Keypoints of Interest Based on Spatio-temporal Feature and MRF for Cloud Recognition System

Takahiro Suzuki and Takeshi Ikenaga
Waseda University, Tokyo, Japan
E-mail: takahir0@toki.waseda.jp Tel/Fax: +81-3-5286-2350

***Abstract*—Keypoint extraction has lately attracted attention in computer vision. Particularly, Scale-Invariant Feature Transform (SIFT) is one of them and invariant for scale, rotation and illumination change. In addition, the recent advance of machine learning becomes possible to recognize a lot of objects by learning large amount of keypoints. Recently, cloud system starts to be utilized to maintain large-scale database which includes learning keypoint. Some network devices have to access these systems and match keypoints. Kepoint extraction algorithm utilizes only spatial information. Thus, many unnecessary keypoints for recognition are detected. If only Keypoints of Interest (KOI) are extracted from input images, it achieves reduction of descriptor data and high-precision recognition. This paper proposes the keypoint selection algorithm from many keypoints including unnecessary ones based on spatio-temporal feature and Markov Random Field (MRF). This algorithm calculats weight on each keypoint using 3 kinds of features (intensity gradient, optical flow and previous state) and reduces noise by comparing with states of surrounding keypoints. The state of keypoints is connected by using the distance of keypoints. Evaluation results show that the 90% reduction of keypoints comparing conventional keypoint extraction by adding small computational complexity.**

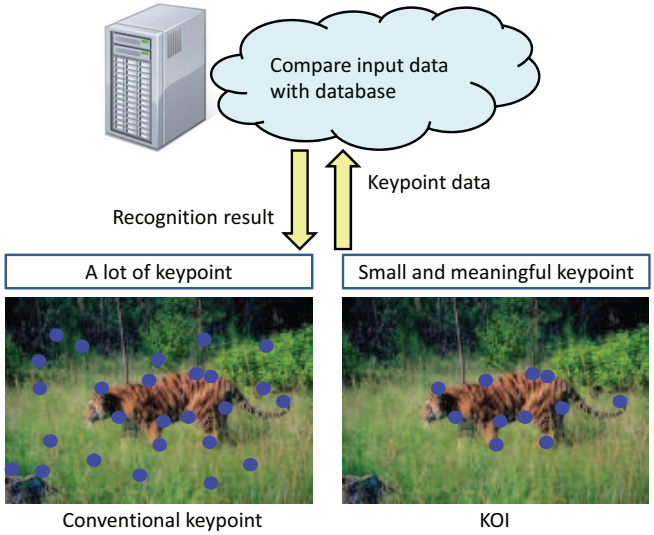

Fig. 1. Conventional keypoints and KOI.

I. INTRODUCTION

Recently, Scale-Invariant Feature Transform (SIFT) [1] has attracted attention in computer vision because of its robustness in keypoint detection. Since SIFT can describe scale, rotation and illumination invariant features from images, matching between distinct images is executed accurately. By fully utilizing this characteristics, wide range of application is being considered. For example, it is used for object recognition [2], human or other object tracking [3], [4], recognizing panorama [5], 3-D reconstruction [6]. In object recognition field, Bag-of-Features (BoF) was proposed by using combinations of SIFT descriptor. It generates one histogram from many keypoints which are extracted from one image. These are breakthrough to recognize generic objects. In addition to feature extraction, Support Vector Machine (SVM) was proposed as a machine learning algorithm. It utilizes non-linear kernel and classify obtained keypoints with a high accuracy. It needs to analyze a lot of keypoints to learn.

Recently, applications whose learned data are stored in cloud system start to be released in relation to image recognition. A lot of keypoints are extracted from input images. All obtained keypoints are communicated with database. In this case, the amount of data is bottleneck of high-speed and stable application. PCA-SIFT [7] which reduces the dimension of SIFT descriptor is also proposed. However, we need only keypoints in interest object parts. We call them Keypoints of Interest (KOI). If only KOI are extracted from input images, it achieves reduction of descriptor data communicated with database and high-precision recognition. The figure 1 shows the concept of the background.

