Pattern Recognition Letters xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Three-way auto-correlation approach to motion recognition

Takumi Kobayashi *, Nobuyuki Otsu

National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1, Tsukuba 305-8568, Japan

ARTICLE INFO

Article history:
Received 26 November 2007
Received in revised form 9 September 2008
Available online xxxx

Communicated by G. Borgefors

Keywords:
Three-way data analysis
Motion feature extraction
Local auto-correlation
Gesture recognition
Gait recognition

ABSTRACT

This paper presents a feature extraction method for three-way data: the cubic higher-order local auto-correlation (CHLAC) method. This method is particularly suitable for analysis of motion-image sequences. Motion-image sequences can be regarded as three-way data consisting of *x*-, *y*- and *t*-axes. The CHLAC method is based on three-way auto-correlations of pixels in motion images. It effectively extracts spatio-temporal local geometric features characterizing the motion, such as gradients (velocities) and curvatures (accelerations). It has also several advantages for motion recognition. Firstly, neither *a priori* knowledge nor heuristics about the objects in question is required. Secondly, it is shift-invariant and thus segmentation-free. Thirdly, its computational cost is less than that of traditional methods, which makes it more suitable for real time processing. The experimental results on large datasets for gesture and gait recognition showed the effectiveness of the CHLAC method.

1. Introduction

Motion recognition is becoming an important area in computer vision, and has attracted an increasing number of researchers. The two approaches being considered are, firstly, recognition of the particular type of motion, and secondly, recognition of the performer of the motion. Gesture recognition corresponds to the former task and this has been developed over the last decade. It is considered to be an effective approach to studying human–machine interactions (Raytchev et al., 2000).

In terms of the latter approach, more recently, gait recognition has become an important focus within the video surveillance community (Nixon and Carter, 2006). The term "gait" refers to the manner of walking, which, if characterized precisely enough, is expected to become a biometric key for human identification. Unlike fingerprinting, this biometric method would have the advantage that human identification could be carried out by observing the gait through a video camera from a distance.

There are several difficult problems associated with motion recognition, namely segmentation, tracking, and analysis both of the human shape and its changes over time. Gait recognition requires an even more detailed analysis of motion and considerable effort has already been invested in these problems. Motion images are characterized as containing both spatial and temporal information that are difficult to treat all together and effectively. In typical approaches to motion-image analysis, these two kinds of information are processed individually: each image frame is processed and usually compressed into a feature vector before a time series of the resulting feature vectors is analyzed. While several previous studies have required specific knowledge of the motion or the performer, some psychological studies have suggested that such specific knowledge may not be required for motion recognition (Johansson, 1973; Cutting and Kozlowski, 1977).

One of the earliest psychological studies related to motion recognition, particularly gait recognition, was that of Johansson (1973) in which an experiment, using "point light display," revealed that we can perceive human motion using only the moving patterns of points of light as a cue (Fig. 1). In terms of employing gait to identify humans, Cutting and Kozlowski (1977) confirmed that humans can recognize a particular person walking by observing the moving points of light, even if familiarity cues are omitted. They also suggested that dynamic cues such as speed, bounciness and rhythm of the walker are more important than static cues such as the height of the walker. It is important to note that human gait can be recognized almost entirely by utilizing only dynamic cues. Cutting and Kozlowski (1977) noted that, "the perception of dynamic forms is probably not derived from the perception of static forms" and "snapshot recognition is a special case of motion recognition, where the dynamic invariance is null."

In this paper, cubic higher-order local auto-correlation (CHLAC) is presented, the basic idea of which is proposed in (Kobayashi and Otsu, 2004, 2006). *Auto-correlation* is associated with *relative* movement of points such as those in point light display. CHLAC is able to cope with static and dynamic cues simultaneously in a natural way using spatio-temporal auto-correlation. The motivation for using

* Corresponding author. Tel.: +81 29 861 5491; fax: +81 29 861 3313.
E-mail address: takumi.kobayashi@aist.go.jp (T. Kobayashi).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.09.006

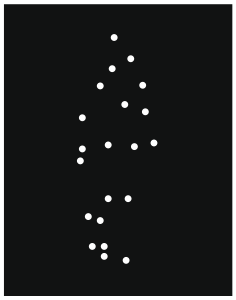

Fig. 1. Point light display. Points of light are perceived as random patterns in the static case, but once the object moves, the movement can be recognized entirely by means of the moving points of light.

