Representing Human Motions in an Eigenspace Based on Surrounding Cameras

Satoshi Houman*, M. Masudur Rahman*, Joo Kooi Tan*, Seiji Ishikawa*

Kyushu Institute of Technology, Department of Control Engineering
Sensuicho 1-1, Tobata, Kitakyushu 804-8550, Japan
(E-mail: {houman, ishikawa}@ss10.cntl.kyutech.ac.jp)

Abstract: Recognition of human motions using their 2-D images has various applications. An eigenspace method is employed in this paper for representing and recognizing human motions. An eigenspace is created from the images taken by multiple cameras that surround a human in motion. Image streams obtained from the cameras compose the same number of curved lines in the eigenspace and they are used for recognizing a human motion in a video image. Performance of the proposed technique is shown experimentally.

Keywords: Eigenspace, Human motion, Motion representation, Motion recognition

1. INTRODUCTION

Automatic human motion recognition by computer has various potential applications such as detecting a person behaving in an abnormal way in a surveillance system, discovering a person who feels bad (and has sat down on a road, for example) in order for an intelligent robot to give him/her a hand, monitoring activities of aged people at home for their safety, etc.

There have been studies on automatic human motion recognition [1-6], but none of them has not yet been put into a practical use. One of the main reasons of this is that the appearance of a human motion differs from each other according to the orientation of observation.

To overcome this difficulty, we propose for human motion representation the employment of an eigenspace based on multiple views. An eigenspace is created from a set of video image frames of a motion taken from multiple orientations by cameras. This means that the eigenspace representation is an appearance-based representation in which various shots or multiple views of a 2-D human motion are memorized. This enables automatic recognition of a human motion from an arbitrary orientation of observation.

Theoretical aspect of the proposed technique is described. Its performance is shown by an experiment employing three human motions and four cameras.

2. AN EIGENSPACE METHOD

An eigenspace method is one of the techniques which recognize a 3-D object from its 2-D image. Since it recognizes a 3-D object as an aggregate of 2-D images, it excels other 3-D object recognition techniques in the point of the amount of calculation and storage capacity.

2.1 Input Image Generation

First, an image portion is extracted from the original image which contains the interested object using the techniques such as spotting and subtraction. A noise reduction filter, a median filter, for example, is then applied to the extracted image. In order to reduce computation time, this image is normalized to a smaller size.

2.2 Image Normalization

The generated input image has variance in average brightness depending on the photographic environment. Therefore the image receives normalization with its brightness.

An input image is represented by a column vector $\hat{\boldsymbol{x}}$ of the form

$$\hat{\boldsymbol{x}} = \left(\hat{x}_1, \hat{x}_2, ..., \hat{x}_N\right)^{\mathrm{T}}, \tag{1}$$

where N is the number of pixels composing the image. Image brightness normalization is performed so that the norm of the image vector $\boldsymbol{x}$ is set to 1 as follows;

$$\boldsymbol{x} = \frac{\hat{\boldsymbol{x}}}{\left\|\hat{\boldsymbol{x}}\right\|}, \quad \left\|\hat{\boldsymbol{x}}\right\| = \sqrt{\sum\nolimits_{i=1}^{N} \hat{x}_i^2} \tag{2}$$

The normalized image vector $\boldsymbol{x}$ is expressed as follows;

$$\boldsymbol{x} = \left(x_1, x_2, ..., x_N\right)^{\mathrm{T}} \tag{3}$$

The input image defined from the rth ($r = 1,2,\cdots,R$) image frame of the hth (h=1,2,…,H) human's motion taken by the pth ($p = 1,2,\cdots,P$) camera is denoted by $\boldsymbol{x}_{r,p}^{h}$.

2.3 Karhunen-Loeve Transform

When a motion image of a person is taken, successive frames normally have high correlation among them, since a motion changes smoothly. From this fact, the motion image stream can be compressed employing Karhunen-Loeve transform. This technique compresses a large dimensional data space into a smaller space called an eigenspace defined by a set of eigenvectors obtained from a data covariance matrix. If one chooses some eigenvectors corresponding to the largest eigenvalues, the original data can be well represented in the reduced eigenspace.

