pretrained CNN models may not fit the task of product search. Inspired by [39], we employ a triplet metric learning model, which characterizes the product similarity relationships with a set of triplets, to further finetune the CNN model.

The triplet ranking model is a kind of pairwise ranking model, whose goal is to learn an embedding function $f(\cdot)$ that maps an image to a point in Euclidean space. The more similar the two images x_i and x_j is, the smaller the distance $D(f(x_i), f(x_j)) = \|f(x_i) - f(x_j)\|_2^2$ is. In the triplet metric ranking model, the training instance is composed by a triplet: the anchor sample which is usually randomly selected from training dataset, a positive sample which is "sharing the same label with an anchor sample, a negative sample is labeled different against the anchor. Compared with Siamese ranking model [4] only consider pairs globally, which can result in insufficient sampling of positive and negative samples, the triplet model utilizes two type of pairwise relationships: a similar pairwise and a dissimilar pairwise in the training process, and samples better potential pairs during optimization.

To be specificly, given a triplet object (x, x^+, x^-), where x, x^+ and x^- are the anchor, the positive sample and the negative sample respectively. The objective of triplet ranking optimization is to reduce the distance between the anchor and the positive sample and increase the distance between the anchor and the negative sample in each iteration. The loss function is defined as:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} [\|f(x) - f(x^+)\|_2 - \|f(x) - f(x^-)\|_2 + m] + \|\theta\|_2, \tag{5}$$

where m is a small margin that enforce minimum inter cluster distance in the embedding space.

Since the number of possible negative pairs are much larger than positive pairs, in this paper, the hard negative mining is utilized to avoid inefficient training on trivial examples. To further restrict the set of all possible models for faster convergence, two additional constraints are enforced here. First, we require the embedding space lies along the surface of a unit hypersphere, *i.e*, $\|f(x)\|_2 = 1$. Second, we enforce the embedding model to belong to the set of models that also perform well on softmax regression by multi-tasks training the softmax loss and the triplet loss.

3.3.3 Low level local descriptors

While CNNs learn high level semantic features very well, their performance degrades when the query object is rotated or when it is in cluttered scene. We extract rotation and scale invraint SIFT features [21] to complement the high level CNN features. We then extract SIFT descriptors at detected key points and used VLAD to encode them into final quantized features. We used opencv's implementation of SIFT and used $nOctaveLayers = 3$, $contrastThreshold = 0.04$, $edgeThreshold = 10$, $sigma = 1.6$.

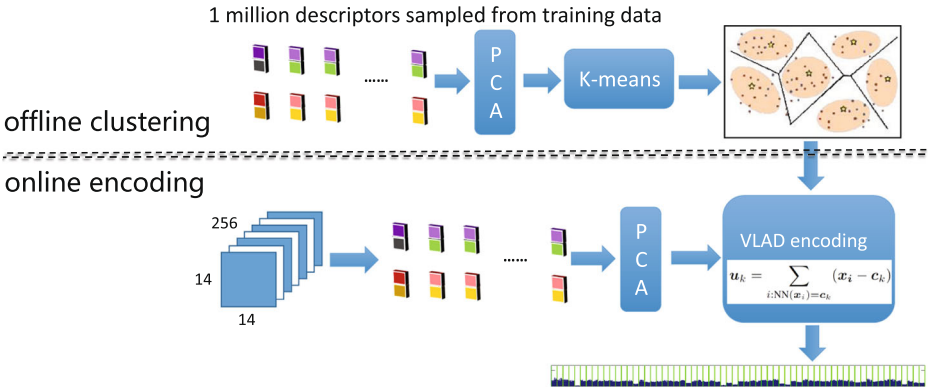


Fig. 5 The illustration of the VLAD procedure. The procedure contains the offline clustering and online encoding. In offline stage, the local descriptors are sampled, and PCA whitening and K-means clustering are performed. In the online stage, the local descriptors are encoded into the VLAD by accumulating the residual between descriptors and cluster centers

