Invariant matching method for different viewpoint angle images

Min Chen,[1] Zhenfeng Shao,[1,*] Dongyang Li,[1] and Jun Liu[2]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, No. 129 Luoyu Road, Wuhan, Hubei 430079, China

[2]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, No. 85 Jinyu Road, Chongqing 400120, China

*Corresponding author: shaozhenfeng@163.com

Received 20 March 2012; revised 17 October 2012; accepted 19 November 2012; posted 29 November 2012 (Doc. ID 165144); published 21 December 2012

In recent years, many methods have been put forward to improve the image matching for different viewpoint images. However, these methods are still not able to achieve stable results, especially when large variation in view occurs. In this paper, an image matching method based on affine transformation of local image areas is proposed. First, local stable regions are extracted from the reference image and the test image, and transformed to circular areas according to the second-order moment. Then, scale invariant features are detected and matched in the transformed regions. Finally, we use epipolar constraint based on the fundamental matrix to eliminate wrong corresponding pairs. The goal of our method is not to increase the invariance of the detector but to improve the final performance of the matching results. The experimental results demonstrate that compared with the traditional detectors the proposed method provides significant improvement in robustness for different viewpoint images matching in the 2D scene and 3D scene. Moreover, the efficiency is greatly improved compared with affine scale invariant feature transform (Affine-SIFT).

OCIS codes: 150.0150, 100.2000.

1. Introduction

Image matching is a fundamental issue in computer vision. It has been used in target tracking [1], image stitching [2], 3D reconstruction [3], and so on. Image matching algorithms are mainly divided into three categories: pixel grayscale-based, transform domain-based, and feature-based. The methods based on image pixel-values are sensitive to image intensity changes and geometric deformations. The transform domain-based methods are suitable for the scale change and rotation of images, but they are not robust to the local distortions. The feature-based algorithms can overcome the disadvantages of the previous two kinds of methods by matching local features extracted from images, such as interest points, lines, and areas.

Generally speaking, the framework of feature-based methods consists of three steps: feature detection, feature description, and feature matching. Researchers have presented many practical methods for the three parts. For feature detection, the famous methods include corner detectors like Harris [4], Harris–Laplace [5], Harris-Affine [5], smallest univalue segment assimilating nucleus [6], blob detectors like difference of Gaussians (DoG) [7], speeded up robust features [8], Hessian [9] and region detectors like maximally stable extremal region (MSER) [10], intensity extrema-based region [11], and so on. For feature description, owing to the robustness about scale, blur, and rotation of images, scale invariant feature transform (SIFT) [7] and histograms of oriented gradient [12] are more popular than other descriptors. Feature matching is to find the nearest correspondence.

1559-128X/13/010096-09$15.00/0

A handful of distances can be used in practice, such as L_1 distance, L_2 distance, and histogram intersection distance [13]. The above three steps determine the performance of image matching together. Although many methods have been proposed and brilliant success has been achieved, there is no feature detector mentioned above that is fully invariant to the change of image view [14,15,16]. This issue motivates us to think and study the matching of different viewpoint images. Apparently, the similarity of the same object shown on images will become smaller and smaller with the increasing of viewpoint change. Therefore, we do not attempt to extract invariant features from the different viewpoint images but focus on how to improve the framework of image matching in order to obtain better results.

In this paper, we propose a novel image-matching framework that is robust to the variation of view. First of all, local regions are extracted from the reference image and the test image and fitted into ellipses. Then, according to the ellipse parameters and the second-order moment, all the elliptical regions are transformed to circular areas. In this case, there are only scale change and rotation left between the corresponding regions where perspective transformation used to exist. Thus we use the scale invariant DoG detector and SIFT descriptor to calculate and compare features in the circular fields. Then the random sample consensus (RANSAC) algorithm based on the fundamental matrix is used to eliminate wrong corresponding pairs from the initial matches. The contribution of our approach is to present a new solution for the matching of different viewpoint images, which ensures satisfactory results and high efficiency at the same time.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction to related work. Then, a novel image matching strategy is presented in Section 3. Contrast experiments between the proposed method and other methods are shown in Section 4. Section 5 concludes this paper and gives insight on future work.

2. Related Work

In order to deal with the matching of different viewpoint images, the traditional methods is to improve the affine invariance of feature detector or descriptor, such as SIFT, Harris-Affine, Hessian-Affine [5], MSER, and so on. In [7], a Gaussian weighting function is used to assign a weight to the gradient magnitude of each sample point when computing SIFT descriptor. It gives less emphasis to gradients that are far from the center of the descriptor. As a result, the problem caused by SIFT without affine invariance can be offset partially. In [5], a set of initial points extracted at their characteristic scales based on the Harris–Laplace detection scheme are input to a Harris-Affine detector, and an iterative algorithm is applied to adjust the location, scale, and local area of every point so as to get an affine invariant point. Through the imitation of Harris-Affine, another affine invariant detector, Hessian-Affine, is proposed. The difference is that it starts from the Hessian rather than the Harris corners. In [10], the concept of terrain watershed is introduced to extract MSERs. The MSERs are the parts of the image where local binarization is stable over a large range of thresholds. The definition of MSER stability based on relative area change is invariant to affine transformations.