This paper proposes the keypoint selection algorithm from many keypoints including unnecessary ones based on spatio-temporal feature and Markov Random Field (MRF). Candidates of KOI are selected by 3 kinds of features (intensity gradient, optical flow and previous state). However, only these feature includes a lot of noise. For example, it misses matching between a keypoint on the current frame and a keypoint on the previous frame. Thus, we propose noise reduction by using MRF that connects adjacent keypoints. This algorithm extracts KOI that have many features including motion and intensity gradient. That part is object that we want to recognize.

II. KEYPOINT EXTRACTION

SIFT is an algorithm which describes scale, rotation and illumination invariant keypoints from images. The algorithm is divided into following two key parts.

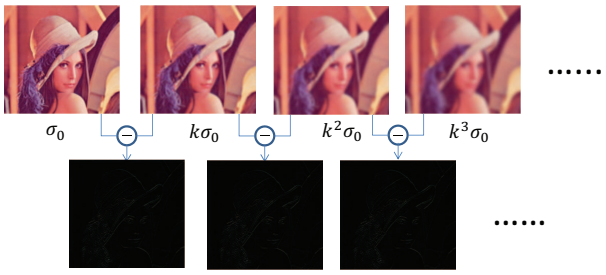

Fig. 2. The DoG detector.

- Keypoint detection
- SIFT descriptor computation

The keypoint detection is the process which decides keypoint's position near characterized region. The SIFT descriptor computation makes the histograms with information about neighboring region. These are primary processes of a keypoint extraction. However, the keypoint detection part of SIFT requires high computational complexity. Thus, this paper utilizes low complexity keypoint extraction based on corner detection and plural images in database [8]. The method performs real-time keypoint extraction maintaining almost same accuracy with SIFT. The keypoint selection process is added to this keypoint extraction.

A. Keypoint Detection

SIFT detects scale invariant keypoints by the DoG function. DoG function computes difference of images convolved by Gaussian filters. An image, $I(x, y)$, a variable-scale Gauss function, $G(x, y, \sigma)$, and a smoothed images, $L(x, y, \sigma)$, define the DoG image, $D(x, y, \sigma)$:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma), \end{aligned} \quad (1)$$

where $*$ is the convolution operation. $D(x, y, \sigma)$ is repeatedly computed by a constant multiplicative factor k. Computational complexity becomes higher and higher when σ increases. Fig.2 is schema of the DoG detector. Thus, this process is very complex. After this, detections of extreme value and localizations of keypoints are performed. It also high computational complexity because localizations use matrix calculation. Thus, we utilized corner detection and integral image to detect keypoints quickly.

B. SIFT descriptor computation

In computation of SIFT descriptor, firstly, keypoint's orientation is obtained. The histogram is calculated by gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$:

$$m(x, y) = \sqrt{L_x(x, y) + L_y(x, y)}, \quad (2)$$

$$\theta(x, y) = tan^{-1} \frac{L_y(x, y)}{L_x(x, y)}. \quad (3)$$

When its sum of magnitude is max, the orientation becomes the keypoint's one. After this, SIFT descriptor is computed.

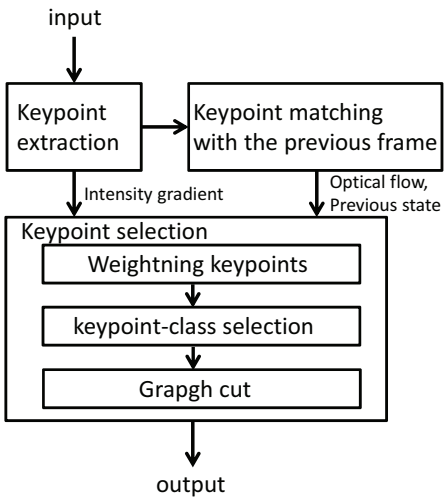

Fig. 3. The flow of entire processing.

The region is rotated by the keypoint's orientation. The size of region depends on scale obtained by the DoG detector. It is divided into 4×4 and histogram is computed by 8 directions in each region. Total 128 dimension vector, SIFT descriptor, is generated. This process's computational complexity changes depending on keypoint's scale. In this paper, the scale is fixed and utilizes plural images in database to deal with scale changes.