3.3.4 VLAD encoding

To encode the local descriptors into global feature representation, Vector of Locally Aggregated Descriptors (VLAD) is employed. As shown in Fig. 5, VLAD has two stages: offline clustering and online encoding. For offline clustering, the local descriptors are sampled from training images, then PCA whitening and K-means clustering are performed. Note that the PCA whitening pre-processing is necessary for a better fit on the diagonal covariance matrix assumption [34]. In this work, efficient K-means in VLFEAT[3] library is conducted to obtain K coarse centers $\{c_1, c_2, ..., c_K\}$, such K centers are then used as vocabulary in the online encoding stage. In the online encoding stage, the local descriptors with PCA pre-processing are assigned to p closest cluster of the vocabulary. For each clustering center in vocabulary, the residuals (vector differences between descriptors and clustering centers) are accumulated by

$$u_k = \sum_{c_k \in NN(x_i)} (x_i - c_k), \tag{6}$$

where $NN(x_i)$ is the p nearest neighbors set of x_i among the K coarse centers. By concatenation of the u_k of all the K centroids, the VLAD is formed with size KN, where N is the dimension of local descriptor after PCA pre-processing. Finally, we apply Intra-normalization [1], Signed Square Root (SSR) normalization, l_2 normalization to the VLAD features.

More specifically, for CaffeNet VLAD and NIN VLAD, we use $K = 32$ and $p = 12$, and PCA reduces the dimension of latent concept descriptors to 192 and 768 respectively, while for the low level local descriptor, we set $K = 200$ and $p = 5$, and PCA reduces the dimension of SIFT to 64.

[3] http://www.vlfeat.org.

The summarization of the procedure about extracting the image features is described in Algorithm 1:

Algorithm 1 The procedure of extracting image feature

Input:
Image im
Caffenet model CAF trained by triplet loss (5) and 10-category softmax,
Network in Network model NIN trained by triplet loss equation (5),
Dic_{CAF}, Dic_{NIN} and Dic_{Sift}, the VLAD dictionaries of CAF, NIN and $Sift$, trained by K-means

Output:
Catefory label C_{CAF},
Caffenet VLAD f_{CAF},
NIN VLAD f_{NIN},
SIFT VLAD f_{SIFT}

Crop im into im_1 by fix size crop;
Crop im into im_2 by MBD crop;
Forward propagate im_1 by CAF, extract local descriptors l_{CAF} and 10-category probability C_{CAF};
Encode l_{CAF} into Caffenet VLAD f_{CAF} by Dic_{CAF}, according to (6);
Forward propagate im_2 by NIN, extract local descriptors l_{NIN};
Encode l_{NIN} into NIN VLAD f_{NIN} by Dic_{NIN}, according to (6);
Forward propagate im_2 by $Sift$, extract local descriptors l_{SIFT};
Encode l_{SIFT} into SIFT VLAD f_{SIFT} by Dic_{Sift}, according to (6);
return C_{CAF}, f_{CAF}, f_{NIN}, f_{SIFT};

3.4 Result reranking and model fusion

Since the query images always belong to the same general class as the results, applying class level classification as a filtering step can significantly improve the retrieval results. We train a classification model based on CaffeNet. This model's parameters are shared with the feature extraction model in Section 3.3.1 to reduce computation time. We achieve 92.5% accuracy on 10 class general classification. However, it may not be accurate enough for using as hard category filter since it will incorrectly remove 8% of valid results from the final results. We treat softmax outputs as another kind of feature and compute the final distance between query image I_q and evaluation image I_e as follow:

$$\begin{aligned} d_{final} = {} & \alpha d_{CaffeNet}(I_q, I_e) + \beta d_{NIN}(I_q, I_e) \\ & + \delta d_{SIFT}(I_q, I_e) + \gamma C_{cat}, \end{aligned} \tag{7}$$

where

$$C_{cat} = \begin{cases} 1, & \text{if } \|x_q - x_e\| > C_{thresh} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

where x_q and x_e are softmax outputs.