In recent years, another feasible solution to cope with the change of view in image matching is simulating the original image to every possible view, extracting features and matching respectively. In [14,15], Morel and Yu have proposed a fully affine invariant framework, Affine-SIFT (ASIFT), for different viewpoint images matching. ASIFT simulates the reference image and the test image to cover the whole affine space. Then, SIFT is used to extract and compare features from these simulations. After feature matching, the correspondent pairs are converted to the original images. ASIFT can find matches from the images even if they are much different in view. However, the important drawback of ASIFT is the computational complexity. In [16], in order to balance the matching results and time efficiency, another matching framework, iterative SIFT (ISIFT), for different viewpoint images is proposed. Through the iterative algorithm, the geometric transformation between the image pair is estimated. According to the estimated model, the test image (or the reference image) is simulated. Then the reference image (or the test image) is matched with the simulated image. And the matching results are converted to the original images.

3. Proposed View Invariant Image Matching Method

A. Analysis of Transformation between Different Viewpoint Images

The transformation between a 3D space point and the corresponding appearance on the image can be described by the Holes perspective projection-imaging model. The relationship between the pixel coordinates and world coordinates can be expressed as

$$s[u, v, 1]^T = \mathbf{M}_1\mathbf{M}_2[X, Y, Z, 1]^T, \tag{1}$$

where s is the depth coordinate in projection direction of the camera coordinate system, the matrix $\mathbf{M}_1$ denotes the internal parameters of the camera including focal length, resolution, and the image plane offset, and the matrix $\mathbf{M}_2$ denotes the translation and rotation between the camera coordinate system and the world coordinate system called extrinsic parameters.

As Eq. (1) shows, the pixel coordinates of one point in the world coordinate system can be determined by the camera intrinsic parameters and extrinsic parameters. When a point in 3D space is viewed from different angles, the relationship between the pixels coordinates of the point in the images can be obtained by the camera imaging model as shown in Eq. (2):

$$\begin{cases} u_2 = a_{11}u_1 + a_{12}v_1 + b_1 \\ v_2 = a_{21}u_1 + a_{22}v_1 + b_2 \end{cases}. \tag{2}$$

The camera intrinsic parameters and extrinsic parameters are equal for every point, which can be seen as constant. But the coefficients a_{ij} and b_i in Eq. (2) are still related to the depth coordinates s_1 and s_2. Obviously, when the ratio s_1/s_2 of the two corresponding pixels in the image pair is constant, the coefficients in Eq. (2) are constant. Then the geometric relationship between all pixels in the image pair can be expressed by an affine transform. In practice, if the fields of view of the cameras are small or the depth variation in local object is little, the projection transformation of the corresponding regions can be approximated by an affine transform because s_1/s_2 is proximate to constant.

According to the analysis above, the geometric transformation between the corresponding local regions shown in Fig. 1 can be approximated by an affine transform that is expressed by matrix **A** [14]:

$$\mathbf{A} = \mathbf{H}_\lambda \mathbf{R}_1(\psi)\mathbf{T}(\theta)\mathbf{R}_2(\phi), \tag{3}$$

where $\mathbf{H}_\lambda$ denotes the scale change, $\mathbf{R}_1(\psi)$ denotes the rotation, and $\mathbf{T}(\theta)\mathbf{R}_2(\phi)$ denotes the variation of view.

In Fig. 1, the two local elliptical regions are transformed to circular areas, respectively, and matrix **B** is used to describe the transformation between them. According to [17],

$$\mu'_L = \mathbf{B}^T \mu'_R \mathbf{B}, \tag{4}$$

where μ'_L and μ'_R denote the second-order moments of the two circular areas. For the circular fields,

$$\mu'_L = \lambda_L E, \qquad \mu'_R = \lambda_R E. \tag{5}$$

It can be deduced from Eqs. (4) and (5) that

$$(\lambda_L/\lambda_R)E = \mathbf{B}^T\mathbf{B}. \tag{6}$$

Therefore, there are only scale change and rotation between the two areas, which could be mathematically expressed by

$$\mathbf{B} = \mathbf{H}_{\lambda'}\mathbf{R}_1(\psi'). \tag{7}$$

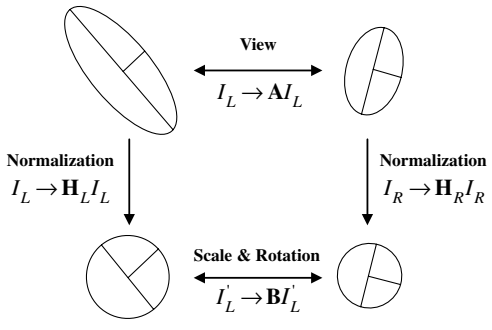


Fig. 1. Transformation between a pair of correspondent local areas.

