OBJECT RETRIEVAL BASED ON SPATIALLY FREQUENT ITEMS WITH INFORMATIVE PATCHES

Ke Gao [1, 2], Shouxun Lin [1], Junbo Guo [1], Dongming Zhang [1], Yongdong Zhang [1], Yufeng Wu [1]

[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 100080
[2]Graduate University of the Chinese Academy of Sciences, Beijing, China, 100080
{kegao, sxlin, zhyd, ts, wuyufeng}@ict.ac.cn

ABSTRACT

Spatial relation of local image patches plays an important role in object-based image retrieval. An approach called spatial frequent items is proposed as an extension of Bag-of-Words method by introducing spatial relations between patches. Spatial frequent items are defined as frequent pairs of adjacent local image patches in polar coordinates, and exploited using data mining. Based on these frequent configurations, we develop a method to encode patches and their spatial relations for image indexing and retrieval. Besides, to avoid the interference of background patches, informative patches are filtrated based on their local entropy and self-similarity in the preprocess stage. Experimental results demonstrate that our method can be 8.6% more effective than the state-of-art object retrieval methods.

Index Terms—Object retrieval, Spatial frequent items, Informative patches

1. INTRODUCTION

OBIR (Object-based Image Retrieval) is an important branch of content-based image retrieval [1, 2, 3]. The goal of OBIR is to find images containing desired object by providing the system a selected region of a query image. It remains a challenging problem because an object's visual appearance may be quite different due to viewpoint, illumination, affine transformation, and even partially occlusion.

The innermost core of OBIR is how to detect and measure the similarities of object regions. Recent work in this field can be divided into two categories: one is based on image segmentation, such as Blobworld and SIMPLcity [3]; the other is so-called BoW (Bag of Words) method, which simulates simple text-retrieval system using the analogy of "visual words" [1, 2]. BoW doesn't rely on the precision of image segmentation, and can deal with a variety of affine transformations. In consequence, it has become increasingly attractive [4, 5].

In BoW, "saliency" patches are detected in images, and a high-dimensional descriptor is computed for each patch. To effectively index these descriptors, they are clustered into a visual vocabulary, and each patch is mapped to its closest visual word. Then an image is represented as a bag of visual words and their frequency of occurrence. Usually, they are organized as an inverted file to facilitate efficient retrieval.

Although BoW is very effective, the spatial information about the image-location of the visual words is ignored, which is similar to retrieve documents only by orderless letters. This will result in false matching such as "abc=cba", and reduce the retrieval precision. To utilize the spatial relation between patches, Sivic et al. [1] uses a search area containing the 15 nearest neighbors of each matched patch, and the neighboring patch which also matches within this area casts a vote for that image. Philbin et al. [2] adds affine matrix verification to nearby patches using LO-RANSAC. To make full use of prior knowledge of image database, Zheng et al. [5] and Quack [6] both use data mining to exploit frequent items based on spatial relation. [5] proposes a visual phrase-based approach using adjacent patch pair, which is the most similar work to us. However, the neighborhood is defined as an intersecting pair of patches, which is hard to satisfy in images with sparse patches, and doesn't contain the information of distance and orientation.

Furthermore, in pretreatment process, patch detector often returns a lot of patches while only a few of them are distinguishable. So these informative patches need to be picked out through a sea of background patches.

To solve the above problems, the benefits of this paper are as follows. First, the spatial relation of patch pairs is described in log-polar space containing both distance and orientation information. Second, data mining method is used to avoid the inefficient pairwise matching over the whole dataset. Third, the pretreatment reserves only informative patches which are distinctive against background patches.

The remainder of the paper is organized as follows. Section 2 describes our method in detail. Experimental results are shown in section 3, and section 4 concludes this paper.

2. OUR METHOD

Our method picks out informative patches in preprocessing stage, and then describes the spatial relation of patch pairs in polar coordinates. Finally, the spatially frequent items are exploited using Apriori algorithm. In this way, spatial relation of patch pairs is constructed and used to conduct object retrieval. Following [7, 8], we use MSER method as region detector [9] and SIFT [8] to describe the image regions. They have been demonstrated to be superior to others used in object retrieval [1, 2].

2.1. Informative patches filtration

"Background patches" we mentioned here includes two kinds of patches: one kind has little information thus can be found in both foreground and background; the other comes from trees or grassplot which has complex texture but seldom occurs in the foreground. Inspired by [10], we use local entropy and self-similarity to remove them separately.

