http://dx.doi.org/10.5392/IJoC.2013.9.2.001

Improvement of ASIFT for Object Matching Based on Optimized Random Sampling

Dung Phan, Soo Hyung Kim, In Seop Na*
School of Electronics and Computer Engineering
Chonnam National University, Gwangju, 500-757, Korea

ABSTRACT

This paper proposes an efficient matching algorithm based on ASIFT (Affine Scale-Invariant Feature Transform) which is fully invariant to affine transformation. In our approach, we proposed a method of reducing similar measure matching cost and the number of outliers. First, we combined the Manhattan and Chessboard metrics replacing the Euclidean metric by a linear combination for measuring the similarity of keypoints. These two metrics are simple but really efficient. Using our method the computation time for matching step was saved and also the number of correct matches was increased. By applying an Optimized Random Sampling Algorithm (ORSA), we can remove most of the outlier matches to make the result meaningful. This method was experimented on various combinations of affine transform. The experimental result shows that our method is superior to SIFT and ASIFT.

Key words: *SIFT, ASIFT, affine invariant, similarity measurement, feature matching, outlier remover, random sampling, ORSA.*

1. INTRODUCTION

Nowadays, image matching becomes an important issue for a large number of computer vision applications: object detection [7] and recognition [8], image categorization, image retrieval [9], image classification [10], image stitching [11], stereo vision [12], [13], 3D modeling [15], etc. One of the popular approaches to this problem consists in using local features around interest points or regions. The local feature should be invariant or robust to various geometrical and scale changes. In the state of the art, the Scale Invariant Feature Transform (SIFT) feature has been proven to be one of the most robust and invariant representation methods. The SIFT method, was proposed by David Lowe [2], is a combination of the Different of Gaussian region detector that is rotation, translation and scale invariant with a descriptor based on the gradient orientation distribution in the region, which is partially illumination and viewpoint invariant. However, the SIFT detector normalizes rotations, translations and simulates all zooms out of query images. So, it is only fully scale invariant method [3]. After that, an affine-invariant SIFT (ASIFT) [1] is proposed being more robust than SIFT. The ASIFT simulates all image views obtainable by varying the two camera axis orientation parameters, namely, the *latitude* and the *longitude* angles, left over by the SIFT method. There is a mathematical proof in [1] to prove that ASIFT is fully invariant on affine transformation. Moreover, by applying a two-resolutions of implementation on matching step, ASIFT computation is accelerated and complexity of this algorithm is about twice of the complexity of a single SIFT[1]. However, we still face with a problem about time complexity, especially on matching step, because of the high dimensional feature vector of SIFT and ASIFT, 128-dimensional vector.

In order to reduce the complexity time for ASIFT algorithm, we used the simple metrics to measure similarity of features and then we applied an optimized random sampling algorithm to remove outlier matches. That makes the matching step more meaningful. In this approach, we combined the Manhattan and Chessboard metric on a linear combination for measuring distance between two features. These metrics are simpler than Euclidean metric, the original metric used in ASIFT algorithm. Therefore, computation time is reduced.

Section 2 introduces two robust local features, SIFT and ASIFT. The efficient matching is explained in section 3. Section 4 explains experimental result that shows our result is the best one and time processing is competitive compare with the ASIFT and SIFT methods.

2. RELATED WORKS

2.1. Scale-invariant feature transform (SIFT)

The SIFT algorithm is an approach for extracting distinctive invariant feature from two images. It is a combination of two

* *Corresponding author, Email: ypencil@hanmail.net*
Manuscript received Feb. 20, 2013; revised Jun 05, 2013; accepted Jun 15, 2013

stages: SIFT detector and SIFT descriptor. The SIFT detector is invariant to rotation, translation and scale [4]. A descriptor based on histogram of gradient is used to build the SIFT descriptor.

The SIFT algorithm applies four steps for two stages: SIFT detector and SIFT descriptor.

