IPSJ Transactions on Computer Vision and Applications Vol.5 80–84 (July 2013)

[DOI: 10.2197/ipsjtcva.5.80]

Express Paper

A Feature Descriptor by Difference of Polynomials

Bo Zheng[1,a)] Yongqi Sun[2,b)] Jun Takamatsu[3,c)] Katsushi Ikeuchi[1,d)]

Received: March 11, 2013, Accepted: April 24, 2013, Released: July 29, 2013

Abstract: In this paper, we propose a novel local image descriptor DoP which is termed as the difference of images represented by polynomials in different degrees. Once an interest point/region is extracted by a common image detector such as Harris corner, our DoP descriptor is able to characterize the interest point/region with high distinctiveness, compactness, and robustness to viewpoint change, image blur, and illumination variation. To efficiently build DoP descriptor, we propose to numerically reduce the computational cost by jumping over the repeatedly calculating polynomial representation. Our experimental results demonstrate a better performance compared to several state-of-art candidates.

Keywords: image feature, feature descriptor, polynomial modeling

1. Introduction

1.1 Motivation and Related Work

Local image feature extraction, including the techniques on detection and description, has attracted the attention of vision researchers, since it often plays an essential role in various applications such as object recognition, 3D reconstruction, image retrieval, robot localization, and video data mining.

Many modern detectors have been developed for extracting interest points or regions of an image, e.g., the popular ones being Harris corner [6], Harris-affine, Hessian, Hessian-affine [15], SIFT [11], SURF [1], MSER (Maximally stable extremal regions) [13], salient region detector [8] and Critical Nets features [5].

Descriptors are developed for the remaining question how to characterize interest points or regions distinctively and robustly. Four types of descriptors have been described in the literature [2], [9]: i) Distribution-based descriptor: spin image Johnson and Hebert [7], SIFT descriptor [11], Shape context [2]. ii) Differential-based descriptor: Gaussian derivatives and complex filters [4], steerable filters [4], SURF descriptor [1]. iii) Learning-based descriptor: one-shot approach [3], Randomized Trees [10]. iv) Others: Generalized moment invariants [18].

Polynomial representation has the potential for describing local images, which has been previously validated by Savitzky-Golay Filters [16] for 1D/2D signals. However, this technique often faces problems such as image smoothing, but has not been designed for feature description. On the other hand, beyond the explicit representation of polynomials, various properties of implicit polynomial have been explored in literals [17], [19]. The areas include fast linear fitting, few coefficients, robustness against noise, etc, but the implicit representation seems only designed for shape description.

1 IIS, U-Tokyo, Meguro, Tokyo 153–8505, Japan
2 BROTHER INDUSTRIES, LTD., Nagoya, Aichi 467–8561, Japan
3 NAIST, Ikoma, Nara 630–0192, Japan
a) zheng@cvl.iis.u-tokyo.ac.jp
b) eikisun@cvl.iis.u-tokyo.ac.jp
c) j-taka@is.naist.jp
d) ki@cvl.iis.u-tokyo.ac.jp

1.2 Contribution

In this paper, we present a novel differential based descriptor by taking advantages of the difference of polynomial representation for local images. To achieve that, we first explore the theory that explicitly represents local image patches using polynomials with various degrees. Then we propose a fast method to calculate the differences of these polynomial representations for characterizing the features of local images.

Over the state-of-the-art descriptors, the main contributions of our method are: 1) it is more robust for repetitive regions, image blur, illumination change, and view variation; 2) it is computationally efficient due to the avoidance of a strict polynomial fitting process and it is independent of the image data; 3) our feature vector is relatively in low dimension, which would be helpful for real-time applications.

2. Difference of Polynomial Representation

As shown in **Fig. 1**, our method characterizes a local interest region of an image with differences of polynomial representation (DoP) in three levels: 1) representing the region with polynomials in different degrees, 2) calculating the representation errors by subtracting the original region from the polynomial representation, and 3) successively subtracting the representation errors in a degree-increment manner.