4 Experimental evaluation

4.1 Experimental dataset

In our experiments, we utilize the ALISC dataset, which contains 2 million training images with tags (10 product categories, 676 subcategories and 6 product attributes), 3 million evaluation images, 3567 evaluation query images and 1417 validation query images with ground truth result sets. All the images are collected from TAOBAO E-commerce platform. The ALISC limits the running time of feature extraction to 1 second per image, based on the a single thread CPU Intel Xeon E5-2420 1.90GHz. Therefore, the efficiency of algorithm is also an important factor to be considered.

4.2 Evaluation metric

There are various metrics to evaluate retrieval results such as precision,recall, F measure, MAP and NDCG. The retrieval engine can present at most 20 images without having the user to scroll.

We use mean average precision (MAP) at 20 for our evaluation. MAP has been shown to be a good discrimination and stability evaluation metric for retrieval. For a given query, average precision is the average of precision values for the set of top k items. MAP@k is the mean of average precision over all queries. Given that Q is the set of all queries and R_{jk} is the set of top retrieved images up to k, MAP@k is calculated as follow.

$$MAP@k(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{k}\sum_{j=1}^{k} Precision(R_{jk}) \tag{9}$$

4.3 Experimental results

4.3.1 Product image search

In this section, we will present the experiments on the proposed solution (NIN_{tpl+v}+ CaffeNet_{tpl+v}+ SIFT_v). We will also compare the experiment results of searching with triplet NIN VLAD (NIN_{tpl+v}), triplet CaffeNet VLAD (CaffeNet_{tpl+v}), triplet NIN VLAD fusion with triplet CaffeNet VLAD (NIN_{tpl+v}+ CaffeNet_{tpl+v}), NIN VLAD (NIN_v), CaffeNet VLAD (CaffeNet_v), NIN VLAD fusion with CaffeNet VLAD (NIN_v+ CaffeNet_v) and SIFT VLAD (SIFT_v). The results of using the average pooling layer of NIN based models ($\text{NIN}_{tpl+avg}$ and NIN_{avg}) and fully connection layer of CaffeNet based models (CaffeNet_{tpl+fc} and CaffeNet_{fc}) are provided as baseline. To generate the triplet training data, first, we generate a small dataset by randomly selecting 1000 image queries and their related from the ALISC development set. Then for each training triplet, we randomly selected an image and one of it's related image as anchor and positive sample, and selected an unrelated image as negative sample.

The saliency based product localization is utilized to detect the product for NIN_v, SIFT_v and NIN_{avg}. For CaffeNet_v and CaffeNet_{fc} fixed size center crop is used to generate the input image. We also find that the configure of taking fixed size center crop for CaffeNet and saliency based method for NIN leads to a better performance than using saliency based method for both of CaffeNet and NIN. It means that the fixed size center crop is a good complementary for saliency based product localization, since more global information is preserved.

For NIN based VLAD, CaffeNet based VLAD and SIFT VLAD, the dimension of local descriptors are reduced to 768, 192 and 64 by PCA pre-process respectively. The number of clustering centers of NIN based VLAD, CaffeNet based VLAD and SIFT VLAD are 32, 32 and 200.

The experimental results are shown in the Table 1. In the experimental result, we can see that the proposed method significantly outperforms the other compared methods, which reaches an MAP@20 of 37.64. It suggested that the proposed methods are effective in solving the product search problems. We further show some results of product image search in the Fig. 7.