III. KEYPOINT SELECTION BASED ON SPATIO-TEMPORAL FEATURE AND MRF

In chapter III, we show the method that extract KOI from an input movie. The entire flow is shown in Fig. 3. This algorithm utilizes keypoint extraction and keypoints are matched with keypoints in previous frame. Using these data, keypoint selection is performed. This paper proposed three methods below.

- Weighting keypoints by three elements
- Statistical keypoint-class selection
- MRF to keypoint class

First, this algorithm weights keypoint by three elements and calculates values which describe likelihood of KOI. Then, these values are arranged and keypoint class is determined by statistical processing. However, the results include a lot of noise that is caused by miss-matching between a keypoint on the current frame and a keypoint on the previous frame. Thus, keypoints are connected by MRF and graph cut algorithm is used to reduce noise from the output keypoints. This chapter shows each algorithm in more detail.

A. Weighting keypoints by three elements

In this paper, we choose three elements for weighting keypoints. The elements are intensity gradient, optical flow and previous state. The reasons are shown next. With respect to intensity gradient, objects which include many intensity gradients stand out among others. For example, book covers, posters and traffic signs are pointed out. With respect to optical

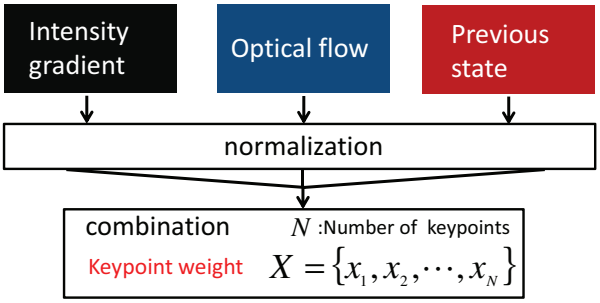

Fig. 4. The flow of weight on keypoints.

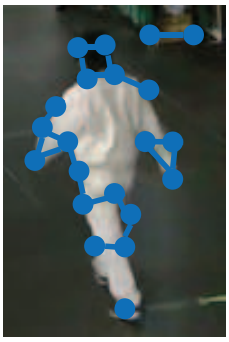

Fig. 5. The connection of keypoints.

flow, objects which move widely stand out among others. For example, human, animals and vehicles are pointed out. Finally, with respect to previous state, objects which stand out in previous frame keep standing out in current frame. These values are normalized and added together. The calculated values are keypoint weight. This flow is described in Fig.4.

The ways to obtain these values are shown. The weight of intensity gradient is calculated by Hessian detector [9]. Hessian detector computes matrix, $\mathbf{H}$, which is composed of elements are the 2nd-order difference of adjacent pixels:

$$\mathbf{H} = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}, \tag{4}$$

$$det(\mathbf{H}) - \omega tra(\mathbf{H}), \tag{5}$$

where ω is a parameter which is determined by experiments. This value of Eq.(5) describes how much the keypoint likes corner. It is obtained by the corner detection part of keypoint extraction. The weight of optical flow is calculated by norm of optical flow. The optical flow is obtained by keypoint matching of SIFT descriptor between previous frame and current frame. The weight of previous state is calculated by counting the state that is kept in plural frames. The value is decided by Gaussian function.

B. Statistical keypoint-class selection

By using values obtained by three, keypoint class is determined. In this paper, a statistical method is proposed. By Bayes' theorem, the equations:

$$P(\omega_0|x_i) = \frac{P(\omega_0)P(x_i|\omega_0)}{P(x_i)}, \tag{6}$$

$$P(\omega_1|x_i) = \frac{P(\omega_1)P(x_i|\omega_1)}{P(x_i)}, \tag{7}$$

are led. The class of not KOI is defined as ω_0. The class of KOI is defined as omega ω_1. If the equation (6) is higher than the equation (7), the keypoint is class ω_0. In other case, the keypoint is ω_1. Next, the right side of these equations are calculated. We assume that ω_0 and ω_1 can be caused equally: $P(\omega_0) = P(\omega_1) = 0.5$. In addition, we assume that posterior possibility comply with gaussian distribution:

$$P(x_i|\omega_1) = \exp\left(-\frac{1}{2}\alpha(x_{max} - x_i)^2\right). \tag{8}$$

This process generates keypoint class $Y = \{y_1, y_2, , y_N\}$ on each keypoint where $y_i \in \{0, 1\}$. x_{max} is the maximum value during normalization.