2.1 Polynomial Representation

Given an interest region Ω detected by detectors, a point set of 2D pixels in the region: $\{\mathbf{x}_k, I(\mathbf{x}_k)\}_{k=1}^{K}, \mathbf{x}_k \in \Omega$, where $I(\mathbf{x}_k)$ is the image intensity at location $\mathbf{x}_k$. Then, a polynomial can be used to approximate this point set as:

© 2013 Information Processing Society of Japan

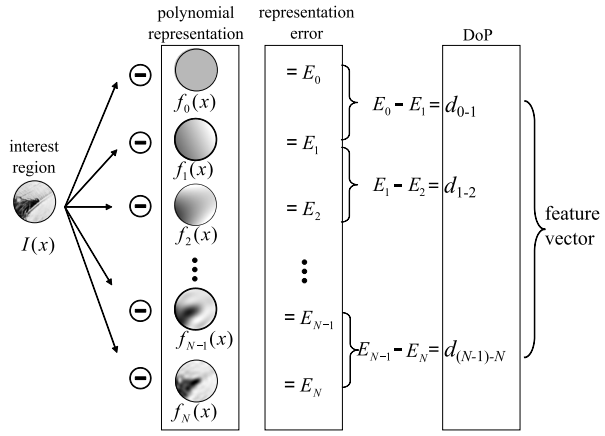

Fig. 1 Differences of polynomial (DoP).

$$f(\mathbf{x}_k) = \sum_{0\le i,j;i+j\le n} a_{ij} x_k^i y_k^j = \sum_{l=1}^{N} a_l m_l(\mathbf{x}_k) \approx I(\mathbf{x}_k), \tag{1}$$

where f is a polynomial of n degree, $\mathbf{x}_k = (x_k, y_k)^{\mathrm{T}}$ is 2D location, and $m_l(\mathbf{x}) = x^i y^j$ is called monomial function accompanying with coefficient a_l. The relationship between indices l and $\{i, j\}$ are determined by the inverse *lexicographical order*: $l = j + \frac{(i+j+1)(i+j)}{2} + 1$, and thus for an n-degree polynomial there are $n + \frac{(n+1)n}{2} + 1$ monomial terms in Eq. (1).

Then the representation accuracy can be evaluated by formulating the approximation errors with mean squares errors (MSE):

$$E = \frac{1}{K}\sum_{i=1}^{K}(I(\mathbf{x}_i) - f(\mathbf{x}_i))^2. \tag{2}$$

Figure 1 shows an example that an interest region can be approximated by several polynomials and where the operation of $\ominus$ is defined in Eq. (2).

2.2 Least-square Solution

In general, building this representation for a local region can be regarded as a linear least-square problem formulated as

$$\mathbf{a} = (M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}}\mathbf{I}, \tag{3}$$

where $\mathbf{a}$ is the unknown coefficient vector; M is a $d \times N$ matrix whose l-th column $\mathbf{m}_l$ is $(m_l(\mathbf{x}_1), m_l(\mathbf{x}_2), \ldots, m_l(\mathbf{x}_d))^T$; and $\mathbf{I}$ is a vector whose k-th entry is $I(\mathbf{x}_k)$.

After solving out coefficient vector $\mathbf{a}$, we obtain the approximated representation for a local image: $M\mathbf{a} \approx \mathbf{I}$. Thus, the representation error Eq. (2) can be rewritten as:

$$\begin{aligned} E(\mathbf{a}) &= \frac{1}{K} \parallel M\mathbf{a} - \mathbf{I} \parallel^2 \\ &= \frac{1}{K}(\mathbf{a}^{\mathrm{T}}M^{\mathrm{T}}M\mathbf{a} - 2\mathbf{a}^{\mathrm{T}}M^{\mathrm{T}}\mathbf{I} + \mathbf{I}^{\mathrm{T}}\mathbf{I}). \end{aligned} \tag{4}$$

2.3 Difference of Polynomial (DoP)

We define the Difference of polynomial (DoP) as the subtraction between the representation errors under different polynomials of different degrees, i.e.,

$$d_{s-t} = |E(\mathbf{a}_s) - E(\mathbf{a}_t)|, \tag{5}$$

where $\mathbf{a}_s$ and $\mathbf{a}_t$ are the optimized coefficient vectors of the polynomials of the s-th and t-th degrees respectively. Note, in this paper, as shown in Fig. 1, we adopt a successive way to calculate DoP, i.e., $|E(\mathbf{a}_{(n-1)}) - E(\mathbf{a}_n)|$.

