FAST AND ROBUST SPATIAL MATCHING FOR OBJECT RETRIEVAL

Wenying Wang[1,2], Dongming Zhang[1], Yongdong Zhang[1], Jintao Li[1]

(1 Advanced Computing Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences , Beijing, 100190)

(2 Graduate University of Chinese Academy of Sciences, Beijing, 100049)

ABSTRACT

Spatial matching for visual words based object retrieval often involves generating affine transformation hypotheses and then choosing the best hypothesis to measure the spatial consistency. In existing methods, generating an affine transformation hypothesis either requires three correspondences or assumes images are taken in restricted range of viewpoints in using a single correspondence. In this paper, we propose a novel spatial matching method, in which the transformation hypothesis can be estimated from only a single correspondence without the assumption of the viewpoints from which the images are taken. Firstly, affine covariant neighborhoods(ACNs) of features are used to eliminate possible false matches. Secondly, we decompose the affine transformation into three sub-transforms and conquer each sub-transform by exploiting the shape information and the ACNs of a single pair of corresponding features. Experiment results demonstrate that this method improves the average retrieval precision evidently with less computation in comparison with the previous methods.

Index Terms: affine transformations, object-based image retrieval, spatial matching, visual words

1. INTRODUCTION

Object-based image retrieval (OBIR) aims to construct a method that can effectively and efficiently pick out a list of images containing the desired target object from a large-scale image database. However, it is still a challenging problem due to the changes of viewpoint, illumination and even partial occlusion in the images. In recent years, the text-retrieval scheme[1-4] with "visual word"(VW) is growing very popular in OBIR. In this method, each image is represented as a bag of visual words (BOVW) and retrieved in a text-retrieval scheme. However, VWs cannot distinguish similar regions from different objects because of the information loss in quantization.

To improve the accuracy, the spatial relations of VWs are utilized in[1, 5, 6]. Sivic *et al.*[1] use a search area from K (=15) nearest neighbors(KNN) of each match, and each region which also matches within this area casts a vote for that image. Zheng *et al.*[7] propose the visual phrase-based approach to retrieve images containing desired objects. However, these methods, using only semi-local spatial constraints, cannot measure the global spatial consistency.

Global spatial consistency can be measured by estimating the affine transformations between the query object and the target images. D.Lowe *et al.*[8] use Hough Transform to estimate the affine transformation. This method is computationally expensive for it needs three matches to provide a solution. J.Philbin *et al.*[5] use the LO-RANSAC algorithm[9]. First, each 3, 4, or 5 dof(degree of freedom) affine transformation hypothesis is estimated by a single pair of corresponding local regions. Then, the 6 dof one is estimated from the inliers by performing least square solution. However, this method has two shortages: first, it assumes that images are taken from a restricted range of canonical views. This assumption limits its applicability, especially when images are taken from different viewpoints. Second, many false matches waste the computation cost.

To overcome the two shortages, we propose a fast and robust spatial matching method, in which each affine transformation hypothesis can be estimated from only a single pair of correspondence without the assumption of the viewpoints from which the images are taken. First, the affine covariant neighborhoods(ACNs) of features are used as semi-local spatial constraints to eliminate possible false matches. Therefore, the number of correspondences considered in estimating affine transformations is decreased. Second, we decompose the affine transformation into three sub-transforms and conquer each sub-transform by exploiting the shape information and the ACNs of a single pair of corresponding features.

The main contribution of this paper is as below: Firstly, we estimates each 6 dof affine transformation hypothesis directly by a single pair of corresponding VWs. Secondly, our method is more robust than Philbin's method, since our method discards the strong assumption of the viewpoints from which the images are taken. Finally, our method eliminates the possible false matches beforehand and does not require computing the least-square solutions in estimating affine transformations, so it decreases the computational cost substantially.

The rest of this paper is organized as follows: In Section 2, we describe our fast and robust spatial matching method for OBIR in detail. In Section 3, we present the experimental results and analysis, and finally we give the conclusion in Section 4.

Supported by National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416),National Nature Science Foundation of China (60873165、60802028),Beijing New Star Project on Science & Technology (2007B071)and Co-building Program of Beijing Municipal Education Commission

2. APPROACH

Our OBIR scheme includes a preprocess stage and a retrieval stage as in Figure 1. In the preprocess stage, first, we extract affine covariant local regions by MSER[10, 11] with high dimensional SIFT descriptors[12, 13]. Then, a number of randomly selected descriptors are clustered to form a VWs codebook by clustering method K-means. Finally, the descriptors of all images are quantified into VWs, and each image is indexed in inverted file[1] by the VWs for retrieval.