For triplet metric learning. The results of $\text{NIN}_{tpl+avg}$ and CaffeNet_{tpl+fc}, which employ the triplet metric learning to fine-tune the deep learning model, are compared with the results of NIN_{avg} and CaffeNet_{fc}. In the comparisons we find that by using the triplet metric learning, the performance of NIN and CaffeNet are boosted about 9.07 and 9.21% respectively. It may because the deep learning models we used here are trained to handle the ImageNet classification task, and they can discriminate the difference between categories. However, the product search tasks require the models are able to discriminate not only the category difference but also the details between the instances in the same category. Nevertheless, as there are some connections between general image classification task and our product search task, in both cases, the CNN are learned to understand some latent patterns lies in the image dataset, such as the $lines$, $corners$, $gredients$ in images. By adopting the triple metric learning, the pertrained CNN are finetuned to understand how these patterns are arranged in the product search dataset. After finetuning by triple metric learning, the features belong to same instance are trend to similar and the features belong to vary instances are trend to differential, and in this way the discriminative of deep learning model are improved.

For spatial quantization. We compared the performance of NIN_{avg} and CaffeNet_{fc} with NIN_{v} and CaffeNet_{v}, which employ the spatial quantization as feature enhancement method. In the comparison we can see the two results generated by CNN with spatial quantization are much better those without, it suggest that the spatial quantization is effectively in improving the deep feature discriminative. There are two reasons for such result: First, the

Table 1 Performance comparison (MAP@20 in percentage)

Methods	MAP@20 (%)
SIFT_{v}	11.09
CaffeNet_{fc}	13.37
NIN_{avg}	15.06
CaffeNet_{tpl+fc}	22.58
CaffeNet_{v}	22.89
$\text{NIN}_{tpl+avg}$	24.13
CaffeNet_{tpl+v}	24.34
NIN_{v}	27.43
NIN_{tpl+v}	29.90
NIN_{v}+CaffeNet_{v}	32.03
NIN_{tpl+v}+CaffeNet_{tpl+v}	33.74
NIN_{v}+CaffeNet_{v}+SIFT_{v}	35.35
NIN_{tpl+v}+CaffeNet_{tpl+v}+SIFT_{v}	37.64

fully connection layer as well as global average pooling in CNN is a kind of summarization of the local feature, which ignore lots of detail information in image. On this other hand, the spatial quantization, which utilizes the feature maps of convolution layer as local descriptors, maintains more detail information and spatial relationships in the final feature. Second, the spatial quantization employed VLAD as coding method, which records the residual between local descriptors and visual vocabulary, in this way, the information loss in quantization are reduced. It's worth noting here that, compareing the results of NIN_{tpl+v} and CaffeNet_{tpl+v} with $\text{NIN}_{tpl+avg}$ and CaffeNet_{tpl+fc}, we can see that even been employed in the triplet finetuned CNN model, the spatial quantization can still improve the discriminative of CNN feature.

For the comparison deep learning feature and traditional feature. In the experiments, we found that the deep learning based methods are better than the traditional feature based methods, even the pretrained model with no finetuning are outperform the SIFT_v by 3%. However, we also found that, even the result of traditional features much is worse than deep learning features, combine these two methods together can still obtain a better performance. For Error analysis. We tested our methods in the validation dataset. In the Fig. 6, we show the performance of our methods on the 10 categories of ALISC. In the experimental results, we find that our method has a better performance in category of top-wear, snacks and cosmetics, and perform worse in jewelry, dress and shoes. According to our observation, compared the categories with better results, the images in jewelry, dress and shoes may have such problems: the simple texture or patterns, the unfixed viewpoints and the unsalient objects. A promising solution for these problems is to utilize some technical like Hough transformation to detect and rectify the viewpoints of objects, and further take some super-pixels methods to enhance the details and patterns in images.

4.3.2 Efficiency analysis

We also evaluate the running time of the proposed method. The total running time of feature extraction is 0.78 second, which is less than the 1 second limit of ALISC, the running time of each step is shown in Table 2. It suggests that the proposed product image search framework is very efficient. Note that it is the running time for single thread on CPU only machine.