3. Efficient Calculation of DoP

The problem of calculating DoP with Eq. (5) is that coefficient vectors $\mathbf{a}_s$ and $\mathbf{a}_t$ are necessary to be solved out in advance. However huge computational costs arise here, if the linear equations of Eq. (3) need to be solved many times for obtaining polynomials of different degrees. In this section, we present an efficient method which can ignore the strict calculation of polynomial coefficient vectors.

3.1 Representation Error by N-O-subspace

First before we calculate representation error using Eq. (4), we carry out QR decomposition on matrix M: $M = QR$, where Q is an orthonormal matrix and R is an upper-triangular matrix.

Suppose the i-th columns of M and Q are $\mathbf{m}_i$ and $\mathbf{q}_i$ respectively, QR decomposition can be interpreted as that the original image patch represented by polynomial subspace $\{\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_N\}$ can be transformed by a orthogonal projection to be represented by a orthonormal subspace $\{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_N\}$ termed as N-O-subspace. That is, $\mathbf{I} = M\mathbf{a} \rightarrow \mathbf{I} = Q\hat{\mathbf{a}}$, where $\hat{\mathbf{a}} = R\mathbf{a}$ and also $\hat{\mathbf{a}} = Q^{\mathrm{T}}\mathbf{I}$.

Now we can re-calculate the MSE in Eq. (4) by subsisting $M = QR$, $\hat{\mathbf{a}} = Q^{\mathrm{T}}\mathbf{I}$ and $\hat{\mathbf{a}}$ as:

$$\begin{aligned} E(\mathbf{a}) &= \frac{1}{K} \parallel M\mathbf{a} - \mathbf{I} \parallel^2 = \frac{1}{K} \parallel Q\hat{\mathbf{a}} - \mathbf{I} \parallel^2 \\ &= \frac{1}{K}(\hat{\mathbf{a}}^{\mathrm{T}} \underbrace{Q^{\mathrm{T}}Q}_{=I(\text{Identity})} \hat{\mathbf{a}} - 2\hat{\mathbf{a}}^{\mathrm{T}} \underbrace{Q^{\mathrm{T}}\mathbf{I}}_{=\hat{\mathbf{a}}} + \mathbf{I}^{\mathrm{T}}\mathbf{I}) \\ &= \frac{1}{K}(\parallel \mathbf{I} \parallel^2 - \parallel \hat{\mathbf{a}} \parallel^2). \end{aligned} \tag{6}$$

3.2 DoP by N-O-subspace

Then to calculate Eq. (5), it can be simplified to

$$d_{s-t} = |E(\mathbf{a}_s) - E(\mathbf{a}_t)| = \frac{1}{K}|(\parallel \hat{\mathbf{a}}_s \parallel^2 - \parallel \hat{\mathbf{a}}_t \parallel^2)|, \tag{7}$$

where $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_t$ are the coefficients according to s-O-subspaces and t-O-subspaces respectively. Suppose $s < t$, then s-O-subspaces is $\mathrm{span}\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_s\}$, and t-O-subspaces is $\mathrm{span}\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_s, \ldots, \mathbf{q}_t\}$. Therefore,

$$d_{s-t} = \frac{1}{K}(\sum_{i=1}^{t}(\mathbf{q}_i^{\mathrm{T}}\mathbf{I})^2 - \sum_{i=1}^{s}(\mathbf{q}_i^{\mathrm{T}}\mathbf{I})^2) = \frac{1}{K}\sum_{i=s+1}^{t}(\mathbf{q}_i^{\mathrm{T}}\mathbf{I})^2. \tag{8}$$

3.3 Feature Vector

In this paper, we choose a successive way for obtaining feature vector by DoP, that is, the feature vector $\mathbf{v}$ can be taken as the square root defined as:

$$\begin{aligned} \mathcal{F} &= (d_{0-1}^{\frac{1}{2}}, d_{1-2}^{\frac{1}{2}}, \ldots, d_{(N-1)-N}^{\frac{1}{2}})^{\mathrm{T}} \\ &= (\mathbf{q}_1^{\mathrm{T}}\mathbf{I}, \mathbf{q}_2^{\mathrm{T}}\mathbf{I}, \ldots, \mathbf{q}_N^{\mathrm{T}}\mathbf{I})^{\mathrm{T}} \end{aligned} \tag{9}$$

Thus our DoP feature vector can be simply calculated using Algorithm 1.