In the retrieval stage, a small amount of candidate images with a high probability containing the query object are firstly obtained by retrieving images in BOVW. Then the candidate target images are re-ranked by our global spatial matching method: First, ACNs of features are used to eliminate possible false matches. Second, the affine transformations are estimated to measure the global spatial consistencies.

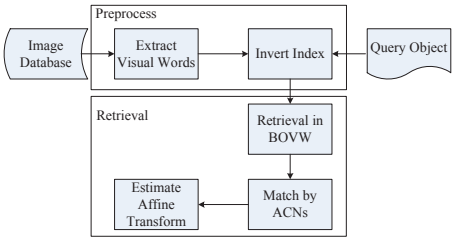

Figure 1: Image Retrieval Scheme

2.1 Matching Affine Covariant Neighborhoods

In the retrieval stage, the query object is represented by the VWs set $\boldsymbol{Q}=\{f_{o1},f_{o2},\ldots,f_{o,nq}\}$, and the candidate image t is represented by the VWs set $\boldsymbol{T}=\{f_{t1},f_{t2},\ldots,f_{t,nt}\}$, where n_q represents the number of the local regions from the query object and n_t, from the image t. The VW i is represented by a triple $f_i=(w_i,\boldsymbol{X}_i,\boldsymbol{\Sigma}_i)$, where w_i represents the corresponding index in the codebook, $\boldsymbol{X}_i=(x_i,y_i)^T$ the coordinate, and $\boldsymbol{\Sigma}_i$ the 2×2 shape matrix describing the elliptical region i, respectively. Then, we obtain a base matches set:

$$M_{init}(\boldsymbol{Q},\boldsymbol{T})=\{\langle f_{oi},f_{tj}\rangle \mid f_{oi}\in\boldsymbol{Q}\wedge f_{tj}\in\boldsymbol{T}\wedge w_{oi}=w_{tj}\} \quad (1)$$

where '$\wedge$' represents logic "AND" operation, w_{oi} and w_{tj} are the corresponding index of f_{oi} and f_{tj} in the codebook, respectively. Obviously, M_{init} contains many false matches.

The ACN uses Mahalanobis distances instead of Euclidean distances to make the K nearest neighbors covariant with affine transformation. For example, the ACN of the VW f_i is the neighboring VWs from the same image with K nearest affine covariant distances to the VW f_i:

$$\mathrm{ACN}(f_i)=\{f_k \mid AD(f_i,f_k)\le d_{Kth}\wedge f_i\in\boldsymbol{I}\wedge f_k\in\boldsymbol{I}\wedge k\ne i\} \quad (2)$$

where $\boldsymbol{I}$ represents $\boldsymbol{T}$ or $\boldsymbol{Q}$, d_{Kth} the threshold that equal to the K_{th} nearest distance, and $AD(f_i,f_k)$ the affine covariant distance from f_k to f_i:

$$AD(f_i,f_k)=[(\boldsymbol{X}_k-\boldsymbol{X}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}_k-\boldsymbol{X}_i)]^{1/2} \quad (3)$$

After obtaining the ACNs, the support set $S(\langle f_{oi},f_{tj}\rangle)$ of the match $\langle f_{oi},f_{tj}\rangle$ is computed to measure the preliminary spatial consistency:

$$S(\langle f_{oi},f_{tj}\rangle)=\{\langle f_{ok},f_{tl}\rangle \mid f_{ok}\in\mathrm{ACN}(f_{oi})\wedge f_{tl}\in\mathrm{ACN}(f_{tj}) \wedge\langle f_{ok},f_{tl}\rangle\in M_{init}\} \quad (4)$$

Therefore, the preliminary spatial consistent matches are obtained as:

$$M_{spa}(\boldsymbol{Q},\boldsymbol{T})=\{\langle f_{oi},f_{tj}\rangle \mid \langle f_{oi},f_{tj}\rangle\in M_{init}(Q,T)\wedge|S(\langle f_{oi},f_{tj}\rangle)|\ge th\} \quad (5)$$

where $|S(\langle f_{oi},f_{tj}\rangle)|$ represents the support set size and th the threshold.

2.2 Estimating Affine Transformations

To measure the global spatial consistency, we generate each 6 dof affine transformation hypothesis using a single pair of correspondences and then evaluate each hypothesis by the number of "inliers" among all features under the hypothesis. In estimating affine transformation, we decompose the affine transformation into three sub-transforms, and then conquer the three one by one.

