MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan

3-3

On Describing Human Motions in an Eigenspace

Satoshi Homan, Takehito Ogata, Joo Kooi Tan, Seiji Ishikawa
Kyushu Institute of Technology, Department of Control Engineering
Sensuicho 1-1, Tobata, Kitakyushu 804-8550, Japan
{homan, Ogata}@ss10.cntl.kyutech.ac.jp, {etheltan, ishikawa}@cntl.kyutech.ac.jp

Abstract

Recognition of human motions using their 2-D appearance images has various applications. An eigenspace method is employed in this paper for representing and recognizing human motions. An eigenspace is created from video images taken by multiple cameras that surround a human in motion. The image streams of the motion compose several curved lines in the eigenspace that are closely situated with each other. The motion is described by this set of lines, which ultimately compose a curved surface. It is used for recognizing a human motion observed from an arbitrary orientation. Performance of the proposed technique was examined experimentally and promising results were obtained.

1 Introduction

Automatic human motion recognition by computer may have various applications such as detecting a person behaving in an abnormal way in a surveillance system, discovering a person who feels bad and finds difficulty in walking in order for an intelligent robot to help him/her, monitoring activities of aged people at home for their safety, etc.

There have been studies on automatic human motion recognition [2-6], but none of them has been put into practical use yet. 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 solve this, one has to recover shape and make a 3-D model of the motion interested. But it certainly increases computation load, resulting in batch procedure [8].

In order to overcome this difficulty, we propose the employment of an eigenspace [1] based on multiple views for human motion representation. An eigenspace is created from a set of video images 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 and contributes to simpler and faster computation than the above 3-D modeling.

The proposed technique is mathematically described and its performance is shown by an experiment employing some human motions and multiple cameras.

2 Eigenspace for Human Motion Description

An eigenspace method [1] is one of the techniques used for recognizing a 3-D object from its 2-D images. Since it recognizes a 3-D object as an aggregate of 2-D images, it excels other 3-D object recognition techniques in less computation time and less storage capacity. Indeed, the eigenspace method employs fewer stages of image processing before making the space describing human motions compared with other existent techniques.

An input image is given the form of a column vector $\hat{\boldsymbol{x}}$ as follows;

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

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

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

The normalized image vector $\boldsymbol{x}$ is expressed by

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

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

When a motion image of a person is taken, successive frames normally have high correlation, since a motion shows continuous change of human form. From this fact, motion image streams 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 appropriate eigenvectors corresponding to the largest eigenvalues, the original data is well represented in the reduced eigenspace.

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

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

This defines an image data matrix X of the form

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

Data matrix X then defines a covariance matrix Q of the form

$$Q = XX^{\mathrm{T}}. \quad (6)$$

Eigenvalues of the covariance matrix Q is obtained from solving the following eigenvalue problem;

$$Q\boldsymbol{u} = \lambda \boldsymbol{u}. \quad (7)$$

According to Karhunen-Loeve transform, the obtained N eigenvalues λ_k (k=1,2,…,N) are arranged in the descending order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq \cdots \lambda_N$,and the k eigenvectors corresponding to the largest k eigenvalues are chosen to define a k-dimensional subspace called an eigenspace. In this way, an eigenspace describing M human motions is made from $RPMH$ image data collected from P cameras and H persons. This is in practice the stage of knowledge basis creation.

Let us denote the chosen k eigenvectors by $\boldsymbol{e}_i$ (i=1,2,…,k). Image $\boldsymbol{x}_{r,p}^{m,h}$ is approximately described by

$$\boldsymbol{x}_{r,p}^{m,h} \approx \sum_{i=1}^{k} a_i \boldsymbol{e}_i + \boldsymbol{c}. \quad (8)$$

The image $\boldsymbol{x}_{r,p}^{m,h}$ is represented by a point (a_1,a_2,…,a_k) in the defined eigenspace. If we denote the point by $\boldsymbol{g}_{r,p}^{m,h}$, from Eq.(8), we have

$$\begin{aligned} \boldsymbol{g}_{r,p}^{m,h} &= \left(\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_k \right)^{\mathrm{T}} \left(\boldsymbol{x}_{r,p}^{m,h} - \boldsymbol{c} \right) \\ &= \left(a_1, a_2, ..., a_k \right)^{\mathrm{T}} \end{aligned} \quad (9)$$

All the $RPMH$ image data are represented in the eigenspace by points.