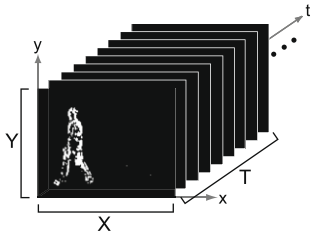

Fig. 2. Cubic data showing motion pixels as white points which are extracted by frame differencing and binarization.

CHLAC is similar to that discussed by Cutting and Kozlowski (1977) which deals with both spatial axes and the time axis equally, not as a compilation of snapshots. The key point is that dynamic perception is connected to the mutual relations among the moving points which we regard as spatio-temporal *auto-correlations* of the moving points (see Section 4 for details). CHLAC requires no *a priori* knowledge about objects and can be used as a method for versatile motion-feature extraction. Furthermore, motion images are analyzed simultaneously within a spatio-temporal context and do not require any two-step analysis for still images and time series.

In this paper, we apply the CHLAC method to the two motion recognition tasks: gesture and gait recognition. The experimental results on large datasets confirmed the effectiveness of the method.

The rest of the paper is organized as follows: the next section reviews previous studies related to gesture and gait recognition. We describe the preprocessing of motion-image sequences in Section 3 and details of CHLAC method in Section 4. In Section 5, the experimental results for gesture and gait recognition are shown. Finally, Section 6 contains our concluding remarks.

2. Previous studies

Firstly, we review related studies of gesture recognition. One of the traditional methods is that of Wilson and Bobick (1999) in which hidden Markov model (HMM) is employed to analyze time series of kinematic parameters of a human body. Recently, Kim et al. (2007) used canonical correlations to calculate the similarity between two video sequences by applying tensor canonical correlation analysis. The method assumes aligned sequences and is affected by background. The method of Dollar et al. (2005) is based on spatio-temporal interest points (Laptev, 2005) and visual code words (Sivic and Zisserman, 2003). In (Jhuang et al., 2007), the interest point detector becomes more sophisticated and is biologically inspired. Although these methods have resulted in good performance, it is difficult to apply them in real time due to the high computational load. The most closely related studies to our work are Raytchev et al. (2000) and Ishihara and Otsu (2004). These methods, however, considered only the first order auto-correlations, which are insufficient to capture details of motion information, as discussed in Section 4.3.

Next, studies related to gait recognition are reviewed. Sarker et al. (2005) used template matching of silhouettes which were roughly extracted by background subtraction. More sophisticated silhouette extraction method was employed together with HMM in (Lee et al., 2003). Tolliver and Collins (2003) applied the refined template matching method by using a variance-weighted metric and detecting key frames in a human gait. Sundaresan et al. (2003) applied HMM to the time series analysis of silhouettes. These methods are all based on human silhouettes and template matching (spatio-temporal *cross*-correlation) for calculating the similarities of frames. Wang et al. (2003) extracted features from the outer contour of silhouettes, while hardly making use of temporal information.

3. Preprocessing

Before applying CHLAC to motion images, pixel values are converted to binary values, based on the motion as follows.

Firstly, as shown in Fig. 2, an image sequence can be regarded as three-dimensional data: x- and y-axes in an image frame ($X \times Y$) and the t-axis (for time) along the frame sequence. Motion is usually composed of characteristic (sub-)motions over certain time durations, such as cyclic motion periods. For capturing the characteristics, we set a time window containing a constant number of frames along the t-axis. The frames within the window are assigned as one unit of three-way data, called "cubic data" ($X \times Y \times T$). A series of such cubic data units is obtained by shifting the window – say one frame at a time – along the time axis, where the width T of the window is a parameter to be determined. Human motion is recognized at each frame t by classifying the CHLAC feature vector of the cubic data.