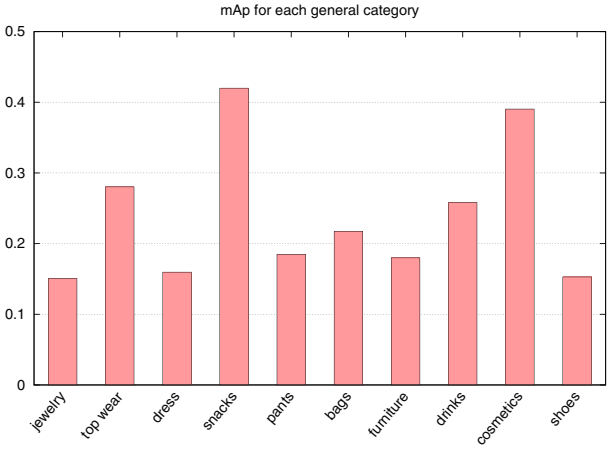


Fig. 6 The performance of proposed paradigm on 10 categories

Fig. 7 Some sample results of product image search. The first column is the query, the other columns are results; bounding boxes of product localization are also provided

The extraction of CNN descriptors, which is the bottle neck of the whole framework, can be accelerated significantly by utilizing GPU.

4.3.3 Product localization

To evaluate the influence of the proposed saliency based product localization, we also test performance of removing product localization and using center crop image as input for NIN VLAD (NIN$_v$ no PL) and SIFT VLAD (SIFT$_v$ no PL). Although the CaffeNet use center crop in this work, we also provide the performance of CaffeNet VLAD with product localization here for comparison. From Table 3. it can be seen that, by using saliency based product localization, the NIN VLAD, CaffeNet VLAD and SIFT VLAD obtain great imporvment of 4.30, 2.85 and 2.26% on MAP@20.

The product localization step takes only 40ms to detect product in the image, which is much lower than the time consumed for the feature extraction step. Moreover, we randomly selected 500 images from training images and labeled the position of product in images. We then use this small image dataset to test the accuracy of the product localization method.

Table 2 Time consuming of each step

Steps	Runing time (s)
Product localization	0.04
NIN VLAD	0.31
Caffenet VLAD	0.29
SIFT VLAD	0.14
Total	0.78

Table 3 The influence of product localization

Methods	MAP@20 (%)
NIN_v with PL	27.43
NIN_v no PL	23.13
$CaffeNet_v$ with PL	25.74
$CaffeNet_v$ no PL	22.89
$SIFT_v$ with PL	11.09
$SIFT_v$ no PL	8.83

The Intersection over Union (IOU) rate of product localization reaches 83% on average. We show some example results of product localization in Fig. 7.

5 Conclusion

We have presented an approach for improving the discriminative capacity of deep features using SPP pooling and VLAD encoding. We have also demonstrated an efficient and effective saliency based product localization approach.

We further note that while deep learning features outperform traditional features such as SIFT, SIFT can still improve the overall performance when used in conjunction with deep features.

We used fusion of multiple models and our results on a challenging real world dataset are significantly better than single models or methods using just deep features without spatial pooled quantization.

Multiple deep learning models were utilized to enhance the robustness and discriminability of image representation. Experiments on ALISC dataset demonstrated the effectiveness of our approach.

While we have demonstrated that significant performance gains can be obtained beyond baseline CNN features via VLAD quantization, further performance gains are obtained when used in conjunction with metric learning models. In the future work, we are expected to exploit a more discriminative CNN model by using some lastest deep learning technical such as generative adversarial networks (GANs) et al.

Acknowledgments This work is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@SC Funding Inititative, by International Exchange and Cooperation Foundation of Shenzhen City under Grant no.GJHZ2015031214149569, by Science and Technology Planning Project of Guangdong Province under Grant no.2016A040403046, the ALISC dataset is provided by Alibaba Group.