For a certain motion m, the average point defined by

$$\boldsymbol{g}_{r,p}^{m,-} \equiv \boldsymbol{g}_{r,p}^{m} = \frac{1}{H} \sum_{h=1}^{H} \boldsymbol{g}_{r,p}^{m,h} \quad (10)$$

represents frame r of motion m observed at camera p. A set of the average points

$$S^m = \left\{ \boldsymbol{g}_{r,p}^{m} \mid r = 1,2,...,R; p = 1,2,...,P \right\} \quad (11)$$

describes motion m. This is a discrete case: If parameters r and p take continuous values, the set S^m will give a surface patch instead of a set of points. 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 after all described as a manifold in an eigenspace. This is, however, not taken into consideration in this particular paper.

The cumulative proportion K is defined by

$$K = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{N} \lambda_i. \quad (12)$$

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

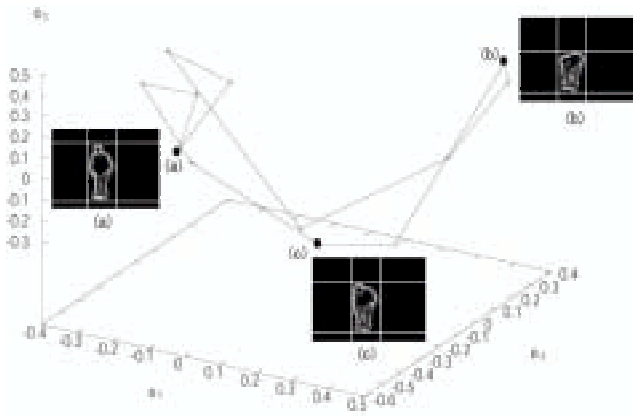


Fig. 1. Example of motion description in an eigenspace. A motion of a person observed from a single camera is described by a single closed curve in the eigenspace.

Figure 1 shows an example of an eigenspace representing a motion in which (a) a person stands straight initially, (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 that compose a closed line segments in the eigenspace. It is noted that we employ differential images of the original gray value images as shown in Fig.1, in order to escape from the dress effect [4].

3 Recognizing Human Motions

Let us assume that the created eigenspace describes M human motions within it. They are given by

$$S = \left\{ S^m \mid m = 1,2,...,M \right\} \quad (13)$$

In a recognition stage, an image stream of an unknown motion is projected into the eigenspace. As the projected image stream is a set of projected points, their proximity to one of the M motions 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 denoted by $\boldsymbol{x}$. An unknown image $\boldsymbol{x}$ is projected onto a point $\boldsymbol{g}$ in the eigenspace by the following formula;

$$\boldsymbol{g} = \left(\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_k \right)^{\mathrm{T}} \left(\boldsymbol{x} - \boldsymbol{c} \right). \quad (14)$$

In the eigenspace, a motion that may contain the projected point $\boldsymbol{g}$ is searched out of S given by (13). The distance between $\boldsymbol{g}$ and $\boldsymbol{g}_{r,p}^{m}$ is evaluated as shown in Eq.(15);

$$d_{\min} = \min_{r,p,m} \left\| \boldsymbol{g} - \boldsymbol{g}_{r,p}^{m} \right\| \equiv d_{r^*,p^*}^{m^*}. \quad (15)$$

If, for a certain small positive threshold ε,

$$d^{m*}_{r*,p*} < \varepsilon \tag{16}$$

holds, the image $\boldsymbol{x}$ is recognized as the r*th image of motion m* observed by camera p*.