3.4 Fast Implementation

Given an image, there are multiple interest regions extracted.

© 2013 Information Processing Society of Japan

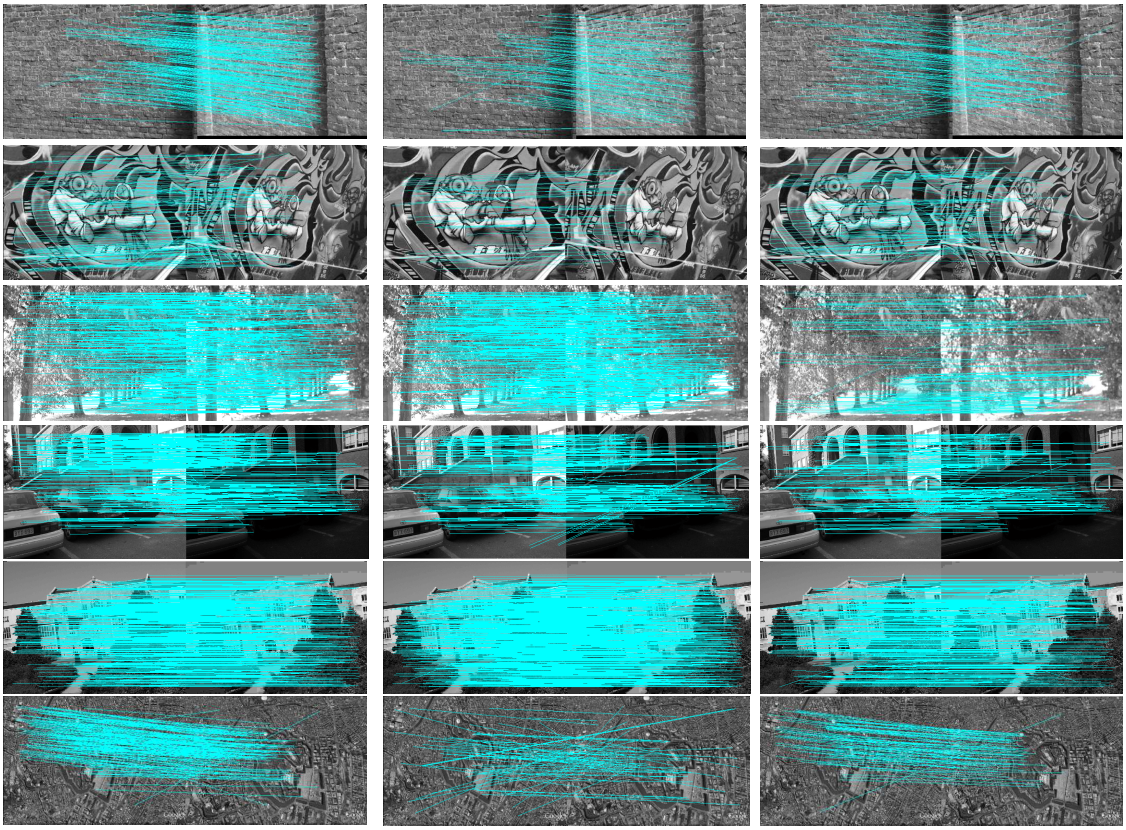

Fig. 2 Image matching results. 1st row: Wall scene with view angle changes and highly repeatable regions. 2nd row: Wall scene with view angle changes. 3rd row: trees scene with image blur. 4th row: building with illumination changes. 5th row: images with JPEG compression noise. 6th row: Google earth scene by view angle changes. Column (a) DoP, (b) SIFT, (c) SURF descriptor.

