Product Image Search with Regional Evidences

Guixuan Zhang, Shuwu Zhang, Zhi Zeng, Hu Guan, Xiaoqian Li
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Email:guixuan.zhang@ia.ac.cn, shuwu.zhang@ia.ac.cn, zhi.zeng@ia.ac.cn,
hu.guan@ia.ac.cn, lixiaoqian2015@ia.ac.cn

***Abstract*—In the task of product image search, the database consists of clean versions of product images, while the query photos are often captured from mobile phone cameras under uncontrolled conditions. Conventional methods usually adopt the SIFT based bag-of-words (BoW) representation of the whole query image, which suffers from the interference of background noise. To address the problem, we extract multiple candidate regions from the query image and compute the regional similarity to database images individually. Then a verification strategy is proposed to evaluate the similarity based on regional semantic evidences. With the proposed method, we can not only improve the search accuracy, but also obtain the location of the product in the query image. Extensive experiments on two public datasets demonstrate the effectiveness of our method.**

Keywords—product image search; bag-of-words; regional semantic evidences; similarity verification

I. INTRODUCTION

This paper considers the task of product image retrieval. With the development of e-commerce and popularity of mobile phones, image-based product search has attracted a lot of attention. In an ideal scenario, the user snaps a photo of a product to identify this product and search the information about it. To describe the products better, the database product images are often photographed professionally, with clean backgrounds. Meanwhile, the query photos are usually taken under uncontrolled conditions with various backgrounds. Some examples are shown in Fig. 1.

Most image retrieval systems rely on local descriptors, such as SIFT [1]. These descriptors are typically used jointly with bag-of-words (BoW) model [2]. In BoW, a visual vocabulary is trained on an independent set of SIFTs with k-means clustering. The SIFTs extracted from an image are quantized to the nearest visual words, then the image can be represented by a vector of the visual word histogram. The similarity between two images can be measured by the cosine similarity of their BoW representations. To make the search efficient, the inverted index is usually employed.

The BoW representation of the whole image is often used for generic image retrieval [3,4]. This kind of global BoW representation is also popular for product image search [5]. Recently, encoding techniques such as Fisher vector [6] and VLAD [7] are applied for product search tasks [8,9]. Though the search accuracy and efficiency have been improved to some extent, these methods still focus on the representation of the entire image. Since the query image may be distracted by the background noise, irrelevant product images are often retrieved when whole image representation is used. An example is illustrated in Fig. 2(a).

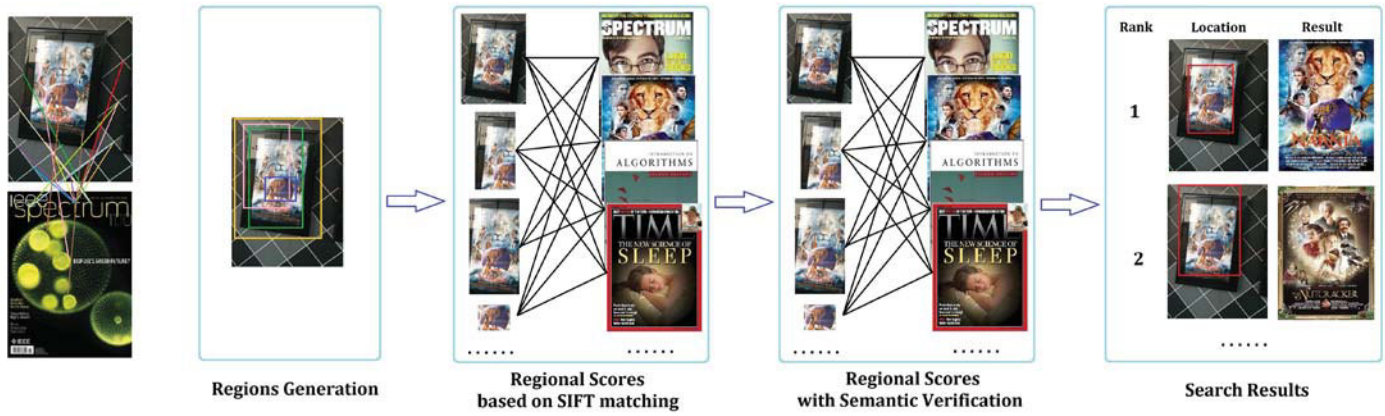
Fig. 1. Examples of product image search. The top two rows show the examples from the Mobile dataset [10], and the last row shows the examples from the Stanford Mobile Visual Search Dataset [11]. All the database images are high-quality without any background. The query images are taken under different lighting conditions with various backgrounds.

