ELSEVIER

Pattern Recognition Letters 17 (1996) 155–162

Pattern Recognition Letters

Moving object recognition in eigenspace representation: gait analysis and lip reading

Hiroshi Murase *, Rie Sakai

NTT Basic Research Laboratories, 3-1, Morinosato Wakamiya, Atsugi-Shi, Kanagawa, 243-01, Japan

Received 25 April 1995; revised 6 November 1995

Abstract

This paper describes a new method to calculate the spatio-temporal correlation efficiently in a parametric eigenspace representation for moving object recognition. A parametric eigenspace compactly represents the temporal change of an image sequence by a trajectory in the eigenspace. This representation reduces the computational cost of correlation-based comparison between image sequences. Experiments for human gait analysis and lip reading show this method is computationally useful for motion analysis and recognition.

Keywords: Motion analysis; Object recognition; Eigenspace representation

1. Introduction

Motion analysis is one of the most active and challenging research areas in computer vision. Many of the goals are tracking the moving objects or recovering the 3D objects (Aggarwal and Nandhakumar, 1986) using optical flow or feature correspondence. In this field, the automatic interpretation of human movement from an image sequence has been gaining more attention, because it has a variety of applications, such as individual recognition, gesture recognition, and lip reading (Niyogi and Adelson, 1994; Murase, 1992; Bregler and Konig, 1994; Del Bimbo and Nesi, 1992; Rohr, 1993).

Some recognition methods used the fact that the human body consists of body parts linked to each other at joints (for example, Rohr, 1993; Del Bimbo and Nesi, 1992); however, these methods require structure models, which should be changed depending on the application. An alternative is to consider the property of the spatio-temporal pattern as a whole. For example, Niyogi and Adelson (1994) used the spatio-temporal edge of the body boundary in spatio-temporal volume, and Murase (1992) proposed the method which observes the silhouette movements of the object in specific feature extraction windows. However, this feature-based method is also specific to the application, because the programmer should design the apropriate features for a specific set. Moreover, features, such as edges, corners, or boundaries cannot be easily extracted from noisy images. Thus, feature-based methods may have some limitation in application.

Polana and Nelson (1993) looked at spatio-temporal Fourier transforms in order to classify activities. Such an approach probably has the strength to reduce noise; however, the low-frequency compo-

* Corresponding author. Email: murase@siva.ntt.jp

SSDI 0167-8655(95)00109-3

nents do not efficiently represent the original patterns. Another simple method is template matching (e.g., spatio-temporal image correlation). It is applicable to various object sets, and is reasonably robust to small noise. The calculation time, however, increases quickly if comparison is performed in a spatio-temporal domain, especially when time-axis stretching is taken into account.

This paper proposes the parametric eigenspace representation for efficient image sequence comparison. We apply this idea to the recognition of people by their walk and to the lip reading problem. The