Algorithm 1: DoP feature extraction

1 **Input**: a local image patch $\{\mathbf{x}_k, I_k\}_{k=1}^{N}$
2 **Output**: feature vector $\mathcal{F}$
3 Calculate matrix M from coordinates $\{\mathbf{x}_k\}$;
4 Carry out QR decomposition on matrix M: $(Q, R) \leftarrow M$;
5 Calculate $\mathcal{F}$ with $\{I_k\}$ by Eq. (9);

Algorithm 2: Accelerated Implementation

1 **Input**: fixed-size image patches $\{\mathbf{P}_i\}_{i=1}^{N}$, each $\mathbf{P}_i = \{\mathbf{x}_k^i, I_k^i\}$
2 **Output**: feature vectors $\{\mathcal{F}_i\}_{i=1}^{N}$
3 **Initialization**: Carry out QR decomposition on matrix M, $(Q, R) \leftarrow M$, using fixed point coordinates
4 for i=1,2,...,N do
5 Calculate $\mathcal{F}_i$ with $\{I_k^i\}$ by Eq. (9)
6 end

The number are according to the feature detector and input data. However, according to Algorithm 1 (in line 4), for each region, it is required to carry out QR decomposition on each matrix M, which is too time-consuming. Fortunately, we observe that QR decomposition is independent to image information, since it is only related to coordinate information of region data points and each local coordinate can be centered by the center of region. Therefore, we derive an accelerated method in Algorithm 2 (with fixed region size). Compared to Algorithm 1, Algorithm 2 reduces the computation of QR decomposition in each loop.

4. Experimental Results

We qualitatively evaluate our method in terms of image matching robustness to 1) view change, 2) image blur, 3) illumination change, and 4) noise caused by JPEG compression. All these evaluations are based on two real image datasets: 1) Oxford affine covariant regions datasets and 2) synthesized image dataset by Google earth. We evaluate the descriptors on real images with different geometric and photometric transformations and for different scene types. The pairs of images shown in **Fig. 2** are some examples of the datasets with different view angle change, blur, illumination change and noise.

In all of our experiments, without loss of generality we choose one of the simplest feature detectors, Harris Corner detector [6]. And the interest region is defined as a fixed size neighborhood of 61×61 pixels centered at a Harris corner point. For comparison, we compare our method to the descriptors employed in SIFT [12] and SURF [1] which are most commonly used and have become a standard of comparison. We select five versions of our descriptor termed 4-degree EP, 8-degree EP, 12-degree EP, 0-2 degree EP and 0-4 degree EP with the vector dimension of 80, 144, 208, 96 and 240 respectively (compared to SIFT: 128 and SURF: 64).

In our experiments, we use nearest neighbor distance ratio matching [14] and our criterion is based on *recall vs.* (*1-precision*) evaluation. Recall is defined as:

$$\text{recall} = \frac{\#\text{correct matches}}{\#\text{correspondences}}$$

Then the number of false matches relative to the total number of matches is represented by 1-precision:

© 2013 Information Processing Society of Japan

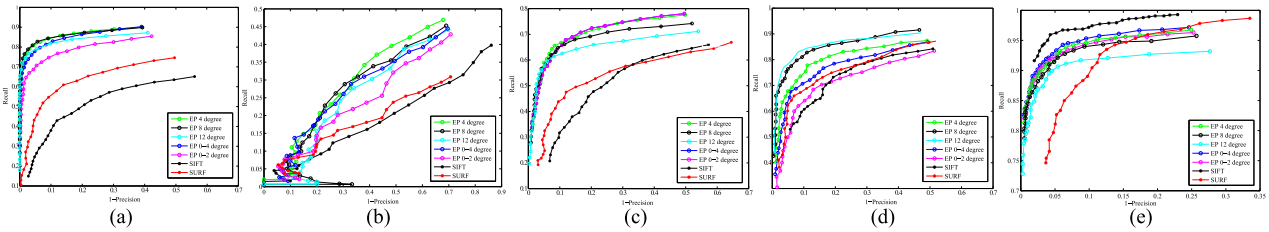

Fig. 3 Graph of recall vs. 1−precision: (a)-(e) are corresponding to the image pairs in the 1st-5th rows of Fig. 2.