Decomposing Affine Transform Matrix: We decompose the affine transform matrix by analyzing the shape matrices of the corresponding local regions. For the elliptical local regions oi and tj with the general elliptical equations

$$(\boldsymbol{X}-\boldsymbol{X}_a)^{\mathrm{T}}\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{X}-\boldsymbol{X}_a)=1 \quad (a=oi \text{ or } tj) \quad (6)$$

we decompose the two elliptical shape matrices as introduced in[14]:

$$\boldsymbol{\Sigma}_a=\boldsymbol{C}_a\boldsymbol{C}_a^T=\boldsymbol{C}_a\boldsymbol{R}_a\boldsymbol{R}_a^T\boldsymbol{C}_a^T \quad (a=oi \text{ or } tj) \quad (7)$$

where $\boldsymbol{C}_a=\boldsymbol{\Sigma}_a^{1/2}$ represents the 2×2 normalization transform matrix and $\boldsymbol{R}_a$ the 2×2 rotational transformation matrix.

If the local region tj is the image of oi taken in different viewpoint, both the two local regions can be transformed to the same unit circle region u as follow:

$$\boldsymbol{X}_u=\boldsymbol{R}_{oi}^{-1}\boldsymbol{C}_{oi}^{-1}(\boldsymbol{X}_o-\boldsymbol{X}_{oi}) \quad (8)$$

$$\boldsymbol{X}_u=\boldsymbol{R}_{tj}^{-1}\boldsymbol{C}_{tj}^{-1}(\boldsymbol{X}_t-\boldsymbol{X}_{tj}) \quad (9)$$

$$\Rightarrow \boldsymbol{X}_t=\boldsymbol{X}_{tj}+\boldsymbol{C}_{tj}\boldsymbol{R}_{tj}\boldsymbol{R}_{oi}^{-1}\boldsymbol{C}_{oi}^{-1}(\boldsymbol{X}_o-\boldsymbol{X}_{oi}) \quad (10)$$

where $\boldsymbol{X}_o$ represents the pixel coordinate on the elliptical region oi, $\boldsymbol{X}_t$, on the elliptical region tj, and $\boldsymbol{X}_u$, on unit circle region u respectively.

Then, we obtain the affine transformation matrix $\boldsymbol{A}$ as follow:

$$\tilde{\boldsymbol{X}}_t=\boldsymbol{A}\tilde{\boldsymbol{X}}_o \quad (11)$$

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{C}_{tj} & -\boldsymbol{C}_{tj}^{-1}\boldsymbol{X}_{tj}\\ \boldsymbol{0} & 1\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{R}_{tj}\boldsymbol{R}_{oi}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{C}_{oi}^{-1} & -\boldsymbol{C}_{oi}^{-1}\boldsymbol{X}_{oi}\\ \boldsymbol{0} & 1\end{bmatrix}$$

$$=\boldsymbol{H}_2^{-1}\boldsymbol{H}_r\boldsymbol{H}_1 \quad (12)$$

where $\tilde{\boldsymbol{X}}$ represents the homogeneous coordinate.

As a result, the affine transformation $\boldsymbol{A}$ is decomposed into three sub-transforms: the normalization transform $\boldsymbol{H}_1$, the rotational transform $\boldsymbol{H}_r$ and the inverse of the normalization transform $\boldsymbol{H}_2$. Though both the $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ can be obtained from the shape matrix $\boldsymbol{\Sigma}_{oi}$ and $\boldsymbol{\Sigma}_{tj}$ directly, the rotational transformation $\boldsymbol{H}_r$ requires estimating the precise rotational angle. The orientations in SIFT descriptors are used in the previous method[15]. However, as quantized into a small amount of orientations, the SIFT orientations can not give a precise rotational angle.

Estimating the Rotation Transform $\boldsymbol{H}_r$: We exploit the support set of the match $\langle f_{oi},f_{tj}\rangle$ to estimate the rotational angle precisely. When transformed by $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ respectively, the elliptical regions oi and tj are transformed to overlap on the unit circle in

different orientations and the neighboring VWs N_{ik} and N_{jk} in the ACNs are moved to the new coordinates L_{ik} and L_{jk} respectively near the origin $\boldsymbol{O}$. As shown in Figure 2, the support match $\langle N_{ik}, N_{jk}\rangle$ in $S(\langle f_{oi}, f_{tj}\rangle)$ provides one rotational angle $\alpha_k=\angle L_{ik}OL_{jk}$. Therefore, the coordinate space of VWs in the support set is converted into the rotational angle parameter space, and the Hough Transform is performed to calculate the rotational angle α_{ij}. This step does not deteriorate the computational complexity since the ACN has only a small number of elements. In short, the rotational angle is estimated by exploiting the ACNs precisely.