Secondly, we apply frame differencing and then automatic-thresholding (Otsu, 1979) in order to detect and binarize motion pixels. These processes also filter out both inherent noise and brightness information, such as that due to clothing, which is irrelevant to the motion itself. Consequently, pixel values in each frame become either 1 (moved) or 0 (static). A moving human contour is visible (Fig. 2), and the contour is sufficient for motion recognition (Veres et al., 2004). A little isolated noise might be left in the resulting frames, but need not be eliminated since CHLAC is robust to such isolated noise (see Section 4.3). In this preprocessing, frame differencing could be replaced by another method, such as silhouette extraction or edge extraction. The extraction of silhouettes, however, requires more complicated processing (background estimation, etc.) whereas frame differencing and binarization are easily processed. It should be noted that CHLAC can deal with these kinds of preprocessing, whereas other methods of motion recognition based on template matching accept only silhouette extraction for preprocessing. Edge extraction and binarization are also easily processed and produce human contours similar to the frame differencing approach. Differences in the effects of these variants of preprocessing are discussed in Section 5.1.

4. Cubic higher-order local auto-correlation (CHLAC)

In this section, the general formulation and practical computation of the cubic higher-order local auto-correlation (CHLAC) (Kobayashi and Otsu, 2004, 2006) are presented. Higher-order local auto-correlation (HLAC) has been proposed for extracting spatial auto-correlations, and its effectiveness has also been demonstrated for static image (two-way data) recognition (Otsu and Kurita,

Please cite this article in press as: Kobayashi, T., Otsu, N., Three-way auto-correlation approach to motion recognition, Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.09.006

1988). We extend this naturally to CHLAC so as to deal directly with three-way data. In this framework, static perception related to HLAC is considered as a special case of dynamic perception related to CHLAC as suggested in (Cutting and Kozlowski, 1977).

4.1. Definition

Let $f(\boldsymbol{r})$ be three-way (cubic) data defined on the region $D: X \times Y \times T$ with $\boldsymbol{r} = (x, y, t)^T$, where X and Y are the width and height of the image frame and T is the time length of the time window. Then the Nth order auto-correlation function is defined as

$$R_N(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N) = \int_{D_s} f(\boldsymbol{r}) f(\boldsymbol{r} + \boldsymbol{a}_1) \cdots f(\boldsymbol{r} + \boldsymbol{a}_N) \mathrm{d}\boldsymbol{r} \quad (1)$$

$$D_s = \{\boldsymbol{r} \mid \boldsymbol{r} + \boldsymbol{a}_i \in D \ \forall i\},$$

where $\boldsymbol{a}_i (i = 1, \ldots, N)$ are displacement vectors from the reference point $\boldsymbol{r}$. Since Eq. (1) can take many different forms by varying N and $\boldsymbol{a}_i$, we limit $N \leqslant 2$ and $\boldsymbol{a}_i$ to a local region by focusing on the high correlation of the local voxels.

A CHLAC feature (vector) consists of $R_N(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)$ with various $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ in the local region and $N \in \{0, 1, 2\}$. However, in the case that the point configuration of $(\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(1)} + \boldsymbol{a}_1^{(1)}, \ldots, \boldsymbol{r}^{(1)} + \boldsymbol{a}_N^{(1)})$ matches that of $(\boldsymbol{r}^{(2)}, \boldsymbol{r}^{(2)} + \boldsymbol{a}_1^{(2)}, \ldots, \boldsymbol{r}^{(2)} + \boldsymbol{a}_N^{(2)})$ by shifting, $R_N(\boldsymbol{a}_1^{(1)}, \ldots, \boldsymbol{a}_N^{(1)})$ takes the same value as $R_N(\boldsymbol{a}_1^{(2)}, \ldots, \boldsymbol{a}_N^{(2)})$. Therefore, we eliminate such duplicate configurations for CHLAC features. The following section describes details of the computation of CHLAC features.