In the proposed technique, an eigenspace is defined using a set of images $\boldsymbol{x}_{r,p}^{h}$ {r=1,2,…,R; p=1,2,…,P; h=1,2,…,H}. In the first place, an average image $\boldsymbol{c}$ is calculated by

$$\boldsymbol{c} = \frac{1}{RPH} \sum_{h=1}^{H} \sum_{p=1}^{P} \sum_{r=1}^{R} \boldsymbol{x}_{r,p}^{h} \tag{4}$$

An image data matrix X is then defined by

$$X = \left(\boldsymbol{x}_{1,1}^{1} - \boldsymbol{c}, \boldsymbol{x}_{2,1}^{1} - \boldsymbol{c}, ..., \boldsymbol{x}_{r,p}^{1} - \boldsymbol{c}, ..., \boldsymbol{x}_{R,P}^{H} - \boldsymbol{c}\right). \tag{5}$$

This data matrix X defines a covariance matrix $\boldsymbol{Q}$ of the form

$$Q = XX^{\mathrm{T}}. \tag{6}$$

Eigenvalue λ of a covariance matrix is obtained from the following eigen-equation;

$$Q\boldsymbol{u} = \lambda\boldsymbol{u}. \tag{7}$$

According to Karhunen-Loeve transform, the obtained eigenvalues are arranged in the descending order, and the k eigenvectors corresponding to the largest k eigenvalues are employed to define a k-dimensional subspace which is called an eigenspace. Let us put the N eigenvalues λ_k (k=1,2,…,N) in the descending order as

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq \cdots \lambda_N. \tag{8}$$

The cumulative proportion K is then defined by

$$K = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{N} \lambda_i. \tag{9}$$

The value K is used for evaluating the degree of approximation.

2.4 Image Representation in an Eigenspace

An appearance of a 3-D object normally changes continuously, if it moves smoothly or a camera position changes gradually. Therefore the 3-D object is expressed as a manifold in an eigenspace.

An image $\boldsymbol{x}_{r,p}^{h}$ that shows an appearance of a 3-D object is projected onto the k-dimensional eigenspace created by a set of eigenvectors $\boldsymbol{e}_i$ (i=1,2,…,k) by the following formula;

$$\boldsymbol{g}_{r,p}^{h} = \left(\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_k\right)^{\mathrm{T}} \left(\boldsymbol{x}_{r,p}^{h} - \boldsymbol{c}\right) \tag{10}$$

Thus an image $\boldsymbol{x}_{r,p}^{h}$ has a one to one correspondence to a point $\boldsymbol{g}_{r,p}^{h}$ in an eigenspace.

Suppose a particular motion of an object h is observed from the pth camera and the motion sequence is given by a series of image frames r (r=1,2,…,R). Then, if the images $\boldsymbol{x}_{r,p}^{h}$ (r=1,2,…,R) are projected into an eigenspace by Eq.(10), they produce R successive points in it. As a motion is normally smooth, these R points may define a smooth curved line segment L_p^h in the eigenspace. Since every camera defines the line segment L_p^h and the camera orientation changes continuously, say every 1 degree, L_p^h (p=1,2,…,P) also changes continuously. This yields a curved surface patch S^h representing the motion of an object h.

After all, a curved surface patch in an eigenspace represents a particular motion of an object and its 2-D multiple appearances, if we create the eigenspace by employing surrounding cameras that gives image streams of multiple views of the interested motion. **Figure 1** shows an example of an eigenspace representing a motion in which (a) a person stands straight first, (b) bends his waist to pick up an object on the ground, and (c,a) stands straight again. The motion is given by 12 successive image frames, which defines a closed line segments in the 3-D eigenspace that approximates a curved line segment $L_{p^*}^{h^*}$, given a particular camera orientation p* and a particular object h*.

3. MOTION REPRESENTATION AND RECOGNITION

This section describes a representing and recognizing method of a human motion by an eigenspace created from multiple views a surrounding cameras system provides.

3.1 Input Image Creation

The creation process of an input trimmed image is shown in **Fig 2**. An original image frame is given in the left of (a) and the background image in the right of (a). First, differentiation is performed to the both images in (a), which yields the differentiated images in (b). Subtraction is done between the two images, yielding the image that contains only a person as shown in (c). The employment of a differentiated image has an advantage that it can extract the interested object easily by subtraction.