To address this problem, we extract multiple candidate regions from the query image. All these regions are considered individually and each region has its own similarity to a database image. These regions are expected to enclose the target product and exclude the noise from the rest. In this paper, we adopt EdgeBoxes [12], which is a proposal method for object detection tasks, to generate the candidate regions.

Decomposing the image into multiple regions has been applied for image retrieval by some works [13,14]. Our work departs from these methods in two aspects. First, we extract candidate regions from the query images, while [13,14] consider multiple regions on the database side. Second, we propose a verification scheme to further check whether one region is similar to a database product image, which helps to improve the accuracy.

As the first contribution in this paper, we propose to extract multiple regions from the query image and conduct the search process for each region independently. By treating the BoW model as matching-based approach, we can compute the similarity scores for all regions efficiently. The second contribution is that a verification method is proposed to evaluate the regional similarity to database images. We adopt semantic evidences from the convolutional neural networks (CNN) [15] to further check whether one query region is relevant to a database image. With these two steps, the product image search accuracy is improved. Moreover, we can locate

978-1-5090-3484-0/16/$31.00 ©2016 IEEE



(a) (b)

Fig.2. (a) An example which fails to retrieve the right database image. The whole BoW representation is used and the background noise leads to the failure. (b) The framework of our method. We first generate multiple candidate regions. Then each region is considered as a query and its similarities to database images are computed. This step is implemented efficiently by adopting the SIFT based matching approach. After that, regional semantic evidences based on CNN are used to further verify the regional similarities. Finally, we return the results by sorting the similarity scores. Our method can also locate the product in the query image.

the product in the query image. Extensive experiments on two public datasets will show the effectiveness of our method.

The remainder of this paper is organized as follows. The proposed approach is described in Section II. The experimental results are shown in Section III. Final conclusions are in Section IV.

II. Proposed Approach

The framework of our proposed approach is illustrated in Fig. 2(b). We first generate multiple regions which are expected to contain the product for the query image. Then each region is considered as an independent query and we compute its similarity scores to database images based on SIFT matching. After that, we adopt semantic evidences to further verify the regional similarities. Finally, we return the results by sorting the similarity scores. For each retrieved database product, we can provide the corresponding location of this product in the query image.

A. Multiple Regions Generation

We expect to concentrate on the content of the product in the query image, so that the interference of background noise can be weakened or eliminated. To this end, we propose to search locally in the query image by evaluating multiple regions. These regions are expected to enclose the product which we want to search. Here we adopt EdgeBoxes [12] to generate candidate regions. EdgeBoxes is designed to detect potential objects in an efficient manner. It is able to achieve a high recall with a reasonable number of regions. For each query image, we extract a set of regions $\mathcal{R}=\{r_1,r_2,\ldots,r_T\}$, where T is the number of regions.

B. Region Search based on SIFT Matching

Let us assume that an image is described by a set $\mathcal{X}=\{x_1,x_2,...,x_n\}$ of n 128-dimensional SIFTs. A visual vocabulary $\mathcal{C}=\{c_1,c_2,...,c_k\}$ has been trained with k-means algorithm, where k indicates the vocabulary size. Each SIFT is quantized to a nearest visual word c by quantizer $q(\cdot)$. Then an image can be represented by a k-dimensional visual word histogram, which is referred to as BoW representation. The similarity between two images is measured by the cosine similarity of their BoW representations.

In our method, each region is considered independently and its similarities to all database images should be computed. If we compute the BoW representations for all regions and then measure the cosine similarity exhaustively, the computation complexity would be very high. To address this problem, a SIFT based matching function is introduced.