2.1.1. SIFT Detector

2.1.1.1. Scale-space extrema detection: In the first stage, interest points called keypoints are identified in the scale-space using a cascade filtering approach. The scale of an image is defined as a function $L(x, y, \sigma)$, that is produced from the convolution of a variable scale Gaussian, $G(x, y, \sigma)$, with the input image, $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

where * is the convolution operation in x and y

and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (2)$$

where σ indicates the standard deviation of the Gaussian function, $G(x, y, \sigma)$.

Using scale-space extrema in the different of Gaussian function convolved with the image $D(x, y, \sigma)$, which can be computed from the difference of two nearby scales separated by a constant multiplicative factor k:

$$\begin{aligned} D(x, y, \sigma) &= \big(G(x, y, k\sigma) - G(x, y, \sigma)\big) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \quad (3)$$

There are a number of reasons for choosing this function. First, it is a particularly efficient function to compute, as smoothed images, L, need to be computed in any case for scale space feature description, and D can therefore be computed by simple image subtraction. Figure 1 shows an illustration for DoG implementation.

In order to detect the local maxima of $D(x, y, \sigma)$, each sample point is compared to 8 neighbors in current image and nine neighbors in the scale above and below, see figure 2. If it is maximum value, it is a keypoint candidate.

2.1.1.1. Keypoint localization: Once a keypoint candidate has been found, the next step is to perform a detailed fit to nearby data for location, scale and ratio of principal curvature. This information allows points to reject that have low contrast or are poorly localized along an edge. Two criteria are used for detection of unreliable keypoints. The first criterion evaluates value of $D(x, y, \sigma)$ at each candidate keypoint. If the value is below a threshold, which means that the structure has low contrast, the keypoint is removed. The second criterion evaluates the ratio of principal curvature along it. Hence, to remove unstable edge keypoints based on the second criterion, the Difference of Gaussian (DoG) region detector which is ratio of principal curvatures of each candidate keypoint is checked. If the ratio is below a threshold, the keypoint is kept, otherwise is removed.

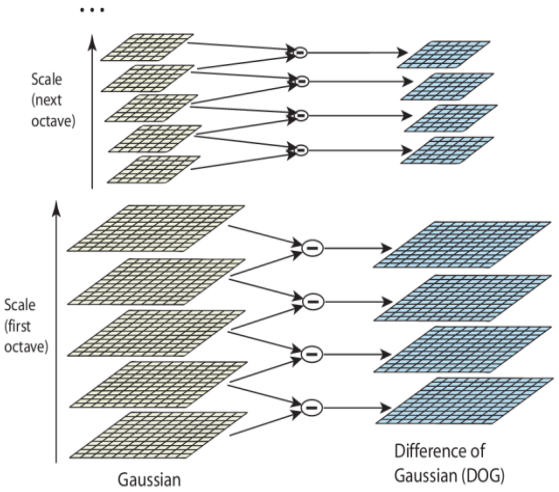


Fig. 1. Illustration for different of Gaussian implementation in multi-scale [3]

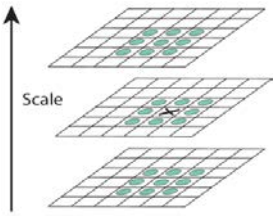


Fig. 2. Maxima of the difference of Gaussian images are detected by comparing a pixel to its 26 neighbors in each 3x3 region [3]

2.1.2. SIFT Descriptor

2.1.2.1. Orientation assignment: An orientation is assigned to each keypoint by building a histogram of gradient orientations $\theta(x, y)$ weighted by the gradient magnitudes $m(x, y)$ from the key-point's neighborhood:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (4)$$

$$\theta(x, y) = \tanh((L(x, y+1) - L(x, y-1))/(L(x+1, y) - L(x-1, y)) \quad (5)$$

where L is a Gaussian smoothed image with a closest scale to that of a keypoint. By assigning a consistent orientation and therefore invariance to image rotation is achieved.