Furthermore, if we use the original gray images instead of differentiated images, different clothes of a person result in the creation of different eigenspaces even if his/her motion is identical. The employment of a differentiated image is reducing this dress problem to a large extent by considering only the principal lines of a person's body.

Before subtraction, a differentiation filter is applied to all the processed image frames and the Gaussian filter is applied to the differentiated image for noise reduction. Then the image is binarized, i.e., it is transformed into a two-valued image. The outline of a person in the binarized image receives dilation and erosion to eliminate discontinuity on the outline. Finally labeling to figure regions is performed and only the region having the largest area is left on the image. In this way, a person is extracted correctly from the original image.

Normalization of the size is applied to the person-extracted image obtained from the above-mentioned processing. This contributes to making the proposed technique indifferent to human physical difference. In the performed experiment, the image size is reduced to 20×20 pixels. This is also advantageous to reducing the computation time.

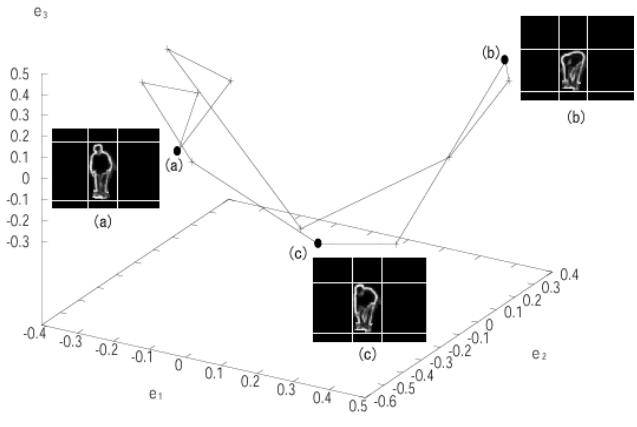

Fig. 1. An example of an eigenspace.

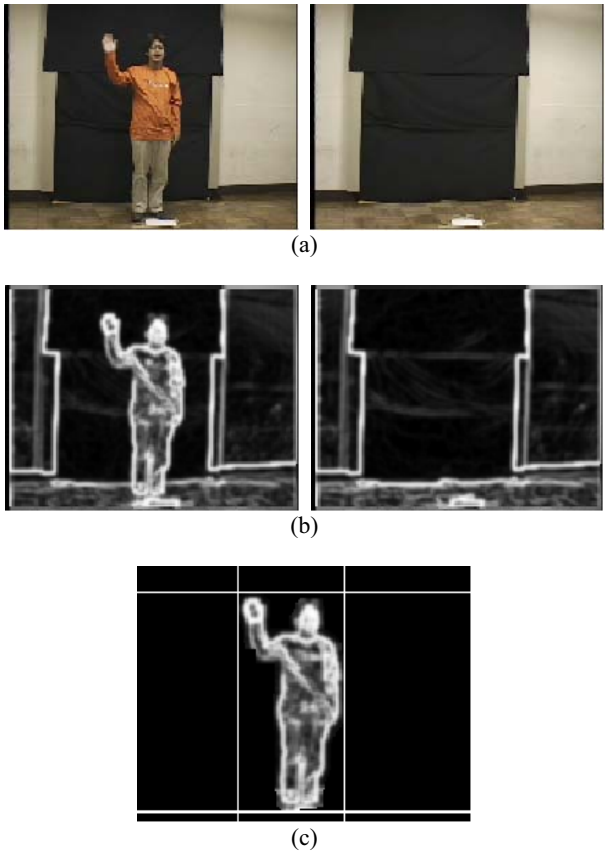

Fig 2. The input image gerneration process.

3.2 Motion Representation Method

The motion image taking system this paper proposes is a surrounding cameras system in which multiple cameras are placed around a person in motion and captures its images from every orientation. Actually the cameras provide frontal views, rear views, and side views of the motion.

Once all of these images are obtained, they are then used for defining a single eigenspace employing Eqs.(4-7). Suppose that an eigenspace has been created from the set of images. A set of image frames representing an image sequence of a motion observed from a certain camera are projected onto the eigenspace by Eq.(10) and it makes a smooth curved line, since smooth motions are considered in this study. P smooth curved lines similar to this particular line are described in the eigenspace, since there are P cameras surrounding the motion concerned. If we assume that many cameras observe the motion, 360 cameras, for example, successive cameras see very similar but slightly different motion sequences. This signifies that P smooth curved lines are continuously changing from a camera to a camera making a smooth curved surface.