C. MRF to keypoint class

To solve noise problem, this paper also proposes MRF to keypoint class. Keypoints are connected by the distance on each other because the keypoint whose adjacent keypoints are KOI tends to be KOI. The example of connections is Fig. 5. We utilize graph cut to reduce noise and determine keypoint classes. Graph cut algorithm minimizes the energy equation:

$$E(X) = \sum_i g_i(x_i) + \sum_{i,j} h_{ij}(x_j, x_i). \tag{9}$$

In this case, global solution is calculated because the keypoint class is binary. To solve this minimization problem, Min-Cut/Max-flow algorithm is used. Each function is defined as

$$g_i(x_i) = \lambda|y_i - x_i|, \tag{10}$$

$$h_{ij}(x_j, x_i) = \kappa \exp(-(dist(i, j))^2). \tag{11}$$

The equation (10) is data term. The outputted x_i change to approximate inputted y_i. The equation (11) is smoothing term. The strengthen of connection depends on distance between kepoints. We assume it as gaussian distribution. The nearer the keypoint distance, the stronger connection this function generate. $dist(i, j)$ represents the distance between keypoint i and keypoint j. λ and κ are parameters which are determined experimentally. They determines the strengthen of data term and smoothing term. The calculated $X = \{x_1, x_2, , x_N\}$ where $x_i \in \{0, 1\}$ is the output keypoint class. If the value of x_i is 1, the keypoint i is KOI. This calculation is faster than noise reduction of image because there are fewer node in MRF of keypoints.

IV. Evaluation Results

This chapter shows the evaluation results that compare the proposed method with general keypoint extraction. The development environment on software is Visual Studio C++ 2008. CPU is Intel Core i5 CPU M 450 2.40GHz. The resolution of the video we used is HD (1280×720), 30 fps taken by fixed camera. In this sequence, the humans are