2.1.2.2. Keypoint description: Local image gradients are measured at selected scale in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

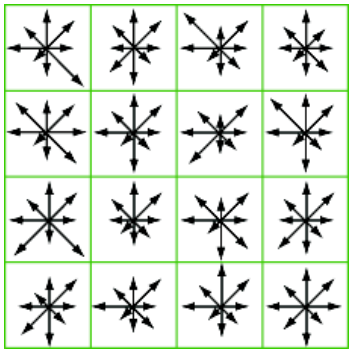


Fig. 3. 128-dimensional SIFT feature vector

In order to achieve orientation invariance, coordinate of the descriptor and the gradient orientations are rotated relative to keypoint orientation. Select a 16x16 rectangular block pixels around a keypoint and divided it into 16 4x4 sub-regions. The gradient magnitude and orientation of each 4x4 sub-region pixels are computed using pixel differences. These samples are accumulated into orientation histogram summarizing contents over 4x4 sub-regions pixels. These are weighted by a Gaussian window, indicated by the overlaid circle. Each orientation histogram is distributed into eight direction projection, length of each angle histogram projection is figure with arrow corresponding to the sum of the gradient magnitudes near that direction within the region. The keypoint is expressed by 128 dimensional feature vectors, see figure 3. Finally, the feature vector is normalized to unit length to reduce the effect of illumination change.

2.2. Affine Scale invariant feature transformation (ASIFT)

ASIFT aims to generate representation of image which is fully invariant to affine transformation. Main idea of this algorithm is combination of simulating all image views and normalizing translation and rotation. Figure 4 describes in short ASIFT simulating and matching.

2.2.1. Affine camera model: Image distortions arising from viewpoint changes can be locally modeled by affine planar transforms. Thus, image deformation model under a camera motion is $u(x,y) \rightarrow u(ax+by+e, cx+dy+f)$, where $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is any linear planar map with positive determinant.

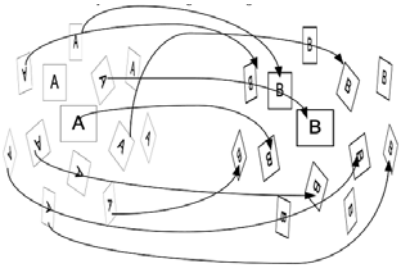
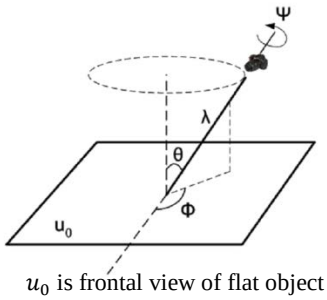


Fig. 5. ASIFT - Simulate latitude and longitude from compared images (A,B images).Then apply SIFT [1]

Any such map has decomposition as follow:

$$A = \lambda \begin{bmatrix} cos\psi & -sin\psi \\ sin\psi & cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} cos\phi & -sin\phi \\ sin\phi & cos\phi \end{bmatrix} \quad (6)$$

where $A = \lambda R(\psi) T_t R(\phi)$,
where $\lambda > 0$, λt is determinant of A, $\phi \in [0, \pi)$, $R(\psi)$ denotes the planar rotation with angle ψ and $T_t (t \geq 1)$ is called the *tilt*.



u_0 is frontal view of flat object

Fig. 6. Geometric Interpretation of Eq. (6)

Figure 5 shows the geometric interpretation of Eq. (6): ϕ *and* $\theta = arccos\left(\frac{1}{t}\right)$ are camera motion viewpoint angles and ψ parameterizes camera spin. In this affine model the camera stands far away from a planar object. Starting from a frontal position, a camera motion parallel to object's plan draws an image translation. The plane containing the normal and the optical axis makes an angle ϕ with a fixed vertical plane. This angle is called *longitude.* Its optical axis then makes a θ angle with the normal to the image plane u. This parameter is called *latitude*. The tilt $t \geq 1$ is defined by $tcos\theta = 1$. The camera can rotate around its optical axis with rotation ψ. Last but not least, the camera can move forward or backward, as measured by the zoom parameterλ. In short, (6) models the image deformation $u(x,y) \rightarrow u(A(x,y))$ induced by a camera motion from a frontal view $\lambda_0 = 1, t_0 = 0, \phi_0 = \psi_0 = 0$ to be an slanted view by λ, t, ϕ and ψ.