Consequently a motion is represented in such a smooth curved surface patch in the eigenspace defined by surrounding cameras. It contains not only frontal views of the motion interested, but also its rear views and side views. The eigenspace representation model is one of the appearance models of a 3-D object.

3.3 Motion Recognition Method

Motion recognition of a person is realized using the proposed eigenspace representation technique. The procedure of the motion recognition is divided into two parts; a learning part and a recognition part.

In a learning part, according to the procedure stated in 2.3 and 2.4, image streams of M kinds of motions are collected to create respective eigenspaces from the image frames within the streams. Once an eigenspace is defined from k chosen eigenvectors, all the image frames are projected onto the respective eigenspace by Eq.(10), yielding M surface patches representing M respective motions.

In a recognition part, on the other hand, an image stream of an unknown motion is projected onto each of the M eigenspaces. As is explained later, the projected image stream is a set of projected points and their proximity to one of the memorized M surface patches is evaluated.

Suppose that a single camera captures an unknown motion. This yields a motion image stream containing R successive image frames, one of which is chosen and trimmed and normalized to provide an input image denoted by $\boldsymbol{y}$. Image $\boldsymbol{y}$ is projected onto a point $\boldsymbol{z}$ in every eigenspace by the following formula;

$$\boldsymbol{z} = (\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_k)^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{c}). \tag{11}$$

Here the average of learning image data is denoted by $\boldsymbol{c}$. Equations (10) and (11) are identical. The former is employed for defining a surface patch of a motion in the learning stage, whereas the latter is used for projecting an image of an unknown motion onto a point in the eigenspace.

In the eigenspace, a surface patch that contains the projected point $\boldsymbol{z}$ is searched. Instead of searching for it on a continuous surface, as it can be approximated by a digital surface composed of projected learning images $\boldsymbol{g}^h_{r,p}$, the distance between $\boldsymbol{z}$ and $\boldsymbol{g}^h_{r,p}$ is evaluated as shown in Eq.(12);

$$d_{\min} = \min_{r,p,h} \left\| \boldsymbol{z} - \boldsymbol{g}^h_{r,p} \right\| \equiv d^{h*}_{r*,p*}. \tag{12}$$

If, for a certain positive threshold ε,

$$d^{h*}_{r*,p*} < \varepsilon \tag{13}$$

holds, the image $\boldsymbol{y}$ is recognized as the r*th image of the motion of a person h* observed by camera p*.

Suppose that the observed image stream of an unknown motion m_{un} contains T successive image frames. Every frame t (t=1,2,…,T) can be recognized by the above stated procedure. Then for overall recognition, we have two rules in this particular study. Note that we have M motions as eigenspaces. Let us denote the number of image frames that have $d_{\min}$ with respect to motion m by a_m (m=1,2,…,M). Let us also denote the number of image frames that have the second minimum value denoted by $d_{\min 2}$ with respect to motion m by a_m' (m=1,2,…,M). Then we have the following rules;

$$\{a_{m*} \geq a_m (m = 1,2,..., M)\} \cap \{a_{m*} \geq T/2\} \Rightarrow m_{\mathrm{un}} = m* \tag{14a}$$

$$\{a_{m*} + a_{m*}' \geq a_m + a_m' (m = 1,2,..., M)\} \cap \{a_{m*} < T/2\} \Rightarrow m_{\mathrm{un}} = m* \tag{14b}$$

In this way, an unknown motion is recognized as a stream of image frames. In the process of an image frame recognition, the minimum and the second minimum distances are calculated by Eqs.(12),(13) and kept for overall recognition by Eq.(14).

4. EXPERIMENTAL RESULTS

4.1 Method

The experiment was conducted by arranging four digital video cameras (P=4) placed in front of a person in motion. Every adjacent view angles between the cameras make 30° degrees. Thus these 4 cameras provide different frontal appearances of the captured motion. Three motions (M=3), i.e., Motion_1: shaking the right hand (abbr. ShakeHand), Motion_2: picking up from a floor (abbr. PickUp), and Motion_3: stepping (abbr. Step), are acted by six (H=6) male students of early twenties.