Suppose that the observed image stream of an unknown motion m_{u} contains T successive image frames, i.e., $\boldsymbol{x}_t^{m_u}$ (t=1,2,…,T). Every frame $\boldsymbol{x}_t^{m_u}$ can be recognized by the above procedure. Then, we have two rules for overall recognition. Note that we have M motions in the designed eigenspace. Let us denote the number of image frames that have $d_{\min}$ with respect to motion m by p_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 q_m (m=1,2,…,M). Then we have the following rules;

$$\begin{aligned}&\{\{p_{m*} \geq p_m (m=1,2,...,M)\} \cap \{p_{m*} \geq T/2\}\} \cup \\ &\{\{p_{m*}+q_{m*} \geq p_m+q_m (m=1,2,...,M)\} \cap \{p_{m*} < T/2\}\} \\ &\Rightarrow m_{\mathrm{u}} = m*\end{aligned} \tag{17}$$

In this way, an unknown motion is recognized.

4 Experimental Results

An experiment was performed in the following way. Four digital video cameras (P=4) were placed fixed in front of a person in motion in a laboratory. Every view angle between successive cameras makes 30 degrees. Thus the 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 something from a floor (abbr. PickUp), and motion 3: stepping (abbr. Step), are acted by six (H=6) male students of early twenties.

We choose one of the 6 persons who acted the three motions and, from his image data, we create an eigenspace describing the 3 motions by employing 4 video images with each motion. For example, an eigenspace description of motion 1: ShakeHand is computed employing 4 video image streams of the motion. Employing Eq.(9), 4 video images of a single motion are projected into the defined eigenspace. Description form of motion 1:ShakeHand is illustrated in **Fig.2**. For simplicity, 3 eigenvalues are chosen for displaying the defined eigenspace.

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 Fig.2, the upper two graphs show respective curved lines containing 22 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 shape of the point set corresponding to motion 1:ShakeHand. Note that a pair of eigenspaces are presented in the figure with certain disparity in order to be observed in a stereoscopic way: The left image is for the left eye, whereas the right one for the right eye.

Three motions (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, an eigenspace is designed from a single person's 3 motion image streams. Next, motions of five persons are projected one by one onto the eigenspace, and they are recognized using Eqs.(15),(16),(17).

In this experiment, 100 eigenvalues (k=100) are chosen, defining the 100-dimensional eigenspace. For k=100, the cumulative proportion becomes 97.4% from Eq.(12).

Results of the recognition are shown in **Table 1**. It is the result with five persons employing the eigenspace defined from the image data of the 6th person. Correct recognition is given a circle, whereas the incorrect represented by X followed by a parenthesis containing the misclassified result.

5 Discussion and Conclusions

We have obtained satisfactory results of recognition with respect to motion 2 and motion 3, as shown in Table 1. On the other hand, there were some misclassification cases with respect to motion 1. This may be because motion 1:ShakeHand is a motion with a narrow range compared with other two motions. Shaking a right hand may not be well observed from the cameras placed at the left of the person. Motion 2 and motion 3 are, on the contrary, recognized exactly as they are wider motions. The number of cameras and their placement are, however, vital in order to achieve better recognition and they should be investigated further.

We have proposed a technique for representing human motions employing an eigenspace defined from video image streams obtained from multiple views based on the camera set surrounding a human in motion. This representation technique is the main issue 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 points set in an eigenspace;

(ii) every appearance can be included in the representation; and

(iii) computation load is much lower than 3-D representation technique, when making an eigenspace as a knowledge basis, as it is 2-D image base computation.

The last advantage may contribute to real-time motion recognition [7].

It can easily be understood that, with the employment of more number of surrounding cameras and with shorter sampling time, the points in the set representing a motion increase to a large extent. Ultimately, human motions will be described as smooth surface patches in an eigenspace. This representation absorbs the difference of observation orientation, since every appearance of the motion concerned will be memorized in the surface patch.

Table 1. Result of motion recognition.

	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

Human motion recognition has broad application fields. An intelligent robot in future, for example, should recognize motions of persons around it in real-time and give them hands for help, if necessary. In order to recognize human motions from every direction of observation, the proposed motion description technique should be employed over others, since other existent techniques have weakness in sensitiveness to the orientation of observation.