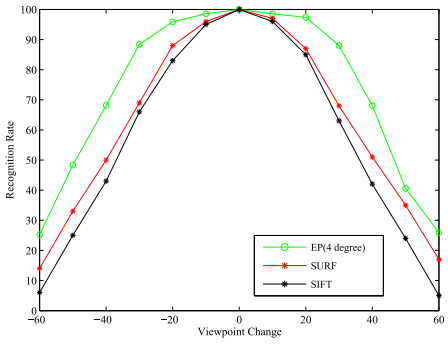

Fig. 4 Evaluation using *Google Earth* sequence (example shown in bottom of Fig. 2) with viewpoint from −60 degree to 60 degree.

$$1 - \text{precision} = \frac{\#\text{false matches}}{\#\text{correct matches} + \#\text{false matches}}$$

Also we design *recognition rate vs. viewpoint* graph on Google earth image dataset by let

$$\text{recognition rate} = \text{recall}\|_{T=1}.$$

We evaluate our method based on the following aspects:

Affine Transformation The top two rows of Fig. 2 and **Fig. 3** (a) and (b) show the result on images with view angle changes from 30 to 50 degree. The top row of Fig. 2 is one of the scene in "Wall sequence" with repetitive textures; and the second row is a structured scene selected from "Graf sequence." In each of them, the left image is with viewpoint of 30 degree, and the right is 50 degree. The results in Fig. 3 (a) and (b) show that 8-degree and 4-degree descriptors perform better than others.

And the results of another evaluation, comparing the recognition rate on images with different viewpoint on Google Earth images, is showed in the bottom row of Fig. 2. The view angles of Google Earth images vary from −60 to 60 degree. In this case, we employed 4-degree descriptor to compare with SIFT and SURF. As shown in **Fig. 4**, our descriptor performs better than others and SURF shows a little better than SIFT.

Image Blur In this experiment, we test the performance on image blur which is caused by variation of camera focus. The third row of Fig. 2 shows the example selected from "tree" sequence, and Fig. 3 (c) shows the corresponding recall vs. 1-precision graph. It shows the result that all the versions of DoP descriptors show much better performance than SIFT and SURF.

Illumination Changes Illumination change is a common image degradation, which can be introduced by variation on weather, light source or camera shutter. The forth row in Fig. 2 shows the results for some images taken with different camera settings. In Fig. 3 (d), 12-degree and 8-degree descriptor shows the best performance. This implies that illumination change might just cause the signals with low frequency, but which might not be sensitive to high-degree polynomial descriptors.

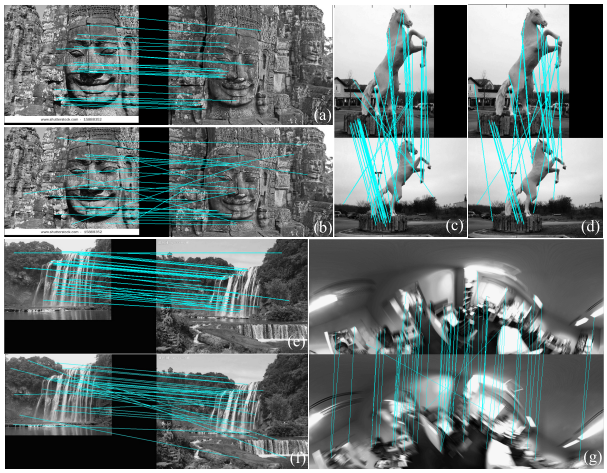

Fig. 5 Other matching results. DoP result: (a), (c), (e) and (g). SURF result: (b), (d) and (f).

JPEG Compression We evaluate the influence of JPEG compression by a building sequence, as shown in the fifth row of Fig. 2 and Fig. 3 (e). In this case SIFT performs best, and our descriptor of 0-4 degree performs as second best, and it seems that the higher degree, the lower the performance is. The reason might be that the most influenced by JPEG compression are some high frequency signals which are sensitive to high degree descriptors.

Other comparable examples We show some other comparable results in **Fig. 5**, in which image data are collected from Internet and omni-camera. Our method shows better performance than SURF descriptor in these cases.