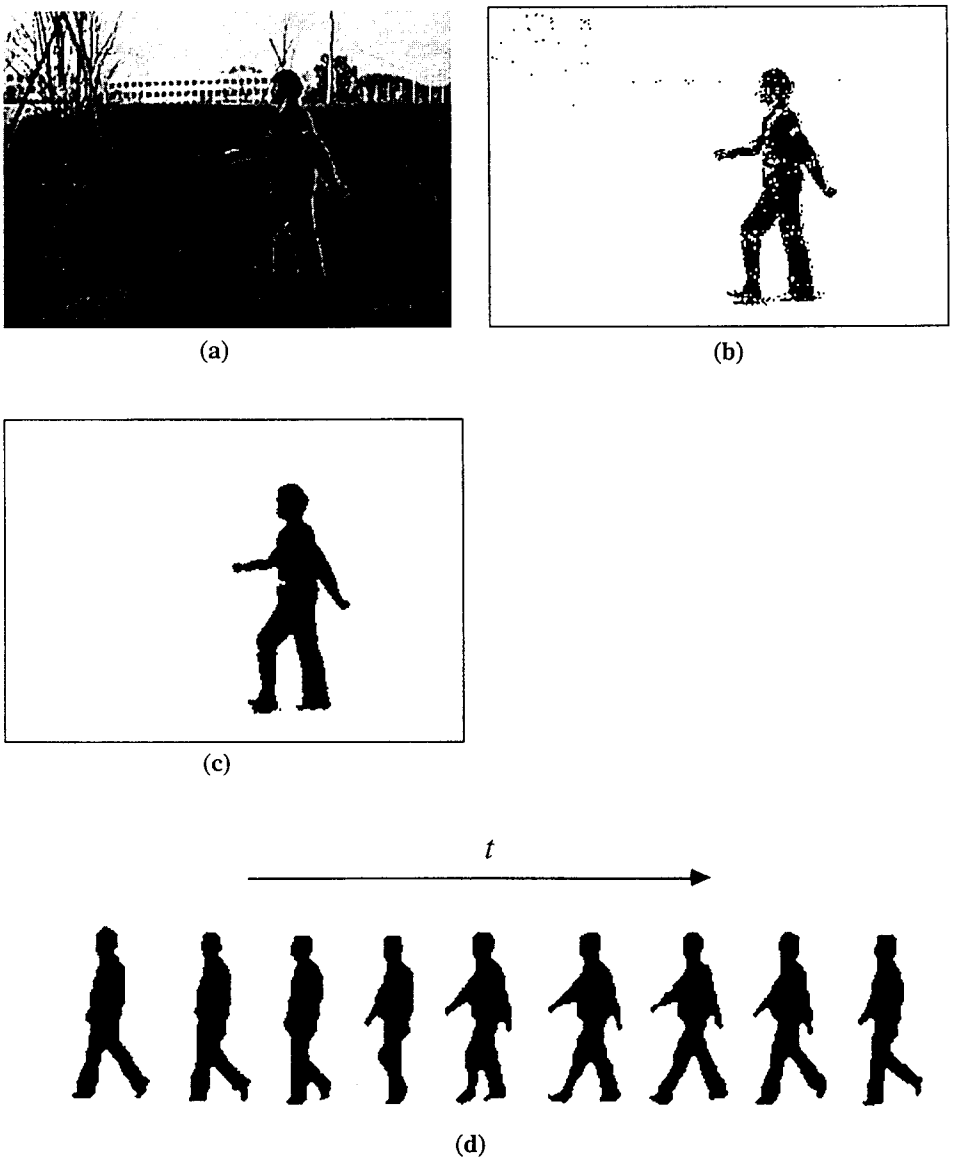

Fig. 1. (a) The original image. (b) The subtracted image followed by binarization. A lot of noise occurs. (c) A noise-reduced image from (b) using smooth filtering. (d) Changes in the gait pattern.

eigenspace representation can represent pattern sets more efficiently than the Fourier representation, and it has been applied for face recognition (Sirovich and Kirby, 1987; Turk and Pentland, 1991), character recognition, etc. (Fukunaga, 1990; Press et al., 1988). Especially the parametric eigenspace representation, which represents a variety of images using a manifold in the eigenspace, can be used for a wide range of applications such as object recognition, object extraction, object tracking, and illumination planning (Murase and Nayar, 1994, 1995; Nayar et al., 1994). We extend this idea to image sequence recognition.

We first address the problem of individual recognition using motion information. This type of problem must be solved in order to build security or other systems that can find a particular person in a crowd. The precision of individual recognition increases when we use a combination of information about the person's face, build, and other factors. Motion information is one of the good cues to recognize individuals. Psychological experiments show that people's gaits contain much information about them. For example, results of a psychological experiment using an MLD (Moving Light Display) show it is possible for humans to identify an individual from paths of lights when a person walks with lit bulbs on his hands and feet (Cutting and Kozlowski, 1977). We show that gait analysis can also be used in a machine recognition of an individual.

Some feature-based recognition methods that extract the boundaries or edges of a figure have been proposed for gait analysis (Niyogi and Adelson, 1994; Murase, 1992). In gait analysis with the parametric eigenspace representation, however, a special feature extraction is not used. Our method is similar to spatio-temporal image correlation, but, using the eigenspace reduces calculations and provides more robustness to noise because of the effective representation of movement. The recognition process is composed of three steps: making a sequence of silhouette images by extracting a walking person from the background, projecting the images to the eigenspace, and comparing the images with reference patterns in a database.