4.2. Computation

Firstly, Eq. (1) is translated from its continuous form to a corresponding discrete version:

$$R_N(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N) = \sum_{x,y,t \in D_s} f(x, y, t) f(x + a_{1x}, y + a_{1y}, t + a_{1t}) \cdots f(x + a_{Nx}, y + a_{Ny}, t + a_{Nt}), \quad (2)$$

where $a_{ix}, a_{iy} \in \{\pm\Delta r, 0\}$, $a_{it} \in \{\pm\Delta t, 0\}$ and $N \in \{0, 1, 2\}$. The parameters Δr, Δt denote the spatial and temporal intervals, respectively. The interval along the x-axis is made equal to that along the y-axis because of isotropy in the x–y plane. On the other hand, the temporal interval Δt may be different from the spatial interval Δr because the resolution of space and time may differ.

The configuration $(\boldsymbol{r}, \boldsymbol{r} + \boldsymbol{a}_1, \ldots, \boldsymbol{r} + \boldsymbol{a}_N)$ is represented by a local mask pattern, e.g., in Fig. 3. Such mask patterns are constructed as follows. There are many possible mask patterns including duplicated patterns in terms of point configurations. The mask patterns which are mutually matched by shifting, e.g., in Fig. 4, can be eliminated; in which case 279 independent mask patterns are obtained for the case of gray-scale image (Appendix A). On the other hand, in the case of binary image ($f(\boldsymbol{r}) = 0$ or 1), the number of mask patterns is further reduced to 251 because $f(\boldsymbol{r})^k = f(\boldsymbol{r})$, $\forall k > 0$. Refer to Appendix A for all of the reduced pattens. The dimension of CHLAC features corresponds to the number of mask patterns. In this paper, we use the latter 251-dimensional features since the voxel values in cubic data are binary values (Section 3).

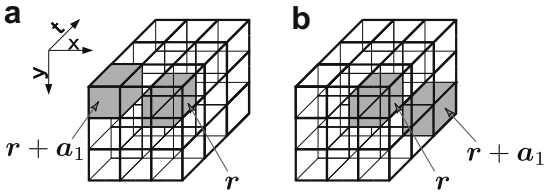

Fig. 4. Examples of duplicated mask patterns: (a) $N = 1$, $\boldsymbol{a}_1 = (-\Delta r, -\Delta r, -\Delta t)^T$, (b) $N = 1$, $\boldsymbol{a}_1 = (\Delta r, \Delta r, \Delta t)^T$. The mask pattern (a) corresponds to a shift of (b) by $(\Delta r, \Delta r, \Delta t)^T$.

In the case of motion images, frames are successively inputted to the system. Considering cubic data with constant time width, the oldest frame in the cubic data is discarded when a new frame is inputted. Similarly, the feature (vector) for the cubic data is updated only by adding CHLAC feature (vector) for the new frame and subtracting one for the oldest frame, not computing for whole cubic data every time. This is due to the "additivity" property of CHLAC as described in the next section. Consequently, the computational cost is significantly reduced. The computational cost of CHLAC is low because it consists of simple multiplications (or AND operations for binary data) and additions, which enables us to use SIMD instructions (Iwata et al., 2007). The computation for a mask pattern is skippable if its center pixel value is 0, which is common in an actual image sequence. In addition, the computation of the mask pattern of $N = 2$ is also skippable if it includes the skippable sub-pattern ($N = 1$). Therefore, we can efficiently compute CHLAC features. For example, it takes about 0.4 ms per frame to compute the CHLAC features for a 320 × 240 image sequence using a Pentium 4 3.8 GHz processor with 3 GB RAM.