4.2 Results

4.2.1 Motion representation

A motion representation is explained. We choose one of the 6 persons and create 3 eigenspaces representing 3 motions by employing the three motion video images taken by the 4 cameras. For example, an eigenspace for representing motion_1:ShakeHand is computed employing 4 video image streams of the motion. For simplicity, 3(=k) eigenvalues are chosen for defining the three eigenspaces. Their cumulative proportion K is given in **Table 1**. Employing Eq.(10), 4 video images of a single motion are projected onto the defined eigenspace. Representation form of the 3 motions is illustrated in **Figs.3-5**: Figure 3 shows Motion_1: ShakeHand, Fig.4 gives Motion_2: PickUp, and Fig.5 depicts Motion_3: Step.

Each video image is sampled by 30fps, which amounts to R frames. (R equals 22, 36, 9 in the case of Motion_1, Motion_2, and Motion_3, respectively.) Therefore a single video image yields R projected points in the eigenspace. In Figs.3-5, the upper two graphs show respective curved lines containing R points with respect to the 4 cameras, whereas, in the lower two graphs, the corresponding projected points representing the same motion frames but different appearances are connected to show the approximate image of a surface patch. Note that a pair of eigenspaces are presented in each figure in order to be observed in a 3-D way: The left eigenspace image is for the left eye, whereas the right one for the right eye.

4.2.2 Motion Recognition

This subsection explains motion recognition using the eigenspace. In this research, three kinds of motion (Motion_1:ShakeHand, Motion_2:PickUp, Motion_3:Step) are employed in the recognition experiment. The judgment to which motion unknown data resembles leads to motion recognition. First, 3 eigenspaces are made from a single person's 3 motion image streams. Next, motions of five persons are projected one by one onto the created eigenspaces, and they are recognized using Eqs.(12),(13),(14).

Table 1. Cumulative propotion K (%) with respect to the 3 motions.

	ShakeHand	PickUp	Step
Cumulative Propotion	67.4	43.2	59.9

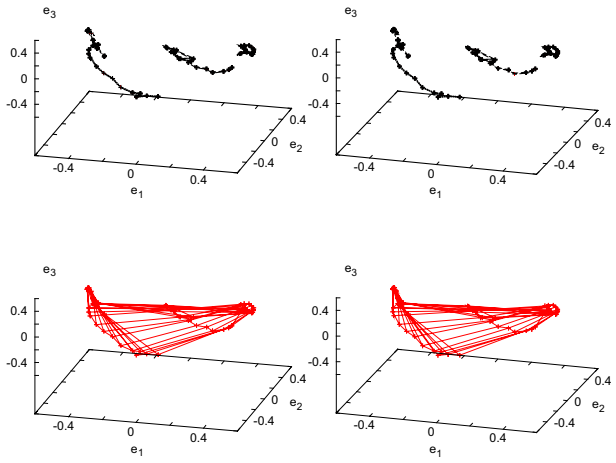

Fig. 3. Representation of motion 1: ShakeHand.

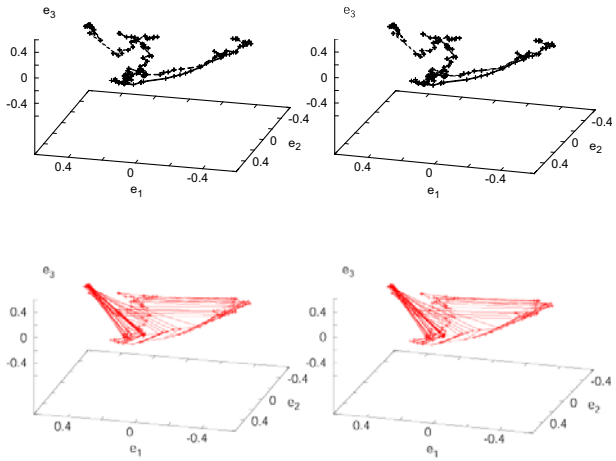

Fig. 4. Representation of motion 2: PickUp.