This method is a general algorithm for moving object recognition, so it can be applied to many problems. We show our method is also applicable to lip reading, which is important because research has shown that efficient lip reading increases the voice recognition precision (Bregler and Konig, 1994).

2. Extracting a gait from the background

We assume that (i) individuals are walking frontoparallel to the camera with a fixed background, and (ii) the body is not occluded. This situation can be easily realized by setting a camera in a proper position. To extract a person area from the background, we can simply take subtraction of two images. The difference between them will be a silhouette of a person. Fig. 1(a) shows an example of an input image, and Fig. 1(b) shows the difference pattern. A lot of noise such as isolated spots or holes occur if we use a simple extraction method like subtraction. It is possible to eliminate small noises by applying smooth filtering and thresholding; however it is not necessary to remove all of them. Fig. 1(c) shows an image where much of the noise has been removed by smooth filtering. Template matching is not sensitive to noise, so this silhouette pattern with some noise was directly applied to the next step.

As a preliminary process, the position and size of the silhouette are normalized to fit the input image frame. The aspect ratio will be kept constant when the size is normalized. Fig. 1(d) shows changes in the gait pattern.

3. Spatio-temporal correlation

Image correlation is a well-know technique for measuring the similarity of images in many practical situations. Spatio-temporal correlation is an extension of 2-dimensional image correlation to 3-dimensional correlation in the space and time domain. Let an input image sequence be

$$\boldsymbol{x}(t) = [x_1(t), x_2(t), \ldots, x_N(t)]^{\mathrm{T}},$$

and a reference image sequence be

$$\boldsymbol{y}(t) = [y_1(t), y_2(t), \ldots, y_N(t)]^{\mathrm{T}}.$$

Here, N is the number of pixels in an image. The elements of these vectors are pixel values of the

image at time t. Spatio-temporal correlation, c, can be written

$$c = \int_{t=0}^{T} \boldsymbol{x}(t)^{\mathrm{T}} \boldsymbol{y}(t)\, \mathrm{d}t.$$

If we consider time shifting and stretching, this equation can be written

$$c = \int_{t=0}^{T} \boldsymbol{x}(t)^{\mathrm{T}} \boldsymbol{y}(w(t))\, \mathrm{d}t,$$

where $w(t)$ is the warp function, which is written $w(t) = at + b$ in the case of time shifting and stretching. Here, a and b in $w(t)$ depend on the velocity change and phase difference, respectively, for each observation. We assumed that the gait is periodic and the velocity is not changed during the observation.

The computational cost for correlation will be high with changing values of a and b in $w(t)$ when the size of images is large. It is possible to shorten the computation time by an orthogonal transformation, which reduces the dimension of an input vector. For example, the transformation using the eigenvector basis $[\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k]$ allows an input image and a reference image to be approximated as k-dimensional vectors, i.e., $\boldsymbol{z}(t) = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k]^{\mathrm{T}} \boldsymbol{x}(t)$ and $\boldsymbol{v}(t) = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k]^{\mathrm{T}} \boldsymbol{y}(t)$ respectively. Here, $k \ll N$. The correlation is rewritten

$$\begin{aligned} c &= \int_{t=0}^{T} \boldsymbol{x}(t)^{\mathrm{T}} \boldsymbol{y}(w(t))\, \mathrm{d}t \\ &\cong \int_{t=0}^{T} \boldsymbol{z}(t)^{\mathrm{T}} [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k]^{\mathrm{T}} \\ &\qquad \times [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k] \boldsymbol{v}(w(t))\, \mathrm{d}t \\ &= \int_{t=0}^{T} \boldsymbol{z}(t)^{\mathrm{T}} \boldsymbol{v}(w(t))\, \mathrm{d}t. \end{aligned}$$

Thus, the computational cost will be reduced because it is calculated in a lower-dimensional subspace.