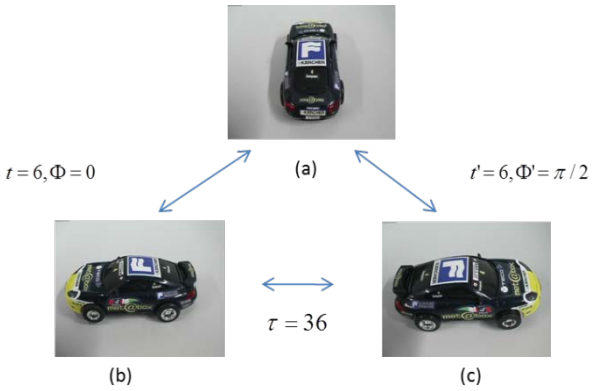


Fig. 4. (a) The frontal image (b,c) 2 slanted views. The absolute tilt is about 6 in each view. The transition tilt from left to right is actually 36

2.2.2. ASIFT algorithm: ASIFT simulates three parameters: scale, camera longitude and latitude angle which related to tilt and normalizes the rest of affine parameters. Simulating whole affine space is not prohibitive at all with

proposed affine space sampling. A two-resolution scheme will further reduce ASIFT complexity to about twice of SIFT.

ASIFT proceeds by the following steps:

- Step1: This step simulates all possible affine distortion of compared image which are caused by the change of camera optical axis orientation from a frontal view. These distortion depend on two parameters: the longitude ϕ and the latitude θ. The image undergo ϕ rotations followed by tilts $= \frac{1}{|cos\theta|}$, which means the convolution by a Gaussian with standard deviation $c\sqrt{t^2-1}$. The value c=0.8 is chosen by Lowe for the SIFT method [2].
The sampling of the latitude and longitude angles is specified below:
  - The latitudes θ are sampled so that the associated tilts follow a geometric series $1, a, a^2, \dots, a^n \; with \; a > 1$. The choice $a = \sqrt{2}$ is a good compromise between accuracy and sparsity.
  - The longitude ϕ are for each tilt an arithmetic series $0, b/t, \dots, kb/t, where \; b \approx 5\pi/12$ seems a good compromise, and k is an integer such that $kb/t < \pi$.

- Step 2: These rotation and tilts are performed for a finite and small number of latitude and longitude angles, sampling steps of these parameters ensuring that the simulated images keep close to any other possible view generated by other value of ϕ and θ.
- Step 3: All simulated images are compared by SIFT matching [2].

2.2.3. Acceleration with two resolutions scheme for ASIFT: Procedure of two-resolution for ASIFT is applying ASIFT algorithm on low resolution and high resolution of compared images. As applying ASIFT on the low-resolution first, if there exists a match on this matching, the affine transforms will be selected. After that it will simulate original image with these selected affine transforms. And finally, comparing these simulated images. The steps of two-resolution method are as follow:

- Subsample the compared images u and v by a KxK factor: $u' = S_K G_K u$ and $v' = S_K G_K v$ where G_K is an antialiasing Gaussian discrete filter and S_K is the KxK subsampling operator.
- Low-resolution ASIFT: apply the ASIFT algorithm to u' and v'.
- Identify the M affine transforms which yields the largest number of matches between u' and v'.
- High-resolution ASIFT: apply the ASIT the u and v with simulating only the M affine transforms.

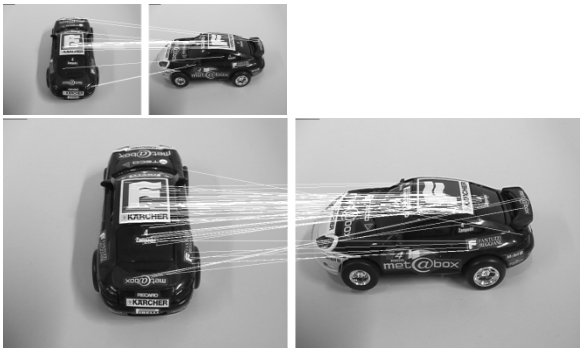


Fig. 7. Two-resolutions ASIFT illustration: Low-resolution ASIFT finds 22 matches in 6s (a), High-resolution ASFIT finds 70 matches in 37s (b)