Given a patch X and its grey level distribution $D = \{d_1 \ldots\ldots d_r\}$, local entropy is defined as:

$$H_X = -\sum_{i=1}^{r} p(d_i) * \log_2 p(d_i)$$

where $p(d_i)$ is the probability of pixel taking the value d_i in patch X. Informative patches often have large entropy, so we remove those patches whose entropy is less than threshold $Entropy_{low}$.

As to the patches from trees or grass, due to their complex texture, entropy filtration itself is not enough. Because they have similar intensity distribution over large ranges of scale, we use self-similarity to remove them. For simplicity the sum of absolute difference of grey-level histograms with different scales is defined as self-similarity.

$$SS_X = \int_{i \in D} \left| \frac{\partial}{\partial s} p_D(s, X) \right| di$$

To prevent deleting informative patches by mistake, we use dual threshold method. Only those patch whose self-similarity is smaller than $SelfSimilarity_{low}$ and local entropy is also smaller than $Entropy_{high}$ will be removed. The proper values of these thresholds are discussed in section 3.

2.2. Spatial relation description

Based on the above stage, we extract SIFT descriptor for each informative patch, and obtain the visual vocabulary using K-Means clustering. Then each patch is denoted as $P_i = (pos_i, ellipse_i, vw_i)$, where pos_i is the coordinate of patch center, $ellipse_i$ includes information of the original MSER region, such as its major and minor axis semi diameters, vw_i denotes the visual word cluster that the patch has been clustered into.

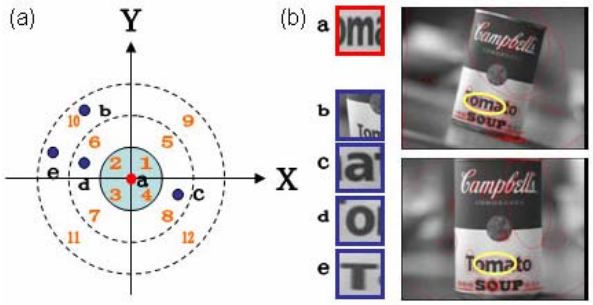

Fig.1 (a) Spatial relation description for central patch a (b) Neighborhood of a and corresponding patches in image

By considering spatial relation of neighboring patches, we can get higher discriminative power compared to individual ones. Considering variety affine transformation, we define spatial relation of patch ***a*** as shown in Fig.1. We use Matrix ***Ha*** to transform $ellipse_a$ to a unit circle $circle_a$ as [2], and calculate the new location of its neighboring patches' centers using ***Ha*** in this new coordinates. Instead of using a k-neighborhood or adjacent patches, all patches lie in R times the radius of $circle_i$ are called its neighbors. In this polar coordinate, we use 3 bins for R and 4 bins for orientation, making the spatial relation more sensitive to position and orientation of patch pairs. Thus the spatial relation of each patch p_i and its neighbor p_j is described as $pair_{ij} = < vw_i, vw_j : SR_{ij} >, SR_{ij} \in [1,...,12]$. Take Fig.1 for example, there are 4 pairs for central patch a: <a, b: 10>, <a, c: 8>, <a, d: 7>, <a, e: 10>. In practice, the patches are described using their visual words, such as <38, 79: 10>, etc. Here the asymmetry should require attention that the former must be the center patch, and different SR_{ij} means different pair. Note how the size and position can be made invariant by aligning the neighborhood structure with the center patch.

2.3. Spatially frequent items

Once the spatial relation description is defined, we can get a lot of patch-pairs from each image. However, this is not optimal, because there are many redundant pairs which seldom occur in the database, and only those frequent ones are distinctive, which are called spatially frequent items in this paper. We cope with this problem by frequent items mining using Apriori algorithm [5, 6].

Let ***I*** be a set of items, and ***A*** be a subset of ***I*** with m items. A transaction $\boldsymbol{T} \subseteq \boldsymbol{I}$ belongs to the transaction database ***D***. The support of ***A*** is defined as follows. An item ***A*** is called frequent in ***D*** if $supp(A) \geq s_{min}$ where s_{min} is the minimal support threshold.

$$supp(A) = \frac{|\{T \in D \mid A \subseteq T\}|}{|D|} \in [0,1]$$

Apriori algorithm takes advantage of the monotonicity property: all m-subsets of frequent (m+1)-sets are also

frequent. We define a patch pair to be spatial frequent items if and only if the number of images containing this pair exceeds a certain threshold θ. The construction algorithm is shown as follows.