B. Two Steps of the Proposed Method

1. *Region Extraction and Transformation*

In order to obtain the circular areas, local regions need to be extracted from the images at first. At present, many methods can be used to get local regions, such as image segmentation and region feature detection. In this paper, we choose the latter because image segmentation is still an unsolved problem and the segmentation accuracy cannot be guaranteed. In consideration of the fact that MSER has the advantages of high efficiency and high feature repeatability rate on the viewpoint change condition [18], it is selected to detect local regions.

In the original MSER algorithm, although the extracted MSERs come in many different sizes, they are all detected at a single image resolution. When a scene is blurred or viewed from increasing distances, many details in the image disappear and different region boundaries are formed. The detected MSERs will be different. In order to improve the scale invariance of MSER, we use the multiresolution strategy to detect MSERs from different resolutions instead of detecting MSERs only in the original input image [19]. First, a scale pyramid is constructed by blurring and subsampling with a Gaussian kernel. Then, MSERs are detected separately at each resolution image according the method proposed in [10]. Finally, duplicate MSERs are removed by eliminating fine scale MSERs with similar locations and sizes as MSERs detected at the next coarser scale. The criteria we used to eliminate MSERs is that the distance between the centroids should be smaller than four pixels in the finer grid and the two areas S_1 and S_2 should be $\mathrm{abs}(S_1 - S_2)/\max(S_1, S_2) < 0.2$. Besides, the minor axis of the elliptical area should be larger than 5 except on the finest scale.

The detected elliptical multiresolution MSERs are transformed to circular areas according to the values of macro axis, minor axis, and the second-order moment by the method as follows:

Assuming the macro axis is l and the minor axis is w, we set the radius of the transformed circular area as $r = \sqrt{l \cdot w}$. Matrix **H** is used to express the geometric transformation between the elliptical region and the circular area, so **H** satisfy the formula

$$[\mathbf{H}(\mathbf{X} - \mathbf{X_g})]^{\mathbf{T}}[\mathbf{H}(\mathbf{X} - \mathbf{X_g})] = r^2, \tag{8}$$

where **X** is a point on the ellipse and $\mathbf{X_g}$ is the center of the ellipse. Since **X** is on the ellipse, thus

$$(\mathbf{X} - \mathbf{X_g})^{\mathbf{T}}\boldsymbol{\mu}^{-1}(\mathbf{X} - \mathbf{X_g}) = 1, \tag{9}$$

where $\boldsymbol{\mu} = [\mu_{20}, \mu_{11}; \mu_{11}, \mu_{02}]$ is the second-order moment of the elliptical region. Calculate the Eqs. (8) and (9):

$$\mathbf{H} = \frac{r}{[\mu_{20}(\mu_{20}\mu_{02} - \mu_{11}^2)]^{1/2}} \begin{bmatrix} (\mu_{20}\mu_{02} - \mu_{11}^2)^{1/2} & 0 \\ -\mu_{11} & \mu_{20} \end{bmatrix}. \tag{10}$$

Then, the elliptical region can be mapped in a circular area with the center $\mathbf{X_g}$ and the radius r by the transformation matrix $\mathbf{H}$.

2. *Feature Extraction and Image Matching*

Because the number of MSERs is small and the localization accuracy of fitted ellipse area is low relatively, we define the circular MSERs as rough features. More point features with higher localization accuracy defined as fine features are detected in the rough features. Because only scale change and rotation exist between the corresponding rough features, we use the scale invariant DoG detector to extract fine features and choose the SIFT feature descriptor to describe and obtain the initial matching results.

There are some incorrect correspondences inevitably in the initial results. The traditional method is to estimate the homography between the two images by using RANSAC algorithm. Those initial matches do not conform the estimated model will be eliminated as false matches. However, the analysis in Section 3.A shows that generally the geometric transformation between all pixels of the two images cannot be approximated by an affine transform. In this condition, if the homography is used to find wrong matches, many correct matches will be eliminated at the same time. In this paper, in order to overcome this problem, epipolar constraint based on the fundamental matrix is used to eliminate wrong corresponding pairs with RANSAC.