4. Learning stage

4.1. Calculating eigenspace from silhouette images

An ith gait image of a person j at time t is represented as $\boldsymbol{y}'_{ij}(t)$. First, the brightness of an image is normalized by

$$\boldsymbol{y}_{ij}(t) = \boldsymbol{y}'_{ij}(t) / \| \boldsymbol{y}'_{ij}(t) \|.$$

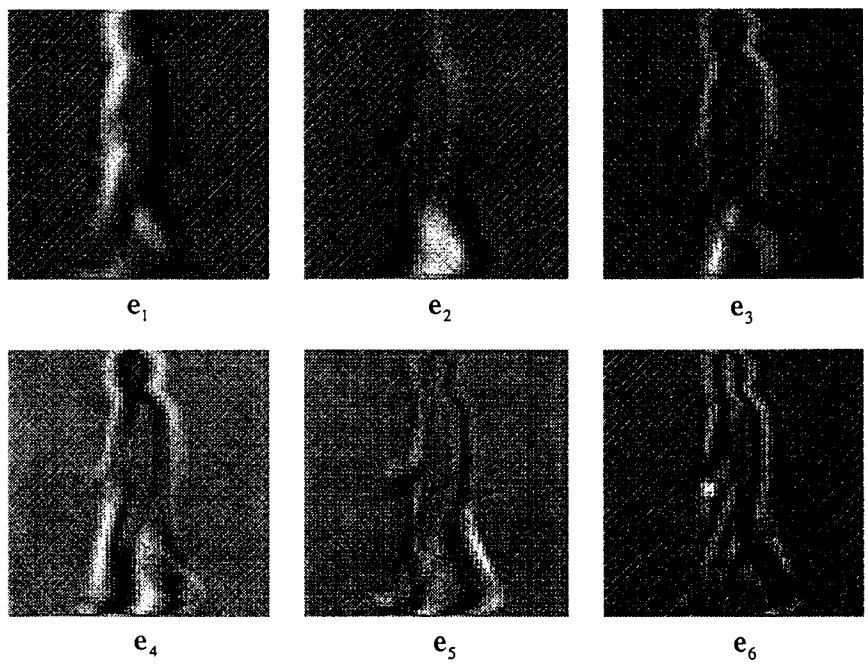

Fig. 2. The first six eigenvectors for gait pattern.

The covariance matrix of an image set $y_{ij}(t)$ is represented by

$$Q = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} (y_{ij}(t) - \bar{y})(y_{ij}(t) - \bar{y})^{\mathrm{T}},$$

where $\bar{y}$ is the mean vector for $y_{ij}(t)$. Next, k eigenvectors $e_1, e_2, \ldots, e_k$ ($\lambda_1 \geqslant \cdots \geqslant \lambda_k \geqslant \cdots \geqslant \lambda_N$) for the image set are calculated by the eigenvalue decomposition (Murase and Lindenbaum, 1995; Oja, 1983):

$$\lambda_i Q = Q e_i.$$

The k-dimensional subspace spanned by these eigenvectors is called the eigenspace. Fig. 2 shows a pattern of eigenvectors. Generally, the first few eigenvectors correspond to large changes in the pattern (low-spatial frequency), and higher-order eigenvectors represent smaller changes (high-spatial frequency).

4.2. Parametric eigenspace representation

An image can be mapped to a point in the eigenspace, therefore a sequential movement can be represented as a trajectory in the eigenspace. This is called the parametric eigenspace representation. An example of a gait pattern is show in Fig. 3. An only 3-dimensional eigenspace is shown here, whereas a more than 10-dimensional eigenspace is used for actual recognition. The trajectory is represented by a vector function $v_{ij}(t) = [e_1, e_2, \ldots, e_k]^{\mathrm{T}} y_{ij}(t)$ parameterized by t in a k-dimensional eigenspace.

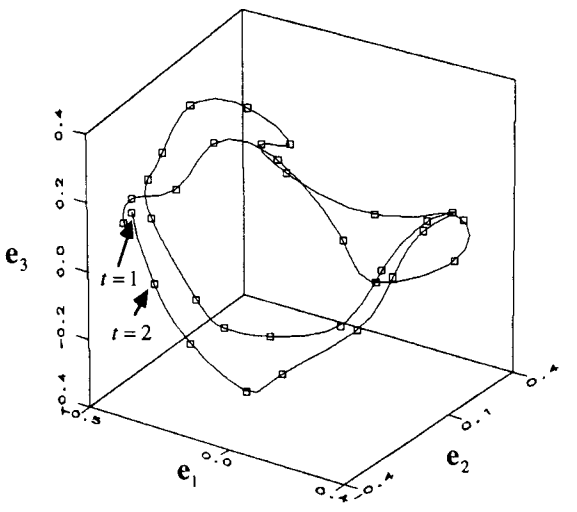

Fig. 3. Parametric eigenspace. An only 3-dimensional eigenspace is shown here. The trajectory is parameterized by time, t.