5. Discussion

We have presented a novel local image feature descriptor using DoP which shows better performance in various image matching cases except JPEG compression. Under Harris corner detection, our descriptor seems more robust against low-frequency signal variation such as the viewpoint, blur and illumination changes, compared to SIFT and SURF. In future direction, we will evaluate the performance under different detectors, e.g., DoG, MSER, etc. Our method also shows the potential capability for applications such as image retrieval, 3D reconstruction, or panoramic image matching as shown in Fig. 5. The convenience of GPU implementation should be also attractive in our future direction.

Acknowledgments This work is supported by Next-generation Energies for Tohoku Recovery (NET), MEXT,

© 2013 Information Processing Society of Japan

and Strategic Information and Communications R&D Promotion Program (SCOPE), Ministry of Internal Affairs and Communications, Japan.

References

[1] Bay, H., Ess, A., Tuytelaars, T. and Gool, L.V.: Speeded Up Robust Features, *ECCV*, pp.346–359 (2006).
[2] Belongie, S., Malik, J. and Puzicha., J.: Shape Matching and Object Recognition Using Shape Contexts, *IEEE TPAMI*, Vol.24, pp.509–522 (2001).
[3] Fei-Fei, L., Fergus, R. and Perona, P.: One-Shot Learning of Object Categories, *IEEE TPAMI*, Vol.28, pp.594–611 (2006).
[4] Freeman, W. and Adelson, E.: The Design and Use of Steerable Filters, *IEEE TPAMI*, Vol.13, No.9, pp.891–906 (1991).
[5] Gu, S., Zheng, Y. and Tomasi, C.: Critical Nets and Beta-Stable Features for Image Matching, *ECCV*, pp.663–676 (2010).
[6] Harris, C. and Stephens, M.: A combined corner and edge detector, *Proc. Alvey Vision Conference*, pp.147–151 (1988).
[7] Johnson, A.E.: Spin-images: A representation for 3-d surface matching, Technical report, Carnegie Mellon University (1997).
[8] Kadir, T., Zisserman, A. and Brady, M.: An Affine Invariant Salient Region Detector, *ECCV* (2004).
[9] Lazebnik, S., Schmid, C. and Ponce, J.: Sparse Texture Representation Using Affine-Invariant Neighborhoods, *CVPR*, pp.319–324 (2003).
[10] Lepetit, V. and Fua, P.: Keypoint Recognition Using Randomized Trees, *IEEE TPAMI*, Vol.28, pp.1465–1479 (2006).
[11] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints, *IJCV*, Vol.60, No.2, pp.91–110 (2004).
[12] Lowe, D.: Object recognition from local scale-invariant features, *ICCV*, pp.1150–1157 (1999).
[13] Matas, J., Chum, O., Martin, U. and Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions, *BMVC* (2002).
[14] Mikolajczyk, K. and Schmid, C.: A Performance Evaluation of Local Descriptors, *IEEE TPAMI*, pp.1615–1630 (2005).
[15] Mikolajczyk, K. and Schmid, C.: Scale & Affine Invariant Interest Point Detectors, *IJCV*, Vol.60, No.1, pp.63–86 (2004).
[16] Savitzky, A. and Golay, M.: Smoothing and differentiation of data by simplied least squares procedure, *Anal. Chem.*, Vol.36, pp.1627–1639 (1964).
[17] Taubin, G.: Estimation of Planar Curves, Surfaces and Nonplanar Space Curves Defined by Implicit Equations with Applications to Edge and Range Image Segmentation, *IEEE TPAMI*, Vol.13, No.11, pp.1115–1138 (1991).
[18] Van Gool, L., Moons, T. and Ungureanu, D.: Affine/Photometric Invariants for Planar Intensity Patterns, *ECCV* (1996).
[19] Zheng, B., Takamatsu, J. and Ikeuchi, K.: An Adaptive and Stable Method for Fitting Implicit Polynomial Curves and Surfaces, *IEEE TPAMI*, Vol.32, No.3, pp.561–568 (2010).

(Communicated by *Hiroshi Ishikawa*)

© 2013 Information Processing Society of Japan