Concerning two images projected from a scene, the point p in the reference image and the point q in the test image are a pair of correspondence related with $p^T F q = 0$, where F is a 3×3 fundamental matrix and p and q are represented with homogeneous coordinate. The Eq. (11) is used to calculate the correspondences set and transformation at each iteration [20],

$$G(F) = \{(p,q) \in S | d^2(q,Fp) + d^2(p,F^T q) < \varepsilon^2\}, \tag{11}$$

where d denotes the distance from the point to the epipolar line, ε denotes distance threshold, F is the fundamental matrix at the current iteration, and S is the correspondences set related to F. The procedure of fundamental matrix computation and mismatched points elimination by applying RANSAC is as follows:

Step 1: Select seven correspondences from the initial matches randomly to construct a sample set;
Step 2: Compute the fundamental matrix F with current sample set;
Step 3: Test all the initial matches and save the correspondences that satisfy the Eq. (11) into a set S;
Step 4: If correspondences in current set S are more than the previous one, the current matrix F and correspondences set S are preserved and discard the previous matrix F and set S;
Step 5: After the iterative calculation, the final matrix F is the fundamental matrix between the two images, and the final correspondences set S contains the correct matches.

In the proposed matching method, region extraction and transformation is a crucial step. We provide an experiment below to demonstrate that more correct matches can be obtained by mapping an elliptical feature area into a circular area. The test images are shown in Fig. 2.

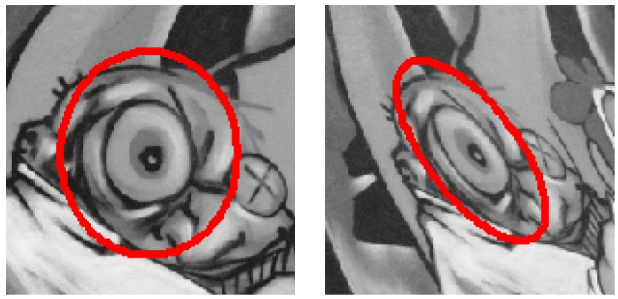
In the two images, the ellipses are two MSERs. The numbers of detected features, repeated features, and correct matches of the method matching in the elliptical areas and the method matching in the circular areas are shown in Table 1.

From the result in Table 1, we can see that more features have been extracted in the elliptical areas. However, the number of repeated features in the elliptical areas is about half of that in the circular areas. As a result, many more correct matches have been found between the circular areas. It demonstrates that mapping an elliptical feature area into a circular area can help matching feature better.

C. Relationship of the Proposed Method, ASIFT and ISIFT

Among the matching methods for different viewpoint images, ASIFT, ISIFT, and the proposed method all focus on the improvement of the matching framework, as shown in Fig. 3. We will analyze the relationship of them as follows.

ASIFT is a fully affine invariant matching framework because the simulations cover the whole affine space and parts of the simulations of the reference image, and the test image should have similar poses in the affine space. The number of the final matches increases with the number of the simulations. ASIFT indeed increases the invariability of the image matching method. However, the transformation between the two images is not considered and exhaustive matching strategy is used in ASIFT. To achieve



(a) Left image **(b) Right image**

Fig. 2. (Color online) Two images with viewpoint change.

Table 1. Test Result of the Two Images in Fig. 2

Matching Method	Detected Features	Repeated Features	Correct Matches
Elliptical area	114/83	28	6
Circular area	111/79	53	44

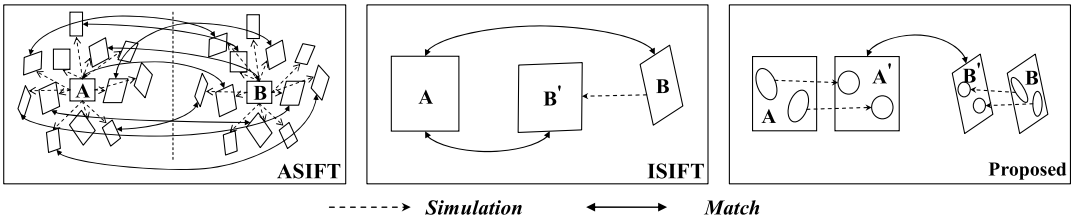


Fig. 3. The relationship among the general framework, ASIFT, ISIFT and the proposed method.

view invariant, the algorithm efficiency is reduced at the cost. Therefore, it is limited in many applications.