Let r be a region in the query image with SIFT set $\mathcal{X}=\{x_1,x_2,...,x_{n_r}\}$ and let I be a database image with $\mathcal{Y}=\{y_1,y_2,...,y_{n_I}\}$. Then the similarity between r and I is

$$sim(r,I)=\frac{1}{\|\mathcal{X}\|\|\mathcal{Y}\|}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}M(x,y)\,, \tag{1}$$

where $M(x,y)$ is a matching function and $1/\|\mathcal{X}\|\|\mathcal{Y}\|$ is the normalization term. In the case of bag-of-words, the matching function is

$$M_{BoW}(x,y)=\begin{cases}1 & \text{if } q(x)=q(y)\\ 0 & \text{otherwise}\end{cases}. \tag{2}$$

With the matching function of M_{BoW}, (1) is equal to the cosine similarity of BoW representation. We compute the score of each SIFT x in the query image with the function of $sim(x,I)=\sum_{y\in\mathcal{Y}}M(x,y)$. Then the similarity score of a certain region is obtained by summing the scores over the SIFTs which locate inside the region. By leveraging the inverted index, we can compute the similarity scores for all query regions efficiently.

Since the conventional bag-of-words model has limited performance, we adopt Hamming embedding (HE) [16] to define a new matching function. HE extends BoW by representing each SIFT x with a 64-bit binary code b_x when quantizing it. Then the match kernel is updated as

$$M_{HE}(x,y)=\begin{cases} e^{-h^2/\sigma^2} & \text{if } q(x)=q(y) \text{ and } h(b_x,b_y)\le\kappa \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

where h is the Hamming distance, κ is the Hamming distance threshold and σ is the parameter of the exponential score function. In the experiment section, we will evaluate both M_{BoW} and M_{HE}.

C. Similarity Verification with Semantic Evidences

For each region, we have obtained its similarities to all database images based on SIFT matching. However, SIFT is not sufficiently distinctive to prevent false matches, since it only describes the gradient distribution in a local patch, without seeing the big picture. So the SIFT based similarity between the query region and a database image may be not convincing. To address this issue, we propose to verify the similarity by employing regional holistic clues.

Recently, convolutional neural networks achieve state-of-the-art performance in many computer vision tasks [15,17,18]. A number of works show that CNN models pre-trained on large and diverse datasets such as ImageNet [19] can be transferred to other image applications. The outputs of fully connected (FC) layers in CNN can be used as semantic-aware features, which describe the image content as a whole. In this paper, we adopt the AlexNet architecture [15] implemented by Caffe [20]. In AlexNet, there are eight layers (5 convolutional layers and 3 fully connected layers), thus we name the three FC layers as FC6, FC7 and FC8. The outputs of FC6 and FC7 are 4096-dimensional, and the dimension of FC8 feature is 1000. We will evaluate the three different FC features in the experiment section.

Considering the efficiency, we transform the high-dimensional FC feature to binary signature with locality-sensitive hashing (LSH) algorithm [21], then every region is associated with a 128-bit signature, which represents the semantic evidence.

Let b_r^c be the semantic signature of region r in a query image, and let b_I^c be the semantic signature of a database image I. Our verification scheme updates (1) as follows

$$sim'(r,I)=sim(r,I)\times e^{-d^2/\theta^2}, \quad (4)$$

where d is the Hamming distance between their semantic signatures and θ is the parameter of the exponential score function. The exponential function rewards a pair (r,I) with high weight if they are close in the semantic space, and punishes the pair if they have large distance in the semantic space.

D. Search Results

Since we generate T regions for a given query image and conduct the search for each region independently, we obtain T similarity scores for every database image. The maximum of the T scores is considered as the final similarity between the query and a database image. Each database image represents an unique product, so the corresponding region in the query image provides the location of this product. We rank the database images by sorting their scores and then return the final results.