References

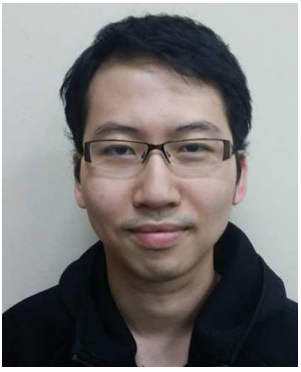
1. Arandjelovic R, Zisserman A (2013) All about Vlad. In: Proceedings of the IEEE international conference on computer vision, pp 1578–1585
2. Bai Y, Yang K, Yu W, Ma WY, Zhao T (2013) Learning high-level image representation for image retrieval via multi-task dnn using clickthrough data. arXiv:1312.4740
3. Bromley J, Bentz JW, Bottou L, Guyon I, LeCun Y, Moore C, Säckinger E, Shah R (1993) Signature verification using a Siamese time delay neural network. Int J Pattern Recognit Artif Intell 7(04):669–688

4. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the IEEE international conference on computer vision, vol 1, pp 539–546
5. Cui J, Liu Y, Xu Y, Zhao H, Zha H (2013) Tracking generic human motion via fusion of low-and high-dimensional approaches. IEEE Trans Syst Man Cybern Syst Hum 43(4):996–1002
6. Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS (2015) Deep learning for visual understanding: a review. Neurocomputing 187(C):27–48
7. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE international conference on computer vision, vol 2, pp 1735–1742
8. Hoffer E, Ailon N (2015) Deep metric learning using triplet network. In: Similarity-based pattern recognition. Springer, pp 84–92
9. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093
10. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of advances in neural information processing systems, pp 1097–1105
11. Lin K, Yang HF, Hsiao JH, Chen CS (2015) Deep learning of binary hash codes for fast image retrieval. In: Proceedings of the IEEE international conference on computer vision workshops, pp 27–35
12. Lin M, Chen Q, Yan S (2013) Network in network. arXiv:1312.4400
13. Liu L, Cheng L, Liu Y, Jia Y, Rosenblum DS (2016) Recognizing complex activities by a probabilistic interval-based model. In: AAAI, pp 1266–1272
14. Liu Y, Cui J, Zhao H, Zha H (2012) Fusion of low-and high-dimensional approaches by trackers sampling for generic human motion tracking. In: International conference on pattern recognition. IEEE, pp 898–901
15. Liu Y, Liang Y, Liu S, Rosenblum DS, Zheng Y (2016) Predicting urban water quality with ubiquitous data. arXiv:1610.09462
16. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2016) Action2activity: recognizing complex activities from sensor data. arXiv:1611.01872
17. Liu Y, Nie L, Liu L, Rosenblum DS (2016) From action to activity: sensor-based activity recognition. Neurocomputing 181:108–115
18. Liu Y, Zhang L, Nie L, Yan Y, Rosenblum DS (2016) Fortune teller: predicting your career path. In: AAAI, pp 201–207
19. Liu Y, Zhang X, Cui J, Wu C, Aghajan H, Zha H (2010) Visual analysis of child-adult interactive behaviors in video sequences. In: Proceedings of the international conference on virtual systems and multimedia. IEEE, pp 26–33
20. Liu Y, Zheng Y, Liang Y, Liu S, Rosenblum DS (2016) Urban water quality prediction based on multi-task multi-view learning. In: Proceedings of the international joint conference on artificial intelligence
21. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
22. Lu Y, Wei Y, Liu L, Zhong J, Sun L, Liu Y (2016) Towards unsupervised physical activity recognition using smartphone accelerometers. Multimedia Tools and Applications, pp 1–19
23. Otsu N (1975) A threshold selection method from gray-level histograms. Automatica 11(285-296):23–27
24. Qi S, Wang F, Wang X, Wei J, Zhao H (2015) Live multimedia brand-related data identification in microblog. Neurocomputing 158(C):225–233
25. Qi S, Zawlin K, Zhang H, Wang X, Gao K, Yao L, Chua TS (2016) Saliency meets spatial quantization: a practical framework for large scale product search. In: Proceedings of the IEEE international conference on multimedia and expo workshops. IEEE, pp 1–7
26. Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE international conference on computer vision workshop, pp 806–813
27. Shah RR (2016) Multimodal analysis of user-generated content in support of social media applications. In: Proceedings of the international conference on multimedia retrieval (ICMR). ACM, pp 423–426
28. Shah RR (2016) Multimodal-based multimedia analysis, retrieval, and services in support of social media applications. In: Proceedings of the international conference on multimedia. ACM
29. Shah RR, Shaikh AD, Yu Y, Geng W, Zimmermann R, Wu G (2015) Eventbuilder: real-time multimedia event summarization by visualizing social media. In: Proceedings of the international conference on multimedia. ACM, pp 185–188
30. Shah RR, Yu Y, Verma A, Tang S, Shaikh AD, Zimmermann R (2016) Leveraging multimodal information for event summarization and concept-level sentiment analysis. Proceedings of the Knowledge-Based Systems 108:102–109