ISIFT is an iterative method. A set of correspondences is obtained at the first iteration, and the geometric transformation between the reference image and the test image is estimated using the initial correspondences. Then, one image of the pair is simulated according to the transformation. SIFT method is used to match the simulated image and another image. ISIFT also constructs simulation to improve the matching results. But generally it simulates the reference image (or the test image) once, and the only one simulation and another image are matched. Compared with the many simulations and exhaustive matching strategy of ASIFT, ISIFT improves the time efficiency greatly. Unfortunately, the drawback of ISIFT is that it does not increase the invariability of the original SIFT method. The final performance depends on the initial matching results. If the transformation between the two images can be approximated by an affine transform and there are more than three correct correspondences in the initial matches, the homography can be calculated and satisfactory results can be obtained with ISIFT. Otherwise, ISIFT fails.

Like the two methods mentioned above, our approach also adopts the idea of simulating to improve the matching performance for different viewpoint images. However, it is different from ASIFT, which simulates every pose in the affine space without the consideration of the transformation between the original images. In our method, each local region is related to one circular simulation. In this way, most superfluous simulations input to feature detection are eliminated. Therefore, the time efficiency of the method is improved greatly. It is also different from ISIFT, which simulates the whole image with the same affine transform. In the proposed method, each local region is simulated with different transformation. Whether the two whole images conform the same transformation or not, the transformations between pixels in the corresponding local regions can be approximated by the same transform. Benefiting from our new solution, the proposed method works well when ISIFT fails since the homography of the two images cannot be obtained.

4. Experimental Results and Analysis

A. Database

In our experiments, in order to evaluate the performance of the proposed image matching framework, first we compare it with the traditional approaches using the detectors SIFT, Harris-Affine, Hessian-Affine, and MSER combined with the most popular SIFT descriptor based on four datasets provided by Mikolajczyk [21], Morel [22], and Matas [23]. Then we compare it with the novel frameworks ASIFT and ISIFT. The reference images of the four datasets are shown in Fig. 4. Dataset (a) is provided by Mikolajczyk, in which the camera varies from a fronto-parallel view to one with significant foreshortening at 60 deg. There is an image in frontal view and eight in slanted view with the angles from 10 to 80 deg in dataset (b) and an image in frontal viewed and four images with viewpoint changes of 45, 65, 75, and 80 deg in dataset (c), respectively. The two datasets are provided by Morel. And the last dataset is provided by Matas, which contains two images of 3D scene with large viewpoint angle.

B. Matching Results

In this paper, we mark the proposed framework based on multi-resolution MSERs as MM-SIFT (MM-SIFT denote multi-resolution MSERs and SIFT). Parameters in the extraction of multi-resolution MSERs are set as follows: the scale pyramid is constructed by blurring and subsampling with a six-tap Gaussian kernel with $\sigma = 1.0$ pixels; the maximum relative area is 0.01; the minimum size of output region is 50 at the first level pyramid image and it decreased at other coarser scales with the image sub-sampling; and the minimum margin is 20.

1. Comparison between Proposed and Traditional Methods

We compare the proposed MM-SIFT with traditional methods using the detectors SIFT, Harris-Affine, Hessian-Affine, and MSER combined with SIFT descriptor on the four image groups to evaluate our method for 2D scene and 3D scene. Figure 5 summarizes the performance of each algorithm in terms of number of correct matches. In order to facilitate display, in Fig. 5(b) only the results of images with viewpoint change beyond 30 deg are presented. For the four datasets, the results of the proposed MM-SIFT between the reference image and the test image with largest view angle are illustrated in Fig. 6.

Experimental results in Fig. 5(a) show that in image group graf the proposed method is very robust to viewpoint change and the correct matches are more than that of the traditional detectors SIFT, Harris-Affine, Hessian-Affine, and MSER combined with

(a) Graf (b) Magazine

(c) Adam (d) Wash

Fig. 4. (Color online) Reference images of the datasets used for the experimental evaluation.