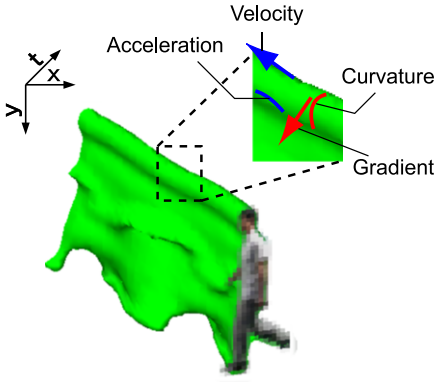

Fig. 5. Manifold of human motion in XYT space. Local geometric characteristics are also described as arrows and curves. Velocities and accelerations are along the time axis while gradients and curvatures are in an image frame.

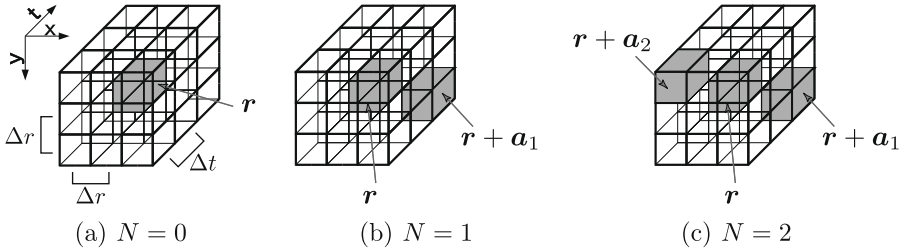

Fig. 3. Examples of independent mask patterns: (a) $N = 0$, (b) $N = 1$, $\boldsymbol{a}_1 = (\Delta r, \Delta r, \Delta t)^T$, and (c) $N = 2$, $\boldsymbol{a}_1 = (\Delta r, \Delta r, \Delta t)^T$, $\boldsymbol{a}_2 = (-\Delta r, -\Delta r, -\Delta t)^T$.

Please cite this article in press as: Kobayashi, T., Otsu, N., Three-way auto-correlation approach to motion recognition, Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.09.006

4.3. Properties

The CHLAC method extracts spatio-temporal features from three-way data in only one step. This is different from the traditional approaches which require two steps: shape feature extraction and temporal feature extraction. Furthermore, it has the following desirable properties with respect to recognition.

- *Shift-invariance* with respect to data: This renders the method *segmentation-free*.
- *Additivity* for data: Suppose we consider disjoint regions A and B ($A \cap B = \phi$); the feature value of the whole region is the sum of those from regions A and B due to the locality of auto-correlations. When different motions, whose features are denoted by $\boldsymbol{f}_A$ and $\boldsymbol{f}_B$, are simultaneously captured in an image sequence, the feature of the whole image sequence is sum of these different motions, $\boldsymbol{f}_T = \boldsymbol{f}_A + \boldsymbol{f}_B$. If we know each motion feature, $\boldsymbol{f}_A, \boldsymbol{f}_B$, through training stage, these different motions can be recognized from $\boldsymbol{f}_T$ simultaneously, such as by multiple regression analysis (Kobayashi and Otsu, 2004).
- *Robustness to noise* in data: Correlation is robust with respect to background noise as follows. Let s_i and n_i be a signal and random noise with mean 0 and variance σ^2 at the ith voxel, respectively; $\boldsymbol{E}(s_i + n_i)(s_j + n_j) = \boldsymbol{E}(s_i s_j) + \sigma^2 \delta_{ij} \fallingdotseq \boldsymbol{E} s_i s_j$ by assuming that $s_i \gg \sigma$. For binary data, isolated noise has hardly any effect on CHLAC feature values because most of the noise points have no correlation with the surrounding points.