For each local patch $P_i \in Patch_{ImageSet}$
  Count P_i.frequency;
  If P_i.frequency $> \theta$
    Add P_i to *1-itemset*;
For each $Pair_{ij} < p_i, p_j : SR_{ij} > \in PatchPair_{ImageSet}$
  If (($P_i \in$ *1-itemset*)&&($P_j \in$ *1-itemset*))
    Count $Pair_{ij}$.frequency;
    If $Pair_{ij}$.frequency $> \theta$
      Add $Pair_{ij}$ to *2-itemset*;
Spatially frequent items = *2-itemset.*

Once spatially frequent items have been constructed, we use them to conduct object retrieval. Considering there might be some single patches which are far from their neighbors thus don't have frequent adjacent patch, we don't use spatially frequent items directly to index images as in [5]. Instead, they are used to re-rank the result based on matching with individual patches as follows.

At first, all of the M images in the dataset are sorted using common Bag of Words algorithm such as [1, 2] to get N images as a small candidate set, N<<M. The process can be completed quickly; however, it is likely to result in many false matching due to the lack of spatial relation verification. Consequently, the spatially frequent items are used to re-rank this candidate, and remove those false matching.

In the re-rank stage, only when some spatially frequent item occurs both in the query region and candidate image ***i***, we consider there is a "successful" spatial verification in ***i***, and increase the weight of corresponding center patch. For each candidate image, the updated weight of successfully verified patch is recorded, and used to re-rank these candidate images. Experimental results demonstrate the effectiveness of our method.

3. EXPERIMENTAL RESULTS

In this section, the experimental result of our method is discussed in detail. The images used here are keyframes extracted from TRECVID 2005 news video retrieval database. Out of which 3000 images are selected. According to the objects they contain, these images are divided into 50 categories. The number of relevant images in each class ranges from about 20 to about 50 images, while the rest are thought to be disturbances. All the subsequent experiments are based on MSER region and SIFT descriptors.

3.1. Informative patches filtration

To measure the efficiency of patch filtration in section 2.1, we adopt exact point-to-point matching with SIFT in this sub-section. As shown in Fig.2, most of the false matchings due to background patches are removed correctly.

There are 3 thresholds in our filtration process, among which $Entropy_{\text{low}}$ influences the performance mostly. So we test its influence in a sample image set including 200 images and about 28k patches are extracted, while we define $SelfSimilarity_{\text{low}}$=0.5 and $Entropy_{\text{high}}$=3.0. The influence on patches quantity and matching precision (the ratio of correct matching and all matching pairs found in each image) are shown separately in table 1 and Fig 3. We can see that when $Entropy_{\text{low}}$=2.0, the best balance can be achieved, while a lot of redundant patches can be removed correctly, and matching precision would be guaranteed at the same time. Accordingly, these filtration thresholds are adopted throughout our subsequent experiment.

(a) Before filtration (b) After filtration

Fig.2 An example for filtration effect.

Tab.1 Comparison of patches quantity with $Entropy_{\text{low}}$

$Entropy_{\text{low}}$	Delete amount	Delete ratio
1.0	589	2.1%
1.5	2407	7.3%
2.0	3786	13.5%
2.5	5076	18.1%

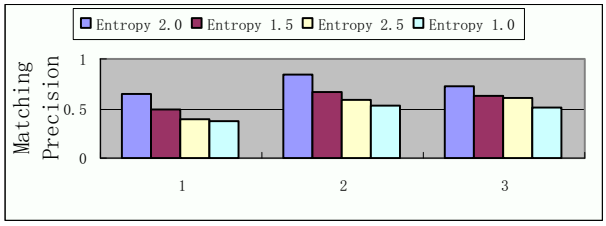

Fig.3 Comparison of matching precision with $Entropy_{\text{low}}$

3.2. Spatially frequent items based object retrieval

To evaluate our spatially frequent items-based approach's effectiveness in object retrieval, 10 categories are selected as query images due to the object's affine transformation, illumination, and even partially occlusion. Using a 3.2G Pentium 4 PC with 1.5G memory, the average spatial re-rank time for each query is 1.29 second. Some examples for object retrieval result are shown as Fig. 4, where query objects are demarcated using yellow rectangle in the left query images. The images of retrieval result are shown with descending object's similarities.

Fig.4 Examples for object retrieval result.