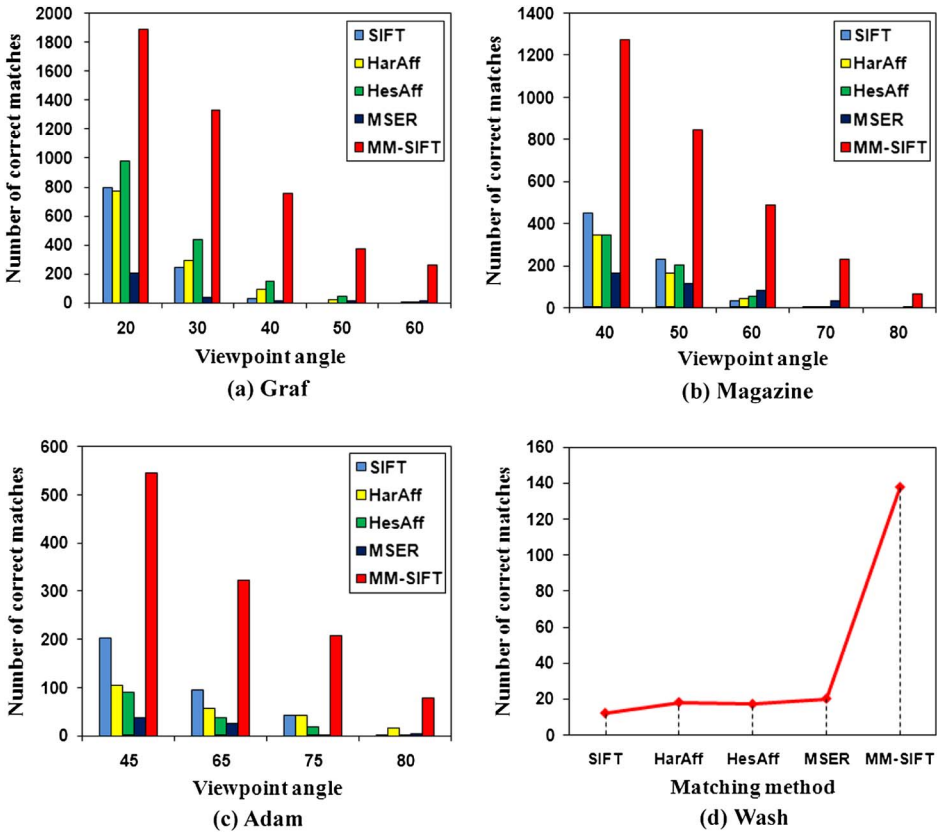
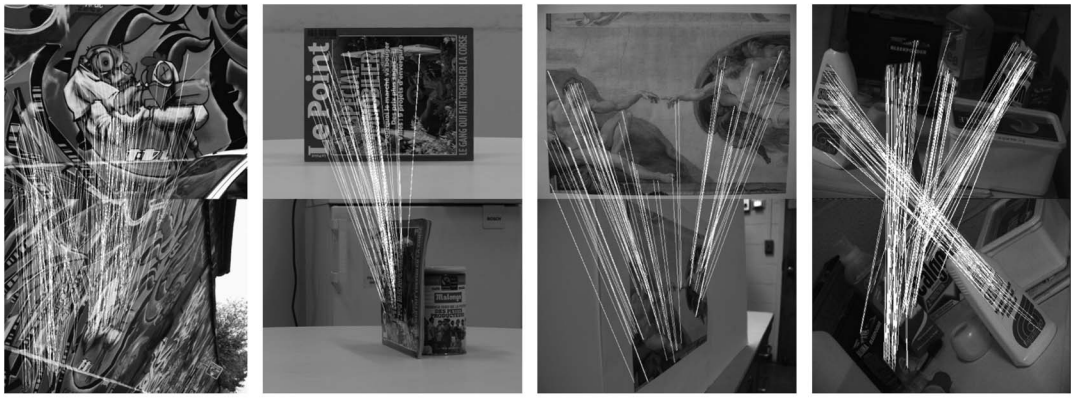


Fig. 5. (Color online) Number of correct matches of different viewpoint images in the four image datasets with approaches SIFT, Harris-Affine, Hessian-Affine, MSER, and MM-SIFT, respectively.



(a) (b) (c) (d)

Fig. 6. Partial matching results of the proposed MM-SIFT among the four image groups. (a) Shows the result of the frontal view and 60 deg in graf. (b) Shows the result of the frontal view and 80 deg in magazine. (c) Shows the result of the reference image and the test image with viewpoint angle of 80 deg in Adam. (d) Shows the result of wash.

SIFT descriptor at every view pose. When the view variation is not larger than 40 deg, a small number of correct matches can be obtained by the traditional methods. With the increase of view angle, the traditional methods can obtain fewer correct matches, and they fail when the viewpoint change is substantial. However, the proposed novel method, by simulating local image regions to cope with the variation of image view, performs well. Moreover, it is still able to get a certain number of correct matches when large variation of view occurs. Although the MSER correspondences are always fewer and show a noticeable decline over 40 deg, a large number of correct matches are obtained by the proposed method benefitting from that we detect fine features in all MSERs, not in the matched MSERs. Similar conclusions are indicated in other three groups of experiments on magazine, Adam and wash. In addition, the conclusion can be drawn that whether the image is 2D scene or 3D scene, the proposed method can achieve satisfactory matching results.

2. *Comparison between the Proposed Method, ASIFT and ISIFT*

Different from the traditional methods, ASIFT, ISIFT, and the proposed method focus on the improvement of the matching framework. In this section, we compare our method with ASIFT and ISIFT based on the four datasets shown in Fig. 4. For the three approaches, the number of image pairs m on which the method still works and the average number of matches n over these m pairs are shown in Table 2 for each dataset.