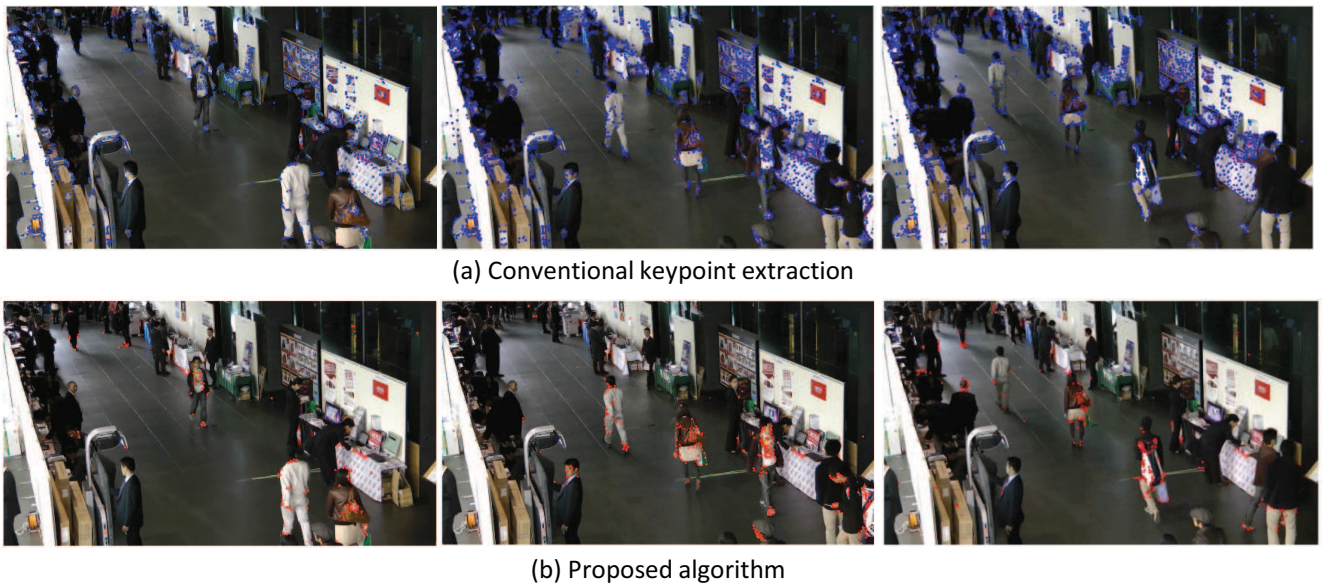

Fig. 6. The comparison between (a) conventional keypoint extraction and (b) proposed algorithm.

TABLE I
THE NUMBER OF KEYPOINT AND PROCESSING TIME BETWEEN CONVENTIONAL METHOD AND PROPOSAL.

	conventional method	proposal
number of keypoints	1962	204
processing time [ms]	431	617

walking on a path and this video is the surveillance-like scene. First, the number of keypoints which are detected by both methods are compared in Tab. I. It shows the average among all frames. The proposed algorithm achieves the 90% reduction of keypoints. In addition, processing time is compared. The proposed algorithm requires about 43% higher computational complexity than the conventional keypoint extraction.
Figure 6 shows the video result of the conventional method and proposed algorithm. The blue keypoints are conventional keypoints and the red ones are keypoints extracted by proposed algorithm. It shows the proposal detects keypoints in human which moves on the path. By using only this keypoints, it is expected to analyze human behaviors in surveillance camera combining motion features.

V. CONCLUSION

The conventional keypoint extractions utilizes only spatial information and extract a lot of unnecessary keypoints. Considering cloud applications, the reduction of data amount of keypoints is needed. This paper proposes the keypoint selection algorithm from many keypoints including unnecessary ones based on spatio-temporal feature and MRF. It calculats weight on each keypoint using 3 kinds of features (intensity gradient, optical flow and previous state) and reducing noise by comparing with states of surrounding keypoints. The evaluation result shows that the proposed algorithm achieves the 90% reduction of keypoints by only requiring low computational time. This algorithm is expected to be applied to surveillance camera and in-vehicle camera when they start to utilize cloud system. This paper utilized only fixed camera. However, it will be possible to apply it to moving camera by using global motion vector and statistical processing.

ACKNOWLEDGMENT

This work was supported by KAKENHI (23300018).

REFERENCES

[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int.Journal of Computer Vision, 60, pp. 91-110, 2004.
[2] D. G. Lowe, “ Object recognition from local scale-invariant features,” In International Conference on Computer Vision, Corfu, Greece, pp. 1150-1157, 1999.
[3] Yuji Tsuzuki, Hironobu Fujiyoshi, Takeo Kanade, “ Mean Shift-based Point Feature Tracking using SIFT,” Journal of Information Processing Society ,Vol. 49, No. SIG 6, pp. 35-45, 2008.
[4] Huiyu Zhou , Yuan Yuan , Chunmei Shi, “ Object tracking using SIFT features and mean shift,” Computer Vision and Image Understanding, v.113 n.3, pp. 345-352, 2009.
[5] Matthew Brown and David G. Lowe, “ Recognising panoramas,” International Conference on Computer Vision , pp. 1218-25, 2003.
[6] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, R. Szeliski, “ Building rome in a day,” In ICCV, 2009.
[7] Y. Ke, R. Sukthankar. PCA-SIFT, “ A More Distinctive Representation for Local Image Descriptors,” Proceedings of Computer Vision and Pattern Recognition, pp. 506-513, 2004.
[8] Takahiro Suzuki, Takeshi Ikenaga, “ SIFT-Based Low Complexity Keypoint Extraction and Its Real-Time Hardware Implementation for Full-HD Video”, APSIPA Annual Summit and Conference (ASC 2012), Dec. 2012.
[9] Beaudet, P. R., “ Rotational invariant image operators,” In Proceedings of the 4th International Joint Conference on Pattern Recognition (ICPR), pp. 579-583, 1978.