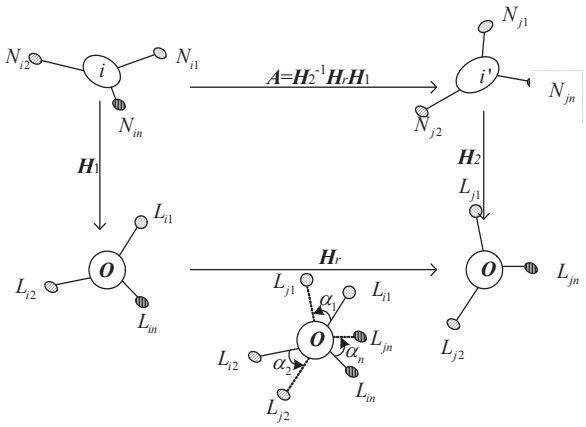

Figure 2: Rotational Angle Estimation Method by ACN

In summary, each 6 dof affine transformation hypothesis, as shown in Figure 3, is estimated from a single pair of corresponding elliptical local regions by using the shape information and ACNs.

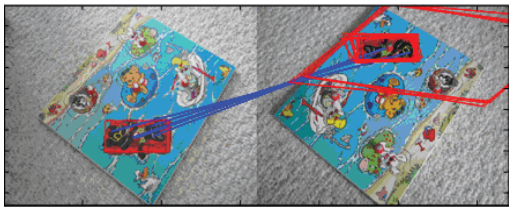

Figure 3: each pair of corresponding local regions provides one affine transform matrix

3. EXPERIMENTS

We compared our spatial matching method with the KNN method [1]and Philbin's method[5] on the Recognition Benchmark [2, 5] image database. This database consists of 10200 images, in which every four images come from the same scene with different conditions such as different viewpoints and light conditions. All experiments are conducted on the Pentium4, CPU3.2 GHz with 2G RAM computer system running windows XP.

3.1 Experiment Method

We randomly choose 120 query regions for the OBIR test and measure the retrieval precision by the percentage of right images in the top 4 ranked target images.

In the preprocess stage, after extracting local regions by MSER[10] with the 128 dimensional SIFT descriptors, we choose 100,000 descriptors randomly for clustering to construct a 1,000 VWs codebook by K-means and quantify all the descriptors into VWs by the codebook. Then, we index each image in inverted file by the VWs. In the retrieval stage, we firstly retrieve images in BOVW as introduced in[1], then re-rank the top 128, 256, 384, and 512 candidate target images respectively by KNN method[1], Philbin's method[5], and our method. We implement Philbin's LO-RANSAC algorithm as introduced in[5] with 5 dof affine transformations hypothesis initially estimated and the 6 dof affine transformation is estimated by performing least-square solution. In our method, the parameter K=10, th=3, and d_{th}= 4.

3.2 Experiment Results and Analysis

The three methods, KNN, Philbin's method, and our method, improve the precision of the BOVW method without spatial constraints substantially as shown in Figure 4. Our method improves the retrieval precision extensively in comparison with Philbin's method[5], since our method discards the strong assumption that Philbin's method makes and many images in the database are taken in different viewpoints. Furthermore, our method and Philbin's method, measuring the global spatial consistency, outperform the semi-local spatial matching method KNN.

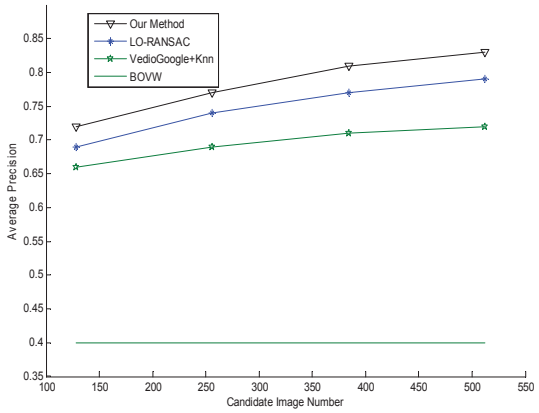

Figure 4: Average Retrieval Precision

The retrieval time, as shown in Figure 5, increases approximately linear to the number of the candidate images and is kept in 1 second. From Figure 5, we also see that our method decreases the computational cost in comparison with Philbin's method. Two factors cause this speedup: firstly, in our method, the semi-spatial constraints ACNs are used to remove the possible false matches and so decrease the number of the corresponding local regions need to be considered in estimating affine transform extensively. Secondly, our method does not perform the least-square solutions to estimate the 6 dof affine transformations.