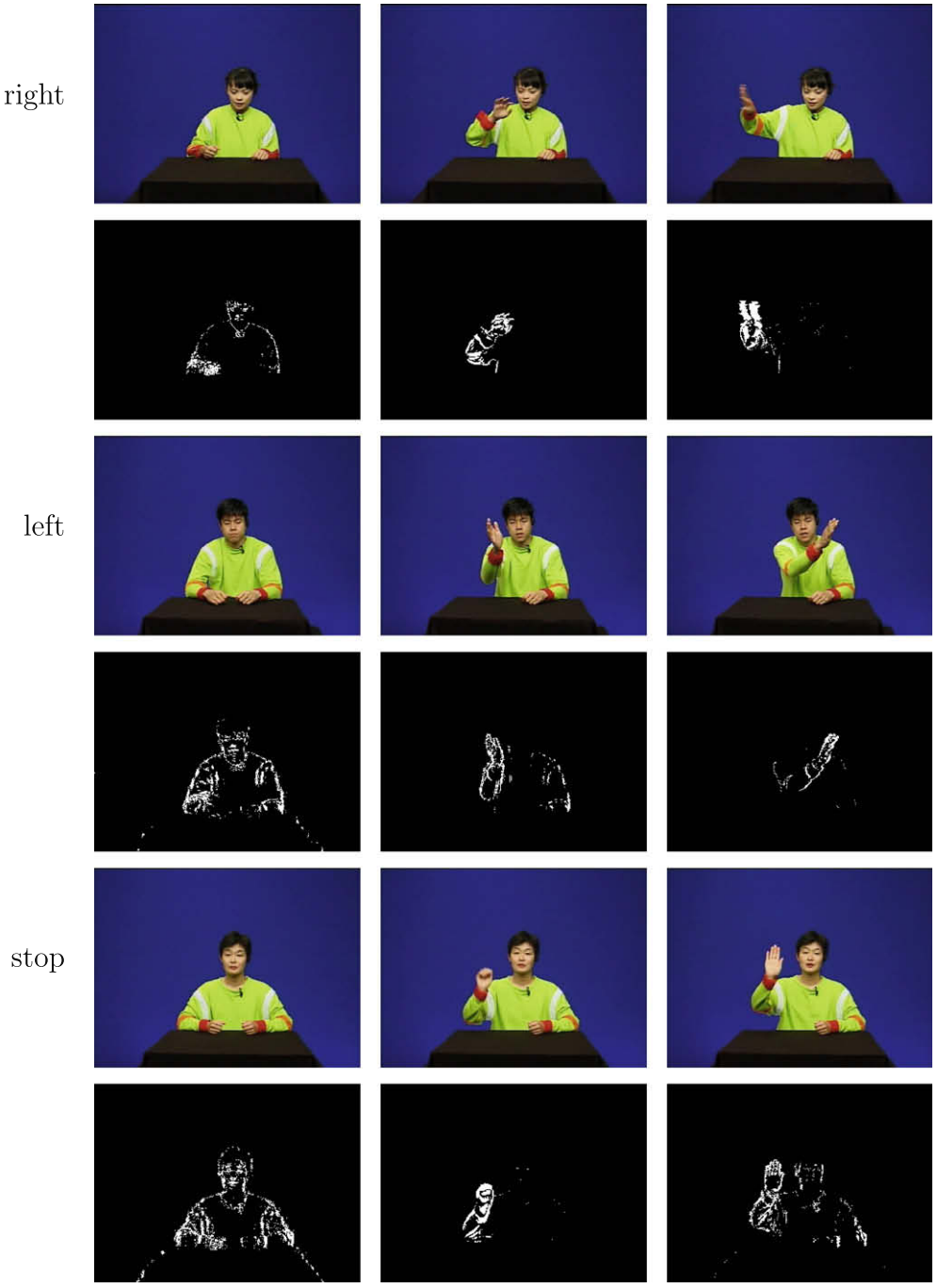

Fig. 6. Snapshots of some gestures and results of preprocessing. Top: "right", middle: "left", bottom: "stop".