III. EXPERIMENTS

A. Datasets

We evaluate the proposed method on two public datasets, the Mobile dataset [10] and Stanford Mobile Visual Search Dataset (SMVS) [11]. The Mobile dataset has 400 database images and 2500 queries, captured by mobile devices. This dataset contains images of books, magazines, movie posters and snacks. The SMVS dataset is designed for different mobile visual search applications. We choose the product categories to evaluate our method. 4 categories are selected, including books, CDs, DVDs, and museum paintings. There are 392 database images and 1567 queries. The query photos are captured from heterogeneous camera phones under widely varying conditions. Some examples are illustrated in Fig. 1. To increase the difficulty, both databases are blended by 10200 images from UKBench [3] as distractors.

Since the objective of product image search is to identify the product in the query image, we care mostly about the top-1 precision (P@1). $\mathrm{P@1}=1$ means that we successfully recognize the query product.

B. Implementation Details

Hessian Affine detector [22] and SIFT descriptor are used for local feature extraction. Following [23], the rootSIFT is used since it is shown to yield superior performance at no cost. A visual vocabulary of size 20k is trained using an independent dataset. Following [16,24], the Hamming distance threshold κ is set to 22, and the weighting parameter σ is set to 16. Moreover, we employ the multiple assignment scheme [16] on the query side, in which a SIFT is assigned to 5 nearest visual words.

C. Parameter Analysis

The impact of the parameter θ associated with (4) is shown in Fig. 3. The matching function of M_{BoW} is used. For both datasets, the top-1 precision rises to the peak at $\theta=40$. Therefore, we set θ to 40.

Fig. 3 also shows the performance of three different FC features. Both FC6 and FC7 work better than FC8. It seems that the 4096-dimensional FC6 and FC7 features provide more sufficient semantic information than the 1000-dimensional FC8 feature. FC6 and FC7 have similar performance. Since FC7 works a little better than FC6, we adopt FC7 feature in the rest of our experiments.

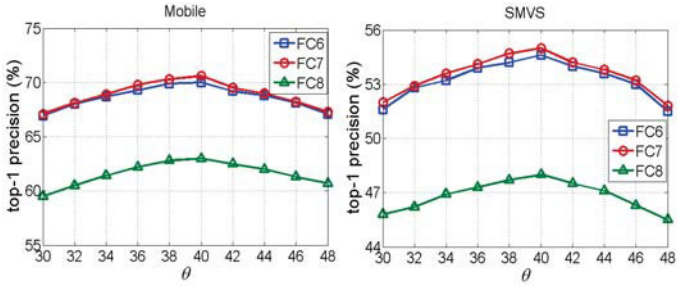


Fig. 3. The impact of the parameter θ. The matching function of M_{BoW} is used.

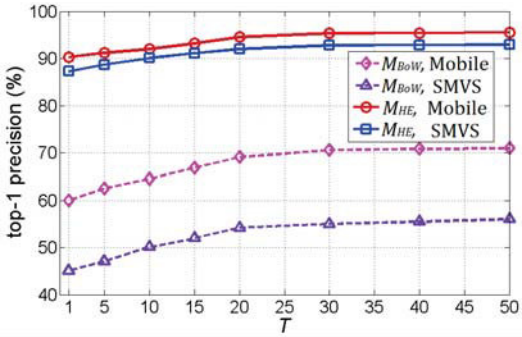


Fig. 4. The impact of region number T. The FC7 feature is used to represent the semantic evidence.

The impact of the region number T is shown in Fig. 4. With the increase of T, the top-1 precision rises for both M_{BoW} and M_{HE}. Note that when T is larger than 30, the precision increases very slowly. Considering the tradeoff between accuracy and efficiency, we choose to use 30 candidate regions for each query image.

D. Evaluation

The effectiveness of the proposed method is demonstrated in Table I. We evaluate the effectiveness of different components in our method. The region search and similarity verification correspond to the algorithms described in Section II-B and Section II-C, respectively. As shown in Table I, the process of region search brings consistent improvements over the original BoW and HE methods. The similarity verification with semantic evidences helps to further improve the performance.