Table 2 shows that the number of image pairs m in our method and ASIFT can achieve its maximum for each test. It means the proposed method and ASIFT are very robust to the variation of image view. Compared with the two methods, ISIFT is sensitive to the viewpoint change. The reason is that the number of correct matches of the basic SIFT algorithm of ISIFT is getting less and the difference of the transformations between pixels is getting larger, with the increase of view angle. The iteration algorithm of ISIFT will stop when the correct matches decrease to fewer than three, or the difference of the transformations between the reference and the test images is too large to be approximated by an affine transformation. However, profiting from the framework of local region simulation, the proposed method is still able to get a certain number of correct matches when ISIFT fails.

Table 2. Average Numbers of Matches over the Image Pairs (m/n)

Matching Methods	Graf	Magazine	Adam	Wash
ASIFT	5/1704	8/4106	4/829	1/285
ISIFT	3/808	6/714	3/166	0/0
MM-SIFT	5/923	8/1451	4/289	1/180

ASIFT obtains more matches for each test. However, these matches are calculated from much more extracted features. Indeed, ASIFT increases the number of correct matches, but they need to extract much more features from images, which cost much time in computation. Detailed analysis for the complexity of ASIFT and the proposed method is as follows:

ASIFT. In [14] a two-resolution procedure is used to accelerate ASIFT, which contains the low-resolution ASIFT and the high-resolution ASIFT. First, the complexity of the low-resolution ASIFT is proportional to the input image area. So estimating the complexity of the low-resolution ASIFT boils down to calculate the image area simulated by the low-resolution ASIFT. According to the simulating rule and parameters of ASIFT, the image area input to low-resolution ASIFT is

$$\frac{1+(|\Gamma_t|-1)180°/72°}{K\times K}=\frac{1+5\times 2.5}{3\times 3}=1.5$$

times as large as that of the original images. Thus the complexity of the low-resolution ASIFT feature computation is 1.5 times as much as that of a single SIFT routine. Low-resolution ASIFT simulates 1.5 times the area of the original images and generates

in consequence about 1.5 times more features on both the reference and the test images. Therefore, the complexity of the low-resolution ASIFT feature comparison is $1.5^2 = 2.25$ times as much as that of SIFT. The complexity of the high-resolution ASIFT depends on the number M of the identified good affine transforms and the values of tilt t in the M affine transforms. In [14], M is set as 5. If we do not consider the impact of t, the complexity of the high-resolution ASIFT feature computation and comparison are about 5 and 25 times as much as that of SIFT, respectively. Therefore, the overall complexity of ASIFT is about

$$6.5 \times \mathrm{SIFT}_{\text{feature-computation}} + 27.25 \times \mathrm{SIFT}_{\text{feature-comparison}}, \quad (12)$$

where $\mathrm{SIFT}_{\text{feature-computation}}$ is the complexity of SIFT feature computation, and $\mathrm{SIFT}_{\text{feature-comparison}}$ is the complexity of SIFT feature comparison.

Proposed method. The procedure of the proposed method mainly includes MSER extraction, SIFT feature detection, and SIFT feature matching. In MSER extraction, since the sort can be implemented as BINSORT and the list of connected components and their areas is maintained using the efficient union-find algorithm, the complexity of MSER extraction is almost linear [10]. MSER extraction is very fast in practice. Thus estimating the complexity of our method can be approximated by estimating the complexity of SIFT feature detection and matching. Assuming single-resolution MSERs is used with the ellipse-scale = 1, the maximum image area input to SIFT feature detection is about as large as the original image. Thus the complexity of the proposed method is about

$$\mathrm{SIFT}_{\text{feature-computation}} + \mathrm{SIFT}_{\text{feature-comparison}}, \quad (13)$$

where $\mathrm{SIFT}_{\text{feature-computation}}$ is the complexity of SIFT feature computation and $\mathrm{SIFT}_{\text{feature-comparison}}$ is the complexity of SIFT feature comparison. When multiresolution MSERs is used and the ellipse-scale is increased, the complexity of our method will increase. But even so, it is foreseeable that the complexity of our method is still far below ASIFT.

The average run times of ASIFT and MM-SIFT for the four experimental datasets in Fig. 4 are shown in Table 3. It indicates our method is more efficient than ASIFT. The computation times mentioned in this table have all been measured on an Intel Core 2 2.1 GHz Windows XP PC. Even though the timings are for not heavily optimized code and may change depending on the implementation as well as on the image content, we believe the table gives a reasonable indication of typical computation times.