Speed of the motion often changes and it is sometimes not an essential factor. To solve this problem, we compare image sequences with linear time stretching. We prepare several reference patterns of $v_{ij}(at)$ in advance of the recognition stage and keep a variety of reference patterns in a database. The time stretching pattern $v_{ij}(at)$ is generated by resampling the original pattern. Here, we used linear interpolation because the sampling rate of the reference pattern is changed by time stretching.

5. Individual recognition using spatio-temporal correlation

We measure similarities between image sequences by spatio-temporal correlation. Let an input image sequence be $x'(t)$. This image sequence is obtained using the same preprocessing method presented in Section 2. First, the brightness of the image is normalized by

$$x(t) = x'(t) / \| x'(t) \|.$$

Then, this image vector is projected into the eigenspace by

$$z(t) = [e_1, e_2, \ldots, e_k]^{\mathrm{T}} x(t).$$

The distance between an input vector sequence, $z(t)$, and reference vector sequence, $v_{ij}(t)$, is

$$d_{ij}^2 = \min_{a,b} \sum_{t=1}^{T} \| z(t) - v_{ij}(at + b) \|^2.$$

This similarity measurement is invariant to time stretching and time shifting. It is known that this distance value is a reversed order of the spatio-temporal correlation, if the norms of the vectors $z(t)$ and $v_{ij}(at + b)$ are unity. This means that computing d_{ij}^2 is equivalent to computing the maximum spatio-temporal correlation:

$$c_{ij} = \max_{a,b} \int_{t=1}^{T} z(t)^{\mathrm{T}} v(at + b)\, \mathrm{d}t.$$

Finally, the recognition result is selected as j which minimizes the distance d_{ij}^2.

6. Experiment

6.1. Collecting data

We collected the gait patterns of seven people (10 each) wearing the same clothes. The sampling rate was 30 frames/second, and the original image size was 320×240. We extracted each person's area and normalized its size, position, and brightness as described in Section 2, then put it in a 64×64 image array. Five of them for each person are used for learning, and the rest are for recognition test.

6.2. Result of recognition tests

Recognition accuracy was estimated with five test data from each person. From Fig. 4 which shows the recognition rate with increasing eigenspace dimension, it is clear that a 16-dimensional eigenspace provides sufficient recognition accuracy.

We compare the computation time between our method and exhaustive spatio-temporal correlation (STC). We assume the following parameters, which are taken from the actual numbers in our experiments; image size: 64×64, the number of references: 7, the number of reference frames: 40, the number of input frames: 90, the number of time-stretch steps: 30, the dimension of the eigenspace: k. The number of pixel operations for our method and the exhaustive STC method are

$$64 * 64 * 40 * (90 - 40) * 30 * 7 = 1720 \text{ M},$$

$$64 * 64 * k * 90 + k * 40 * (90 - 40) * 30 * 7 = k * 0.778 \text{ M},$$

respectively. The numbers of operations and recognition rates for each method are listed in Table 1. Our method with 16 eigenspace dimensions reduces the computation time at the rate of one 136th without changing the recognition accuracy. We can reduce the computation time more using less eigenspace dimensions (i.e., 8); however, too much reduction lowers the recognition accuracy.