3. EFFICENT LOCAL FEATURES MATCHING

In the original ASIFT matching [1], the nearest neighbor algorithm was used to find the best candidate match for each key-point. The nearest neighbor is defined as the keypoint with minimum Euclidean distance. As we know, the ASIFT feature vector is 128-dimensional data. So the computational cost is very high. In this paper, the Manhattan and chessboard distance are proposed to measure the similarity of features instead of the Euclidean distance to reduce the computational cost. Then part characteristics of 128-dimensional feature vector take part in the calculation gradually. By using the proposed metric, the outlier margin of matching distances are reduced. Applying the same threshold with original method we can obtain a large number of matches. Then, applying the Optimized Random Sampling algorithm (ORSA) instead of RANSAC, used on the original ASIFT, the result achieved more good matches.

3.1. Similarity Measurement

We can know that computing of Manhattan (8) and chessboard distance (9) is simpler than Euclidean distance (7), and easily we can verify expression $L_{Chess} \le L_E \le L_{Manhattan}$. Therefore, we use a linear combination of Manhattan and chessboard distance for similarity measurement of features.

$$L_E = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{7}$$

$$L_{Manhattan} = \sum_{i=1}^{n} |x_i - y_i| \tag{8}$$

$$L_{Chess} = \max_{1 \le i \le n} (|x_i - y_i|) \tag{9}$$

The similarity metric is defined as:

$$L = \alpha L_{Manhattan} + \beta L_{Chess} \tag{10}$$

Computing of the new distance was less than computing of the original. It's obvious that we can save the computation time.

3.2. Removing outlier using Optimized Random Sampling algorithm (ORSA)

The Optimized Random Sampling algorithm is able to estimate the epipolar geometry between two images even for a very large proportion of outliers. The implementation relies on a stochastic algorithm following the Random Sampling Consensus (RANSAC) [6]. In the RANSAC algorithm, a threshold has to be preset arbitrarily to decide whether a set of point matches is compatible with a fundamental matrix F.

The algorithm relies on the idea that on average the proportion of outlier should be smaller than p among the most meaningful rigid sets. This suggests that the final optimization step below should be added at the end of previous Random Sampling Algorithm.

ORSA algorithm

Set $\bar{\varepsilon} = \varepsilon_2(U)$
Repeat
  Generate a random set T of 7-point matches among U
    For each fundamental matrix F associated to T
      Compute the most meaningful rigid set $\bar{S} = \bar{S}(F)$ associated to F
      If $\varepsilon_2(\bar{S}) < \bar{\varepsilon}$ set $\bar{\varepsilon} = \varepsilon_2(\bar{S})$ and $U = \bar{S}$
End
Until the number of trial T exceed N_{opt}

In practice set $N_{opt} = \frac{N}{10}$ and apply this optimization step to the first absolute meaningful set found by random sampling.

4. EXPERIEMTAL RESULT

In order to evaluate the proposed method, we have experimented on the Morel Yu's Dataset [17] which includes many various simulations of affine transform with the absolute tilt and transition tilt. Then we compared the proposed method with the original SIFT [2], ASIFT [1].

According to the below chart [figure 7,8,9], we can see the result of SIFT at all of cases was very bad compare with ASIFT method and proposed method. Figure 7 shows that ASIFT and our method get the good result when comparing the original image to the slanted image with absolute tilt and rotation. However, our method gave the best result when rotation angle is large. In case the image transforms with the transition tilt, the proposed method shows the best result at all rotation case. The visual matching result in figure 10, 11, 12 shows the contrast of our method compare with ASIFT and SIFT.

Our method not only increases the number of matches but also eliminates false matches by using ORSA method. By applying a simpler metric, our system saved the time about 10% compare with the ASIFT.