31. Shah RR, Yu Y, Zimmermann R (2014) Advisor: personalized video soundtrack recommendation by late fusion with heuristic rankings. In: Proceedings of the international conference on multimedia. ACM, pp 607–616
32. Shah RR, Yu Y, Zimmermann R (2014) User preference-aware music video generation based on modeling scene moods. In: Proceedings of the multimedia systems conference. ACM, pp 156–159
33. Shen X, Lin Z, Brandt J, Wu Y (2012) Mobile product image search by automatic query object extraction. In: European conference on computer vision, pp 114–127
34. Snchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: theory and practice. Int J Comput Vis 105(3):222–245
35. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE international conference on computer vision, pp 1–9
36. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE international conference on computer vision, pp 1701–1708
37. Van De Sande KE, Gevers T, Snoek CG (2010) Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 32(9):1582–1596
38. Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the international conference on multimedia. ACM, pp 157–166
39. Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, Chen B, Wu Y (2014) Learning fine-grained image similarity with deep ranking. Computer Science, pp 1386–1393
40. Wang W, Yan Y, Zhang L, Hong R, Sebe N (2016) Collaborative sparse coding for multiview action recognition. IEEE Multimedia 23(4):80–87
41. Xu Z, Yang Y, Hauptmann AG (2015) A discriminative cnn video representation for event detection. In: Proceedings of the IEEE international conference on computer vision, pp 1798–1807
42. Zagoruyko S, Komodakis N (2015) Learning to compare image patches via convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision
43. Zhang H, Shang X, Luan H, Wang M, Chua TS (2016) Learning from collective intelligence: feature learning using social images and tags. ACM Trans Multimed Comput Commun Appl 13(1):1
44. Zhang H, Shen F, Liu W, He X, Luan H, Chua TS (2016) Discrete collaborative filtering. In: Proceedings of the international ACM SIGIR conference on research and development in information retrieval, pp 325–334
45. Zhang H, Yang Y, Luan H, Yang S, Chua TS (2014) Start from scratch: towards automatically identifying, modeling, and naming visual attributes. In: Proceedings of the international conference on multimedia. ACM, pp 187–196
46. Zhang H, Zha ZJ, Yang Y, Yan S, Gao Y, Chua TS (2014) Attribute-augmented semantic hierarchy: towards a unified framework for content-based image retrieval. ACM Trans Multimed Comput Commun Appl 11(1s):21
47. Zhang J, Sclaroff S, Lin Z, Shen X, Price B, Mech R (2015) Minimum barrier salient object detection at 80 fps. In: Proceedings of the IEEE international conference on computer vision, pp 1404–1412
48. Zhang L, Gao Y, Hong C, Feng Y, Zhu J, Cai D (2014) Feature correlation hypergraph: exploiting high-order potentials for multimodal recognition. IEEE Transactions on Cybernetics 44(8):1408–1419
49. Zhang L, Gao Y, Ji R, Xia Y, Dai Q, Li X (2014) Actively learning human gaze shifting paths for semantics-aware photo cropping. IEEE Trans Image Process 23(5):2235–45
50. Zhang L, Gao Y, Xia Y, Lu K (2014) Representative discovery of structure cues for weakly-supervised image segmentation. IEEE Trans Multimedia 16(2):470–479
51. Zhang L, Gao Y, Zimmermann R, Tian Q, Li X (2014) Fusion of multichannel local and global structural cues for photo aesthetics evaluation. IEEE Trans Image Process 23(3):1419–29
52. Zhang L, Han Y, Yang Y, Song M, Yan S, Tian Q (2013) Discovering discriminative graphlets for aerial image categories recognition. IEEE Trans Image Process 22(12):5071–5084
53. Zhang L, Hong R, Gao Y, Ji R (2015) Image categorization by learning a propagated graphlet path. IEEE Transactions on Neural Networks and Learning Systems 27(3):674–685
54. Zhang L, Li X, Nie L, Yan Y, Zimmermann R (2016) Semantic photo retargeting under noisy image labels. ACM Trans Multimed Comput Commun Appl 12(3):37
55. Zhang L, Li X, Nie L, Yang Y, Xia Y (2016) Weakly supervised human fixations prediction. IEEE Transactions on Cybernetics 46(1):258–269
56. Zhang L, Song M, Li N, Bu J, Chen C (2009) Feature selection for fast speech emotion recognition. In: Proceedings of the international conference on multimedia, pp 753–756