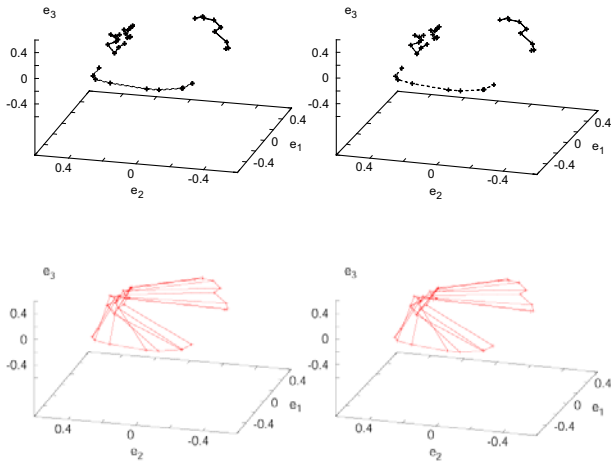

Fig. 5. Representation of motion 3: Step.

The graph showing the change of cumulative proportion when generating an eigenspace using learning image data is shown in **Fig 6**. From this figure, an eigenspace expression can be effective if 100 eigenvalues (k=100) are chosen defining the 100-dimensional eigenspace. For k=100, the cumulative proportion becomes 97.4% from Eq.(9).

Result of the recognition employing 100-dimensional eigenspace is shown in **Table 2**. The incorrect recognition case is marked × and the misclassified motion is given in the parenthesis. The recognition results corresponding to camera 1, camera 2, camera 3 and camera 4 are also shown in **Table 3**, **Table 4**, **Table 5** and **Table 6**, respectively.

5. DISCUSSION

We have proposed a technique for representing and recognizing human motions employing an eigenspace defined from video image streams obtained from multiple views based on the camera set surrounding a human in motion. Examples of motion representation by an eigenspace are given in Figs.3-5, in which particular motions are expressed as a digital curved surface patch. This representation technique is what we propose in this paper. The advantages of the technique over others include that (i) a human motion can be dealt with numerically as a curved surface in an eigenspace, (ii) every appearance is included in the representation, and (iii) the computation load is lower than 3-D representation technique as the computation is 2-D image base. The last advantage largely contributes to real time motion recognition [7].

It can easily be understood that, with the employment of more number of surrounding cameras, this surface patch becomes smoother. In this way, a human motion is represented by a smooth surface patch in an eigenspace. This representation absorbs the difference of observation orientation, since every appearance of the motion concerned is memorized in the surface patch.

As shown in Figs.3-5, the surfaces patches seem somewhat like a hammock. However it is actually an approximation in the 3-D eigenspace. Real shape reveals itself when we employ more number of eigenvalues and eigenvectors for defining the eigenspace; say, 100 eigenvalues as is indicated in Fig.6. Not very smooth part of the surface patch may disappear in the real shape having a smooth surface.

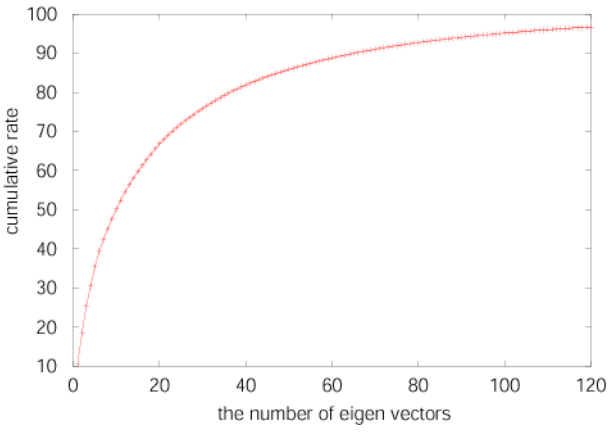

Fig. 6. Change of the cumulative proportion in the performed experiment.