E. Complexity

In our method, 4 bytes are allocated to store image ID in the inverted index for each SIFT. When HE is used, 8 bytes are needed to store the 64-bit binary SIFT feature. Each database image costs 16 bytes to store its 128-bit semantic signature. Assume that every image has 1000 SIFTs, our method consumes 1.16 GB memory for a 100k database. The average query time for a database of 100k product images is 0.09s.

F. Comparison

Table II compares our approach with some existing methods. The proposed method achieves the best performance for both Mobile and SMVS datasets. Moreover, our method is able to provide the location of the product in the query image. As shown in Table II, the efficiency of our method is very competitive with other methods.

TABLE I. THE EFFECTIVENESS OF DIFFERENT COMPONENTS IN OUR METHOD

Methods	Region Search	Similarity Verification	Mobile P@1 (%)	SMVS P@1 (%)
M_{BoW}			51.9	42.2
	√		63.5	48.6
	√	√	70.6	55.8
M_{HE}			88.2	86.5
	√		91.5	88.3
	√	√	95.3	92.8

TABLE II. COMPARISION WITH EXISTING METHODS

Methods	Mobile P@1 (%)	SMVS P@1 (%)	Search Time (s)
BoW[2]	51.9	42.2	0.12
HE[16]	88.2	86.5	0.08
REVV[9]	79.5	76.8	0.05
FV+SAA[8]	83.5	81.6	0.05
c-MI [25]	93.9	87.5	0.06
Ours	**95.3**	**92.8**	0.09

IV. CONCLUSION

In this paper, we propose a product image search method with regional evidences. We generate multiple candidate regions for the query image and conduct the search process for all regions individually. This step helps to weaken the noise interference from the background. Moreover, we employ regional semantic evidences extracted by CNN model to verify the similarity, which further improves the performance. Extensive experiments on two public datasets demonstrate the effectiveness of our method.

ACKNOWLEDGMENT

This work has been supported by the National Science and Technology Supporting Program of China under Grant No.2015BAH49F01 and the Key Technology R&D Program of Beijing under Grant No.D161100005216001.

REFERENCES

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, pp. 1470–1477, 2003.

[3] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, pp. 2161–2168, 2006.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, pp. 1-8, 2007.

[5] B. Girod, V. Chandrasekhar, and D. Chen, "Mobile visual search," *IEEE Signal Process. Mag.*, 2011.

[6] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, June 2007.

[7] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE*

Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1704-1716, Sept. 2012.

[8] G. Zhang, Z. Zeng, S. Zhang, and Q. Guo, "Transmitting informative components of fisher codes for mobile visual search," in *ICASSP*, pp. 1136-1140, 2015.

[9] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grezeszczuk, and B. Girod, "Residual enhanced visual vector as a compact signature for mobile visual search" *Signal Processing*, 2012.

[10] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T.X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *ICCV*, pp. 209-216, 2011.

[11] V. Chandrasekhar, D. Chen, S. Tsai, N.M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod. "The stanford mobile visual search data set," in *ACM Conference on Multimedia Systems*, pp. 117-122, 2011.

[12] C. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *ECCV*, pp. 391-405, 2014.

[13] R. Arandjelovic and A. Zisserman, "All about VLAD", in *CVPR*, pp. 1578-1585, 2013.

[14] R. Tao, E. Gavves, C. Snoek, and A. Smeulders, "Locality in generic instance search from one example," in *CVPR*, pp. 2099-2106, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks." in *NIPS,* pp. 1106-1114, 2012.

[16] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vis*., vol. 87, pp. 316–336, 2010.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS,* pp. 91-99, 2015.

[19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F.F. Li, "ImageNet: A large scale hierarchical image database," in *CVPR*, pp. 248-255, 2009.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," CoRR abs/1408.5093, 2014.

[21] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *ACM Symposium on Theory of Computing*, pp.380-388, 2002.

[22] M. Perdoch, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval." in *CVPR*, pp. 9-16, 2009.

[23] R. Arandjelović, and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, pp. 2911-2918, 2012

[24] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *CVPR*, pp. 1169-1176, 2009.

[25] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *CVPR*, pp. 1947-1954, 2014.