57. Zhang L, Song M, Liu Z, Liu X, Bu J, Chen C (2013) Probabilistic graphlet cut: exploiting spatial structure cue for weakly supervised image segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1908–1915
58. Zhang L, Wang M, Hong R, Yin B (2015) Large-scale aerial image categorization using a multitask topological codebook. IEEE Transactions on Cybernetics 46(2):1
59. Zhang L, Yang Y, Gao Y, Yu Y, Wang C, Li X (2014) A probabilistic associative model for segmenting weakly-supervised images. IEEE Trans Image Process 23(9):4150–4159
60. Zhang L, Yang Y, Wang M, Hong R, Nie L, Li X (2015) Detecting densely distributed graph patterns for fine-grained image categorization. IEEE Trans Image Process 25(2):1–1
61. Zhang R, Lin L, Zhang R, Zuo W, Zhang L (2015) Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. IEEE Trans Image Process 24(12):4766–4779

Shuhan Qi received his M.S. degree in Computer Sciences from the Harbin Institute of Technology in 2011. Since 2011, he has been a Ph.D. degree candidate in Computer Sciences from Harbin Institute of Technology Shenzhen Graduate School. His research interests include computer vision, multimedia and pattern recognition.



Zawlin Kyaw received his B.Sc in school of computing from National University of Singapore. He has been working as a senior engineer at Kaisquare Company Since 2012, and now he is also working as a research assistant at lab of media search, National University of Singapore. His research interests include computer vision, multimedia and pattern recognition.

Xuan Wang received his M.S. and Ph.D. degrees in Computer Sciences from Harbin Institute of Technology in 1994 and 1997 respectively. He is a professor and Ph.D. supervisor in the Computer Application Research Center, Harbin Institute of Technology Shenzhen Graduate School. His main research interests include artificial intelligence, computer vision, computer network security and computational linguistics.

Zoe L. Jiang received the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2010. She is currently an Assistant Researcher with School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China. Her research interests include computer vision and pattern recognition.

Jian Guan received the B.Sc. and M.Sc. degrees from the College of Computer Science and Technology, Jilin University, China, in 2005 and 2010, respectively. He is now pursuing the Ph.D. degree in Computer Science at Harbin Institute of Technology Shenzhen Graduate School. His research interests include blind signal processing, machine learning, and sparse signal processing.