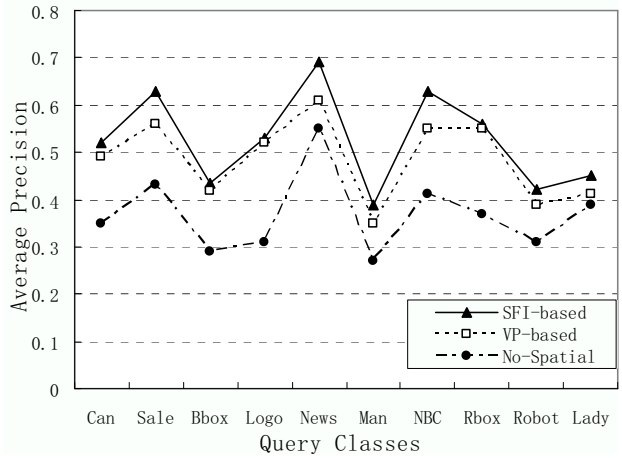

Fig.5 Average retrieval precision comparison

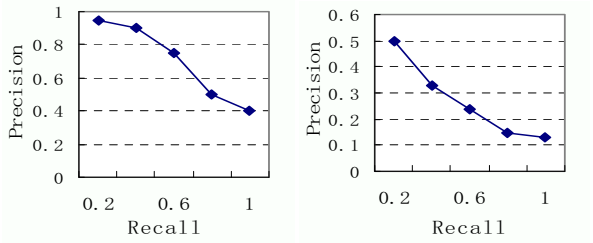

Fig.6 Precision–Recall curses for "News" (left) and "Man" (right)

We compare our SFI-based approach with VP-based method [5] and No-spatial result (BoW without spatial information) judged by AP (Average retrieval precision), where n denotes the amount of relevant images for query image $\boldsymbol{q}$ in each class, and p_j is the $\boldsymbol{j}$th retrieval result image:

$$AP_q = \frac{1}{n}\left(\sum_{i=1}^{n}\left(\frac{1}{i}\sum_{j=1}^{i}\psi(p_j,q)\right)\right),\quad \psi(p_j,q)=\begin{cases}1, & p_j \text{ is relevant to q}\\ 0, & p_j \text{ is not relevant}\end{cases}$$

As shown in Fig.5, our SFI-based method generally outperforms visual phrase- based approach with 8.6% mean average precision improvement, and both of them are much better than visual word-based approach. We attribute this to that spatially frequent items contain abundant spatial information between patches such as distance and orientation, while visual phrase only contains jointed patches. However, compare the best class "News" with the worst class "man", we can find that our method is more adopted for dense patch images such as "News" which including many distinctive and adjacent patch pairs, while in "man" there are few distinctive patches and they are often far from each other. Precision-Recall curves for the two classes are shown as Fig.6.

4. CONCLUSION

We have presented an approach called spatially frequent items for object retrieval. The spatial relations of adjacent patch pairs are described with their distance and orientation in polar coordinates. Based on this, we use Apriori algorithm to exploit spatially frequent items, and re-rank the BoW result with this information. Besides, we also pick out informative patches in preprocess stage using entropy and self-similarity, to remove background patches. Experimental result demonstrates that our method is efficient and outperforms the state-of-art object retrieval methods.

5. ACKNOWLEDGEMENT

This work was supported in part by the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416) and the National Basic Research Program of China (973 Program, 2007CB311100), the National Nature Science Foundation of China (60773056), the Beijing New Star Project on Science & Technology (2007B071).

6. REFERENCES

[1] J.Sivic, A.Zisserman. "Video Google: A Text Retrieval Approach to Object Matching in Videos", *ICCV 2003*

[2] Jams Phibin, Ondrej Chum, *et al*. "Object retrieval with large vocabularies and fast spatial matching", *CVPR 2007*

[3] C. Carson, et al. "Blobworld: A System for Region-based Image Indexing and Retrieval". *In 3rd Int. Conf. on Visual Information Systems*, Amsterdam, pp. 509-516, 1999

[4] S. Lazebnik, et al. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", *CVPR 2006*

[5] Qing-Fang Zheng, *et al*. "Effective and efficient object-based image retrieval using visual phrases", *ACM Multimedia*, Santa Barbara, USA, pp77-80, 2006

[6] Till Quack, Vittorio Ferrati, *et al*. "Efficient Mining of Frequent and Distinctive Feature Configurations", *ICCV07*

[7] K.Mikolajczyk, T. Tuytelaars, *et al*.. "A comparison of affine region detectors", *IJCV2006*, pp: 43-72

[8] K. Mikolajczyk, C. Schmid. "A performance evaluation of local descriptors". *IEEE Transaction on PAMI*. 2005

[9] J.Matas, O. Chum, *et al*. "Robust wide baseline stereo from maximally stable extremal regions", *BMVC 2002*, pp 384-393

[10] Timor Kadir, Michael Brady. "Saliency, Scale and Image Description", *International Journal of Computer Vision*. 45 (2):83-105, 2001