The experiment performed at the moment is still a preliminary experiment. We have been collecting motion data from more persons. They will be employed for defining an eigenspace describing the chosen motions and for performing the recognition by introducing the leave-one-out method that will yield more reliable results.

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.

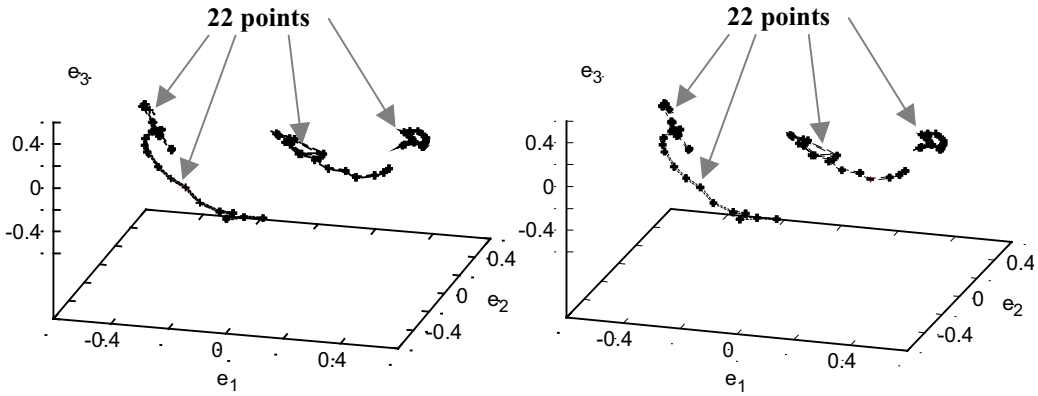
[4] M. M. Rahman, S. Ishikawa: "Robust appearance-based human action recognition", *Proc. of 2004 International Conference on Pattern Recognition,* CD-ROM, Cambridge, 2004.

[5] M. M. Rahman, S. Ishikawa: "Human motion recognition using an eigenspace", *Pattern Recognition Letters,* Elsevier Science. (to appear)

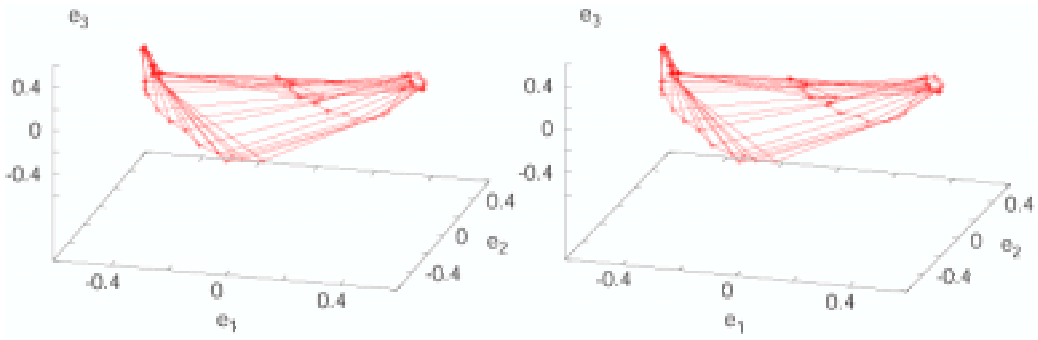
[6] O. Masoud, N. Papanikolopoulos: "Recognizing human activities", *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pp.157-162, 2003.

[7] T. Ogata, M. M. Rahman, J. K. Tan, S. Ishikawa: "Real time human motion recognition based on a motion history image and an eigenspace", *Proc. of SICE Annual Conf.*, 1901-1904, Sapporo, 2004.

[8] M. Yamamoto, et al.: Incremental tracking of human actions from multiple views, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.2-7, 1998.



(a)



(b)

Fig. 2. Representation of motion 1: ShakeHand: (a) Projected image points, and (b) the image corresponding points are connected. Stereoscopic images are provided.