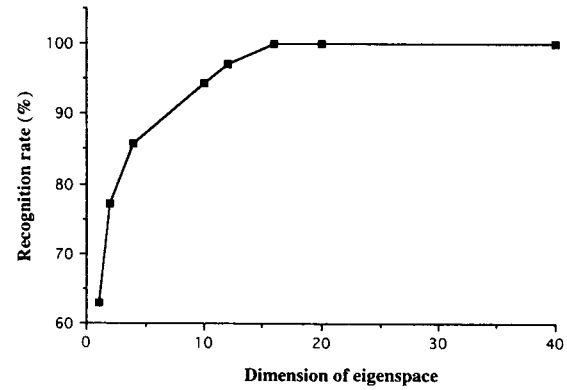

Fig. 4. The recognition rate with varying the dimension of the eigenspace. A 16-dimensional eigenspace provides sufficient recognition accuracy.

We tested our method's robustness to noise. A binary random dot pattern (black/white) $\boldsymbol{r}$ was generated, whose black area is $p\%$ of the whole image area. A test pattern was made by the Exclusive-OR operation of the two patterns $\boldsymbol{r}$ and $\boldsymbol{x}(t)$, and this pattern was fed to our recognition system instead of

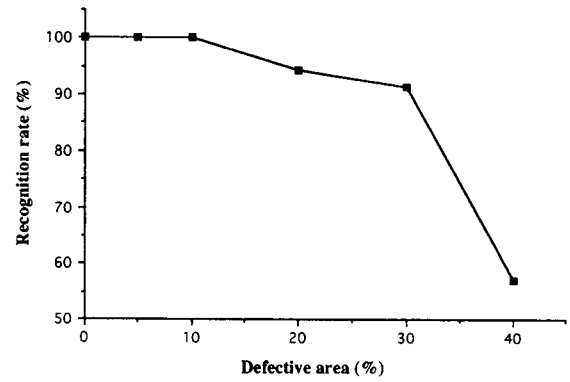

Fig. 5. The recognition rate with varying the defective percentage p, where a 16-dimensional eigenspace is used.

Table 1
Comparison between our method and spatio-temporal correlation

	Our method ($k = 16$)	Our method ($k = 8$)	Exhaustive STC method
No. of operations	12.6 M	6.3 M	1720 M
Recognition rate	100% (35/35)	88% (31/35)	100% (35/35)

$x(t)$. Here, p is a parameter that makes defects in the original pattern. Fig. 5 shows the recognition rate for increasing defective percentage p. This result shows that the recognition association is still high, even if 10% of the image area is defective due to noise. This means we need not be too careful with noise reduction in the preprocessing.

In this experiment, walkers wore the same shoes and clothes. From the practical perspective of human identification, we have to investigate the effect of their shoes and clothes. These are part of future research.

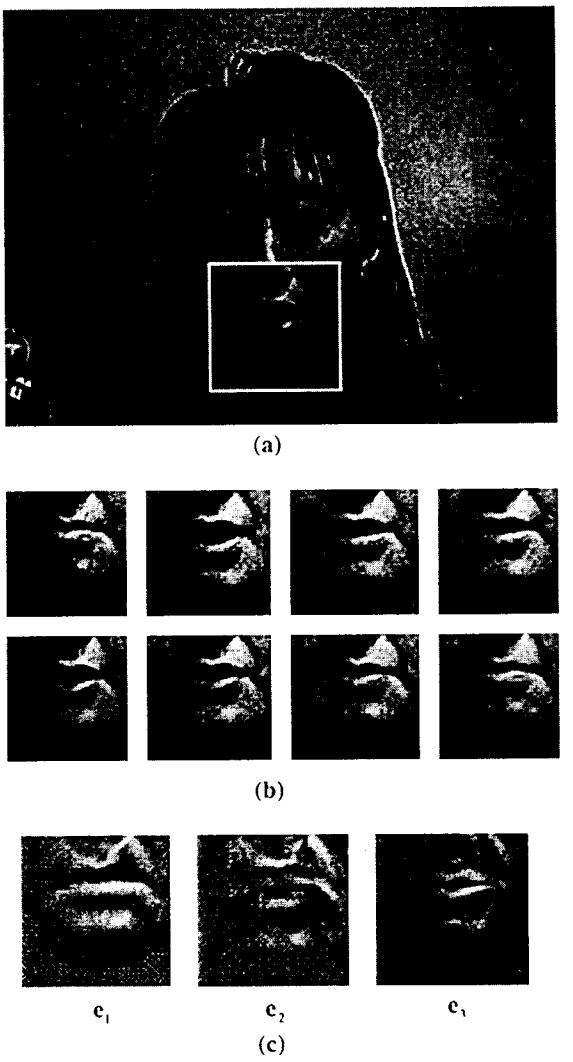

Fig. 6. (a) An extracted mouth area. (b) A sequence of images (saying "seven", sample rate: 7.5 frames/second). (c) The first three eigenvectors for this area.