Table 2. Result of motion recognition (all data)

Motion	Motion 1	Motion 2	Motion 3
Person 1	×(Motion_3)	O	O
Person 2	×(Motion_3)	O	O
Person 3	×(Motion_3)	O	O
Person 4	O	O	O
Person 5	O	O	O

Table 3. Result of motion recognition (camera 1)

Motion	Motion 1	Motion 2	Motion 3
Person 1	O	O	O
Person 2	×(motion2)	O	O
Person 3	×(motion3)	×(motion3)	O
Person 4	O	×(motion1)	O
Person 5	O	O	O

Table 4. Result of motion recognition (camera 2)

Motion	Motion 1	Motion 2	Motion 3
Person 1	×(motion2)	O	O
Person 2	O	O	O
Person 3	×(motion3)	O	O
Person 4	O	O	O
Person 5	O	O	O

Table 5. Result of motion recognition (camera 3)

Motion	Motion 1	Motion 2	Motion 3
Person 1	×(motion3)	O	O
Person 2	×(motion2)	×(motion3)	O
Person 3	×(motion3)	O	O
Person 4	×(motion3)	O	O
Person 5	×(motion3)	O	O

Table 6. Result of motion recognition (camera 4)

Motion	Motion 1	Motion 2	Motion 3
Person 1	×(motion3)	×(motion3)	×(motion2)
Person 2	×(motion3)	O	O
Person 3	×(motion3)	O	O
Person 4	×(motion3)	O	O
Person 5	×(motion3)	×(motion3)	O

The overall result of the recognition, given in Table 3, is described on the basis that a motion is regarded as recognized exactly, if two or more cameras recognize it exactly. It tells that Motion_2:PickUp and Motion_3:Step are well recognized among 5 persons. This may come from the fact that they are large motions compared to Motion_1. They can be discriminated well from every camera. On the other hand, the proposed technique does not give good performance with Motion_1:ShakeHand, which is also indicated by Tables 3-6. Actually cameras 3 and 4 didn't recognize Motion_1 at all. One of the main reasons of this may be that the motion where a person shakes his right hand is not a large motion enough to be recognized by the cameras located at the opposite side of the right hand. To solve this difficulty, we need employing more number of cameras in front of the person and increasing the number of the cameras which recognize rather local motions.

The experiment performed at the moment is only a preliminary experiment. We are now collecting motion data from larger number of persons. They will be employed as learning data for defining the eigenspaces representing respective motions and for performing the recognition experiment by introducing the leave-one-out method that will yield more reliable results.

6. CONCLUSION

A technique was proposed for representing and recognizing human motions. An eigenspace was introduced for the representation of a motion. The eigenspace was defined by multiple image streams of a motion obtained from a set of cameras that surround a person in the motion. A motion was described as a curved surface patch in the eigenspace. To recognize a motion, the image stream of an unknown motion was projected onto an eigenspace yielding a line describing the motion. The distance was computed between the projected line and the surface patches of the stored motions. Along with some rules on judgment, the surface patch having the shortest distance was judged as the motion to be recognized.

The advantages of the proposed technique over others are that, since a human 3-D motion is described numerically as a curved surface patch in an eigenspace, the representation fits for numerical analysis and evaluation of a motion, and that the computation load is much lower than other 3-D representation technique as the computation is 2-D image base in the technique.

Future problems include to increase motion data so that the learned eigenspace less depends on individual person's motion data, and to employ more number of cameras to surround a person in motion in order to create a smoother and more exact surface patch that represents the motion.

REFERENCES

[1] H. Murase, S. K. Nayar: "3D object recognition from appearance – parametric eigenspace method", *Trans. on IEICE*, **J77-D-II**, 11, pp.2179-2187, 1994.

[2] H. Murase, R. Sakai: "Moving object recognition in eigen-space representation: Gait analysis and lip reading", *Pattern Recognition Letters*, **17**, pp.155-162, 1996.

[3] T. Watanabe, M. Yachida: "Real time gesture recognition using eigenspace from multi-input image sequences", *Trans. on IEICE*, **J81-D-II**, 5, pp.810-821, 1998.

[4] M. M. Rahman, S. Ishikawa: "A mean eigen-window method for partially occluded/destroyed objects recognition", *Proc. of 7th Int. Conf. on Digital Image Computing: Techniques and Applications*, pp. 929-936, 2003.

[5] M. M. Rahman, S. Ishikawa: "A robust recognition method for partially occluded/destroyed objects", *Proc. of the Sixth Asian Conf. on Computer Vision*, pp. 984-988, 2004.

[6] O. Masoud N. Papanikolopoulos: "Recognizing human activities", *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pp.157-162, 2003.

[7] T. Ogata, J. K. Tan, S. Ishikawa: "Real time motion recognition employing an eigenspace", *Proc. of SICE2004*, 2004. (to appear)