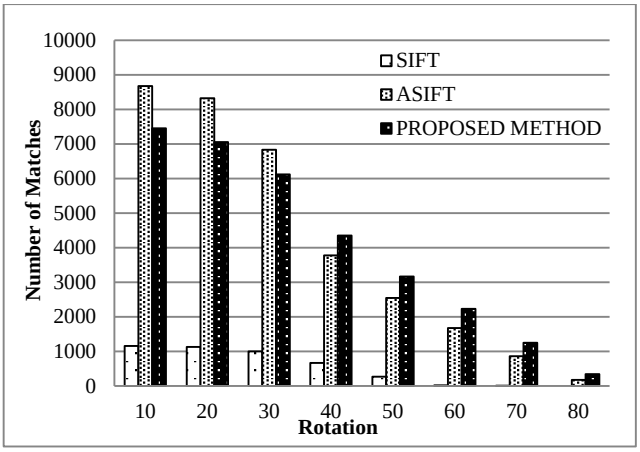


Fig. 8. Number of matches between the front image and slanted images with absolute tilt.

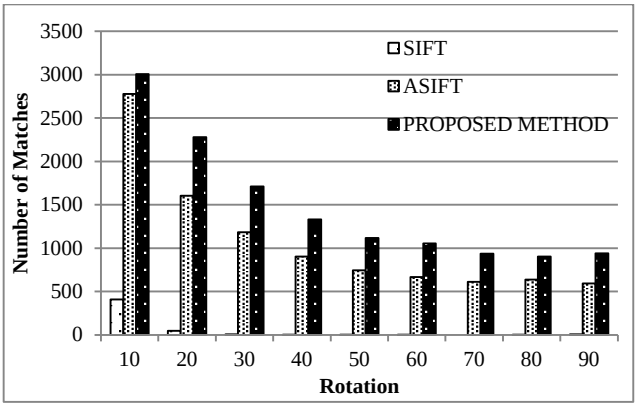


Fig. 9. Number of matches between the front image and slanted images with transition tilt (t = 2).

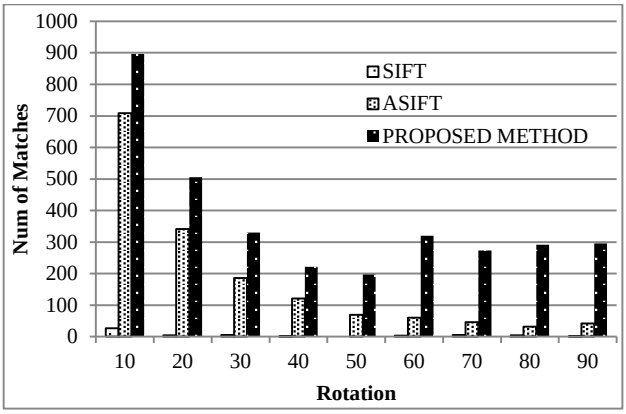
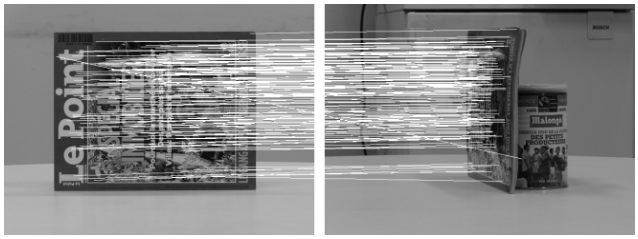


Fig. 10. Number of matches between the front image and slanted images with transition tilt (t = 4).

SIFT: 0 matches

ASIFT: 179 matches



PROPOSED METHOD: 344 matches

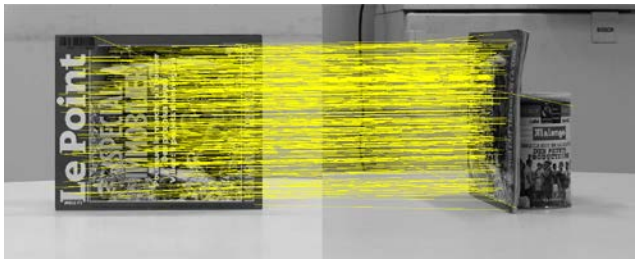


Fig. 11. Contrast of feature matching with absolute tilt (t=0) and rotation (80º) images using SIFT, ASIFT and our proposed method.