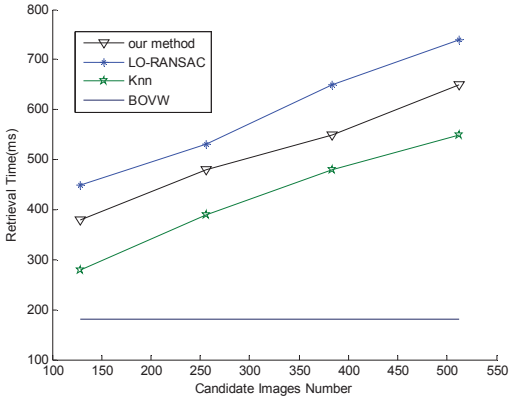

Figure 5: Average Retrieval Time

Some retrieval results are shown in Figure 6; the results demonstrate that our method estimates the affine transform matrices robustly, although some images are rotated and even flipped. In contrast, the images with "X" are not detected by Philbin's method. We can see that our method is more robust than Philbin's method, especially when images are taken from different viewpoints.

Figure 6: Some Retrieval Results

4. CONCLUSION

In this paper, we propose a fast and robust spatial matching method for OBRI. We estimate the affine transformation matrices novelly exploiting the ACNs and the shape information of local regions. Our method outperforms previous methods. Firstly, our method estimates each 6 dof affine transformation hypothesis by using only a single pair of corresponding local regions, while Lowe's method[8] needs three matches to provide a solution. Secondly, our method is more robust than Philbin's method especially when the photos are taken from different viewpoints. Finally, as our method eliminates possible false matches beforehand and does not require performing the least-square solutions, it decreases the computational cost. The experiment results demonstrate that our method improves the performance in both precision and speed.

5. REFERENCE

[1] J. Sivic, A. Zisserman. Video google: A text retrieval approach to object matching in videos; ICCV, Washington D C, 2003.

[2] D. Nister, H. Stewenius. Scalable recognition with a vocabulary tree; CVPR, New York F,2006.

[3] J. Sivic, A. Zisserman. Efficient visual search of videos cast as text retrieval. PAMI, 2009, 31(4): 591 - 606.

[4] O. Chum, M. Perdoch, J. Matas. Geometric min-hashing-finding a (thick) needle in a haystack; CVPR, Miami, 2009.

[5] J. Philbin, O. Chum, M. Isard, et al. Object retrieval with large vocabularies and fast spatial matching; CVPR, Minneapolis, 2007.

[6] H. Jegou, M. Douze, C. Schmid. Hamming embedding and weak geometric consistency for large scale image search; ECCV, Marseille, 2008.

[7] Q.-F. Zheng, W.-Q. Wang, W. Gao. Effective and efficient object-based image retrieval using visual phrases. ACM MM. Santa Barbara. 2006: 77-80.

[8] D.G. Lowe. Object recognition from local scale-invariant features; ICCV, Kerkyra, 1999.

[9] O. Chum, J. Matas, and ˇ S. Obdrˇz´alek. Enhancing RANSAC by generalized model optimization. ACCV, 2004.

[10] J. Matas, O. Chum, M. Urban, et al. Robust wide baseline stereo from maximally stable extremal regions. Image and Vision Computing, 2004, 22(10): 761-767.

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, et al. A comparison of affine region detectors. IJCV, 2006, 65(1): 43-72.

[12] K. Mikolajczyk, C. Schmid. Scale & affine invariant interest point detectors. IJCV, 2004, 60(1): 63-86

[13] D.G. Lowe. Object recognition from local scale-invariant features; ICCV, Kerkyra, F, 1999.

[14] J.R.ı. Matas, P. B´ılek, O.R. Chum. Rotational invariants for widebaseline stereo. CVWW. Wien, Austria. 2002.

[15] G. Carneiro, A.D. Jepson. Flexible spatial configuration of local image features.PAMI, 2007, 29(12): 2089-2104.