Please cite this article in press as: Kobayashi, T., Otsu, N., Three-way auto-correlation approach to motion recognition, Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.09.006

In the case of binary three-way data where dot patterns form a manifold of the object, CHLAC can extract local geometric characteristics: gradients and curvatures. These two characteristics are discretized and described by the first and second order mask patterns, respectively. The first and second order patterns compose local lines and curves, respectively, and the gradients and curvatures of the manifold are approximated by such mask patterns. From this viewpoint, a CHLAC feature may be interpreted as a histogram of orientations of these local characteristics. In this study of motion images, the preprocessed frame contains dot patterns of human contours, and the manifold is formed by successive human contours (dot patterns) in three-dimensions (x,y,t) (see Fig. 5). The geometric characteristics of this manifold are spatio-temporal. In particular, along the temporal axis, gradients and curvatures correspond to velocities and accelerations of individual points which characterize the motion. Therefore, CHLAC for (binary) motion images can effectively and simultaneously capture the characteristics not only of the shape but also its motion. It should be noted that Raytchev et al. (2000) and Ishihara and Otsu (2004) used only first order mask patterns indicating gradients for simplicity, which are not sufficient for fully describing the characteristics of motion.

5. Experimental results

5.1. Gesture recognition

For gesture recognition, we used a Multimodal Database of Gestures with Speech (Hayamizu et al., 1996). This database contains time-varying image sequences of 17 gesture classes by 48 subjects, consisting of 25 women and 23 men (see Fig. 6). Each gesture class of each subject contains 4 sequences and thus the total number of sequences is 3264. The size of the image frames is 320×240.

The scheme for recognizing gestures in image sequences is based on Fisher discriminant analysis (FDA) of CHLAC features and k-NN decision rule in the discriminant space. At each time t in an image sequence, the CHLAC feature of cubic data is classified by k-NN decision, and then the whole sequence is classified by accumulating successive decision results of frames. FDA is preferable for recognition because feature vectors of each gesture class are clustered in the reduced dimensional discriminant space, and the computational cost of k-NN search is lessened.

Leave-one-out cross-validation was performed to test the proposed method, i.e., for each run the sequences of 47 subjects were exploited for training and the sequences of the remaining subject were used for test.

The results with various parameter values are shown in Fig. 7. In this experiment, there were two parameters; the time width T of cubic data and the spatial interval Δr in CHLAC (setting the temporal interval Δt to 1). The optimal set of the parameter values was found by grid search, and the partial results with respect to each parameter is shown in Fig. 7. As for the time width T, it is taken as a moving average along the trajectory of the CHLAC feature sequence. Cubic data of greater width may decrease discrimination between trajectories, while those of smaller width do not include sufficient motion information. As for the spatial interval Δr, the correlations between very close points become meaningless because points that are close together are obviously highly correlated in every direction. On the other hand, points that are distant are not correlated at all, and thus the appropriate interval Δr to obtain effective correlations will depend on the scale of the object and the movement. These trade-offs in T and Δr can be determined by the peak of recognition rate in Fig. 7, which occurred when $T = 30$, $\Delta r = 4$. It is noted that the best recognition rate of 95.86% is slightly superior to the 95.65% rate reported by Ishihara and Otsu (2004) who applied a more complicated time series analysis with HMM for the auto-regressive model.

Next, we demonstrate the effectiveness of *higher-order* auto-correlations which is primary advantage of CHLAC. For this purpose, only first order correlation, i.e., $N \leqslant 1$ in Eq. (2), is applied, and then the performance is compared with that of CHLAC of $N \leqslant 2$. For roughly making uniform the dimensionality of these two types of feature, spatial interval $\Delta\ r$ is varied from 1 to 20 and then extracted first order features are concatenated to form 261-dimensional feature vector which is comparable to 251-dimensional vector of CHLAC with $\Delta r = 4$. The recognition result of first order feature with $T = 30$ is 93.01% and it is inferior to 95.86% of CHLAC. This indicates that CHLAC of *higher-order* is superior, extracting detailed characteristics of motion as described in Section 4.3.