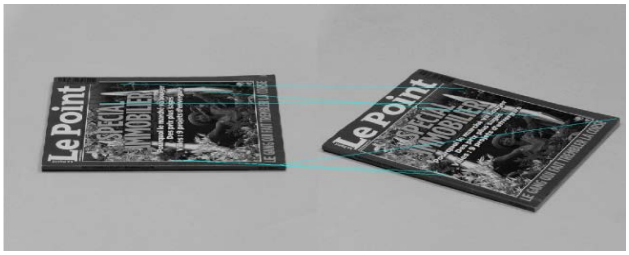
5. CONCLUSIONS

Our method got an improvement on the results compared with two well-known methods on the state-of-art of object matching SIFT and ASIFT. The proposed method got optimal number of matching and eliminated the wrong matches to make more meaningful the matching. Also the computation time is reduced by applying a linear combination of distance metrics. In the future work, we will optimize the system and also experiment with many large sets of database to meet the requirement of real-time online system.

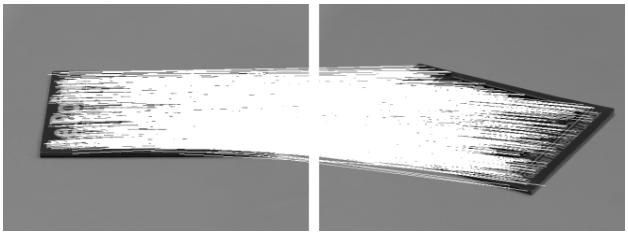
ACKNOWLEDGEMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science and Technology (2012-047759) and the MSIP (Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2013-H0301-13-3005) supervised by the NIPA(National IT Industry Promotion Agency).

SIFT: 6 matches



ASIFT: 1183 matches



PROPOSED METHOD: 1711 matches

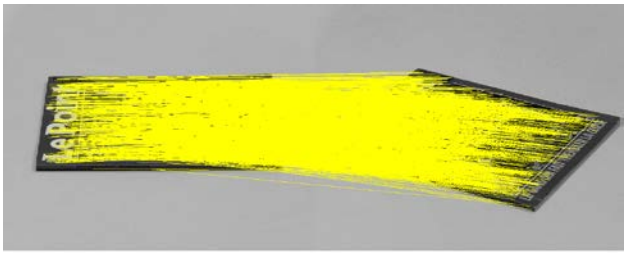
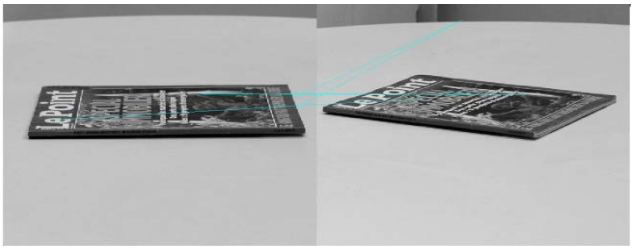
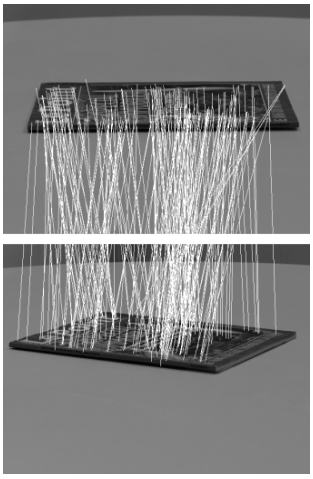


Fig. 13. Contrast of feature matching with transition tilt (t=2) and rotation (30º) images using SIFT, ASIFT and our proposed method.

SIFT: 5 matches



ASIFT: 186 matches



PROPOSED METHOD: 330 matches

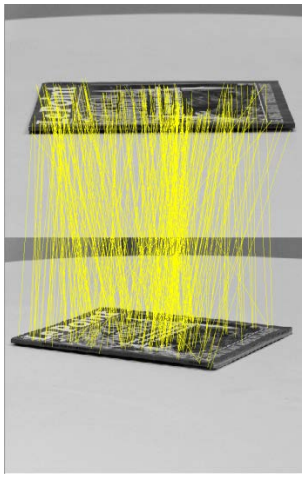


Fig. 12: Contrast of feature matching with transition tilt (t=4) and rotation (30º) images using SIFT, ASIFT and our proposed method.