7. Application for lip reading

Spatio-temporal image correlation can be used in many fields. This means our parametric eigenspace method is also widely applicable to moving object recognition. We conducted a simple experiment to investigate its applicability to lip reading. Input data was an image sequence of a mouth pronouncing the number one through ten in English. In the experiments, we took ten samples for each number, five for learning and the other five for recognition. The image sampling rate of inputs was 30 frames/second. First, the face area was detected by the difference from the background, then the mouth area was extracted using the positional relation in the face. The size of this part was normalized to a 64×64 image. We used the grey-level image directly in this case. Learning and recognition procedures were the same as in the gait experiment. Fig. 6 shows an extracted mouth area, an image sequence of a mouth pronouncing "seven", and the eigenvectors for the mouth area. The result of this experiment using 50 test samples tells us that a recognition rate of 76% can be achieved with a 16-dimensional eigenspace, whereas the same rate of 76% is obtained by spatio-temporal correlation. The computation time with our method, on the other hand, is 100 times faster than the exhaustive spatio-temporal correlation method.

8. Conclusion

We have described a moving object recognition method using spatio-temporal correlation and showed that the computational cost can be reduced by introducing the parametric eigenspace representation without losing recognition accuracy. The results of human gait and lip reading experiments show this method is robust to noise in an input image. We conclude that the parametric eigenspace representation is widely applicable to moving object recognition.

Acknowledgment

We would like to thank Dr. K. Ishii and Dr. S. Naito for encouragement.

References

Aggarwal, J.K. and N. Nandhakumar (1986). On the computation of motion from sequences of images – A review. *Proc. IEEE* 6 (1), 90–99.

Bregler, C. and Y. Konig (1994). Eigenlips for robust speech recognition. *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing*.

Cutting, J.E. and L.T. Kozlowski (1977). Recognizing friends by their walk: gait perception without familiarity cues. *Bull. Psyconomic Soc.* 19, 353–356.

Del Bimbo, A. and P. Nesi (1992). Behavioral object recognition from multiple image frames. *Signal Processing* 27, 37–49.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, London.

Murase, H. (1992). Recognizing individuals from the silhouettes of their walk. *IEICE Trans.* J75-DII (6), 1096–1098.

Murase, H. and M. Lindenbaum (1995). Spatial temporal adaptive method for partial eigenstructure decomposition of large images. *IEEE Trans. Image Processing*.

Murase, H. and S.K. Nayar (1994). Illumination planning for object recognition using parametric eigenspace. *IEEE Trans. Pattern Anal. Mach. Intell.* 16 (12), 1219–1227.

Murase, H. and S.K. Nayar (1995). Visual learning and recognition of 3D objects from appearance. *Internat. J. Computer Vision* 14, 5–24.

Nayar, S.K., H. Murase, and S.A. Nene (1994). Learning positioning, and tracking visual appearance. *IEEE Internat. Conf. on Robotics and Automation*.

Niyogi S.A. and E.H. Adelson (1994). Analyzing and recognizing walking figures in XYT. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 469–474.

Oja, E. (1983). *Subspace Methods of Pattern Recognition*. Research Studies Press, Herfordshire.

Polana, R. and R. Nelson (1993). Detecting activities. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2–7.

Press, W., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling (1988). *Numerical Recipes in C*. Cambridge Univ. Press, Cambridge.

Rohr, K. (1993). Incremental recognition of pedestrians from image sequences. *IEEE Conf. on Computer Vision and Pattern Recognition*, 8–13.

Sirovich, L. and M. Kirby (1987). Low dimensional procedure for the characterization of human faces. *J. Opt. Soc. Amer.* 4 (3) 519–524.

Turk, M.A. and A.P. Pentland (1991). Face recognition using eigenfaces. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 586–591.