From the analysis above, it can be seen that there is a balance between the number of correct matches and the computational efficiency in the proposed MM-SIFT method. The proposed MM-SIFT method extracts fewer correct matches than the ASIFT method. However, it is as robust as ASIFT with the increase of viewpoint angle. It can still obtain sufficient correct matches to satisfy the requirement of many applications such as registration, stitching and so on.

Table 3. Computation Times for ASIFT and MM-SIFT for the Datasets of Fig. 4 (s)

Matching Method	Average Computation Times for the Four Datasets			
	Graf	Magazine	Adam	Wash
ASIFT	85.0	34.3	69.0	43.0
MM-SIFT	51.6	9.2	25.5	14.9

Objectively, if time efficiency is not important in application, both the ASIFT and MM-SIFT can be selected. However, if time efficiency and robustness are both non-negligible factors in application, the proposed MM-SIFT method is the better choice.

5. Conclusions and Future Work

In this paper, a novel matching framework for different view angle images is proposed based on the analysis of the geometric transformation between the reference image and the test image. The multiresolution MSERs are extracted from the input images and transformed to circular areas according to the second-order moment. Then DoG detector and SIFT descriptor are combined to compute and compare features in the transformed regions. After initial matching, epipolar constraint is used to reject the incorrect matches from the initial correspondences based on the fundamental matrix. The experimental results demonstrate that the proposed method is very robust for the variation of image view in 2D scene and 3D scene. Although ASIFT is able to obtain more correct matches, many simulations and exhaustive matching strategy lead to the computing time increase greatly. Compared with that, the proposed framework simulates local areas to balance the computation efficiency and the matching result. It can obtain satisfactory matching results and ensure high time efficiency at the same time. We have to point out that the number of regions extracted by MSER detector strongly depends on the image content and the proposed method works best for structured images that can be segmented well. A possible future work is to improve the matching performance for the textured scene for large viewpoint angle.

Our research was supported by the National Basic Research Program of China (973 Program, No. 2010CB731800), the National Natural Science Foundation program (Nos. 61172174 and 10978003), the Fundamental Research Funds for the Central Universities (No. 201161902020014) and the National Science & Technology Specific Projects (Nos. 2012YQ16018505 and 2013BAH42F03).

References

1. S. N. Sinha, J-M. Frahm, M. Pollefeys, and Y. Genc, "Feature tracking and matching in video using programmable graphics hardware," Mach. Vis. Appl. **22**, 207–217 (2011).

2. M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," Int. J. Comput. Vis. **74**, 59–73 (2007).
3. B. E. Kratochvil, L. X. Dong, L. Zhang, and B. J. Nelson, "Image-based 3D reconstruction using helical nanobelts for localized rotations," J. Microsc. **237**, 122–135 (2010).
4. C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference* (Plessey, 1988), pp. 147–152.
5. K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," Int. J. Comput. Vis. **60**, 63–86 (2004).
6. S. Smith and J. Brady, "Susan: a new approach to low-level image-processing," Int. J. Comput. Vis. **23**, 45–78 (1997).
7. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis. **60**, 91–110 (2004).
8. H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," Comput. Vis. Image Underst. **110**, 346–359 (2008).
9. K. Mikolajczyk, "Interest point detection invariant to affine transformations," Ph.D dissertation (Institut National Polytechnique de Grenoble, 2002).
10. J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," Image Vis. Comput. **22**, 761–767 (2004).
11. T. Tuytelaars and L. V. Gool, "Matching widely separated views based on affine invariant regions," Int. J. Comput. Vis. **59**, 61–85 (2004).
12. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2005), pp. 886–893.
13. A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Proceedings of IEEE International Conference on Image Processing* (IEEE, 2003), pp. III-513-16.
14. J. M. Morel and G. Yu, "ASIFT: a new framework for fully affine invariant image comparison," SIAM J. Imaging Sci. **2**, 1–31 (2009).
15. G. Yu and J. M. Morel, "A fully affine invariant image comparison method," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, 2009), pp. 1597–1600.
16. Y. Yu, K. Huang, W. Chen, and T. Tan, "A novel algorithm for view and illumination invariant image matching," IEEE Trans. Image Processing **21**, 229–240 (2012).
17. A. Baumberg, "Reliable feature matching across widely separated views," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2000), pp. 774–781.
18. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," Int. J. Comput. Vis. **65**, 43–72 (2005).
19. P.-E. Forssen and D. G. Lowe, "Shape descriptors for maximally stable extremal regions," in *Proceedings of IEEE International Conference on Computer Vision* (IEEE, 2007), pp. 1–8.
20. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision* (Cambridge University, 2000).
21. http://www.robots.ox.ac.uk/~vgg/research/affine.
22. http://www.ipol.im/pub/algo/my_affine_sift/.
23. http://cmp.felk.cvut.cz/~wbsdemo/demo/.