REFERENCES

[1] J.-M. Morel and G. Yu, "*ASIFT: A New Framework for Fully Affine Invariant Image Comparison*", SIAM Journal on Imaging Sciences, vol. 2, 2009, pp. 438-469.

[2] D.G. Lowe, "*Distinctive image features from scale-invariant keypoints"*, International Journal of Computer Vision, 60(2): 91-110, 2004.

[3] J.M. Morel and G. Yu, *"Is SIFT scale invariant?"* Inverse Problems and Imaging, 5(1): 115-136, 2011.

[4] J. Morel and G. Yu, "*On the Consistency of the SIFT Method"*, Technical report, CMLA, ENSCachan, Cachan, France, 2008.

[5] ETHZ Toys V 1.0 Dataset: http://groups.inf.ed.ac.uk/calvin/datasets.html

[6] A. Desolneux, L. Moisan, J.M. Morel, "*From Gestalt Theory to Image Analysis: A Probabilistic Approach"*, Springer; 2008 edition.

[7] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "*Local features and kernels for classification of texture and object categories: A comprehensive study."* Int. J. Comput. Vision, vol. 73, no. 2, 2007, pp. 213–238.

[8] V. Ferrari, T. Tuytelaars, and L. Gool, "*Simultaneous object recognition and segmentation from single or multiple model views*," Int. J. Comput.Vision, vol. 67, no. 2, 2006, pp. 159–188.

[9] N. Snavely, S. M. Seitz, and R. Szeliski, "*Modeling the world from internet photo collections*", International Journal of Computer Vision, 2008.

[10] A. Bosch, A. Zisserman, and X. Munoz, "*Scene classification using a hybrid generative/discriminative approach*," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 4, 2008, pp. 712–727.

[11] M. Brown and D. G. Lowe, "*Automatic panoramic image stitching using invariant features*", International Journal of Computer Vision , vol. 74, no. 1, 2007, pp. 59–73.

[12] J. Jia and C.-K. Tang, "*Image stitching using structure deformation*", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 4, 2008, pp. 617–631.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla, "*Robust wide baseline stereo from maximally stable extremal regions*", British Machine Vision Conference, 2002, pp. 384–393.

[14] R. Deriche, Z. Zhang, Q. Luong, and O. Faugeras, "*Robust recovery of the epipolar geometry for an uncalibrated stereo rig*", Proc. European Conference on Computer Vision, 1994, pp. 567–576.

[15] A. Kushal and J. Ponce, "*Modeling 3D objects from stereo view and recognizing them in photographs*", Proc. European Conference on Computer Vision, 2006.

[16] Morel Yu's Dataset: http://www.ipol.im/pub/art/2011/my-asift/dataset_Morel_Yu_09.zip

Dung Phan

She received her B.S. degree in Computer Science of University of Science– Ho Chi Minh City, Vietnam in 2009. After that she has worked at eSilicon and Iritech Company for 2 years. Since 2012, she has been a master process student in the Department of Computer Science, Chonnam National University, Korea. Her main research interests include image processing, pattern recognition and object matching.

Soo Hyung Kim

He received his B.S. degree in Computer Engineering from Seoul National University in 1986, and his M.S. and Ph.D degrees in Computer Science from Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. From 1990 to 1996, he was a senior member of research staff in Multimedia Research Center of Samsung Electronics Co., Korea. Since 1997, he has been a professor in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, document image processing, medical image processing, and ubiquitous computing.

In Seop Na

He received his B.S., M.S. and Ph.D. degree in Computer Science from Chonnam National University, Korea in 1997, 1999 and 2008, respectively. Since 2012, he has been a contract professor in Department of Computer Science, Chonnam National University, Korea. His research interests are image processing, pattern recognition, character recognition and digital library.