Finally, for preprocessing of motion images, edge extraction is compared with frame differencing. In this experiment, we applied Sobel filters to calculate gradients and then binarized them to extract edges. The recognition results from edge extraction are shown in Fig. 8 in a manner similar to that of Fig. 7. The best result from edge extraction was 94.12%, when $T = 40$, $\Delta r = 4$, and it can be obviously seen that frame differencing is superior to edge extraction. The reason is as follows. Firstly, the process of edge extraction produces contours of objects whether or not they move. The resulting frames include information that is not relevant for gesture motion, such as the static human shape, which may result in a decrease in performance. Secondly, human contour "curves" are obtained from edge extraction whereas moving "regions" are constructed by frame differencing (Fig. 9). The movements of curves may be captured by CHLAC features with constant intervals $\Delta\ r$

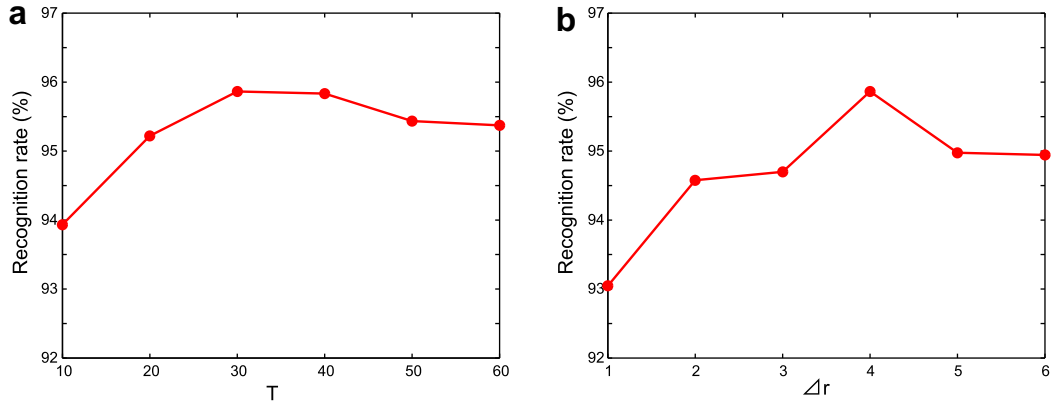

Fig. 7. Recognition results. The optimal parameter values by grid search are $T = 30$, $\Delta r = 4$. (a) Varying T from 10 to 60 with $\Delta r = 4$. (b) Varying Δr from 1 to 6 with $T = 30$.

Please cite this article in press as: Kobayashi, T., Otsu, N., Three-way auto-correlation approach to motion recognition, Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.09.006

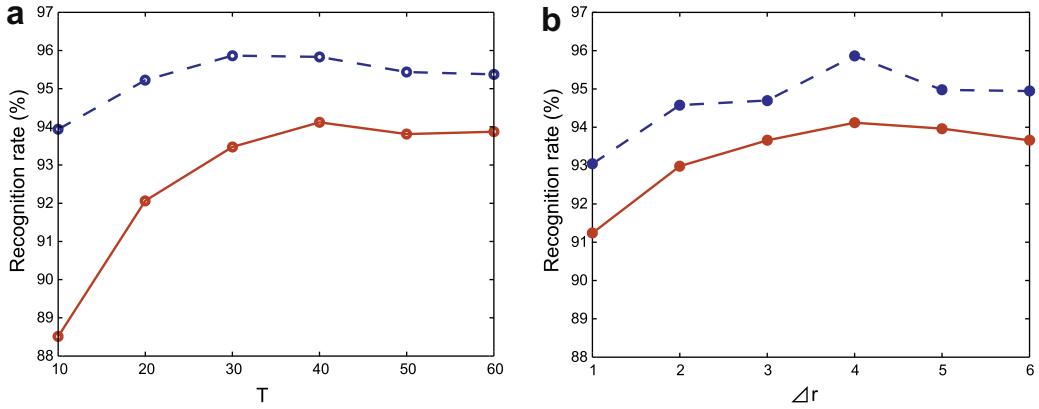

Fig. 8. Recognition results when edge extraction is used for preprocessing (solid line). The optimal parameter values by grid search are $T = 40$, $\Delta r = 4$. (a) Varying T from 10 to 60 when $\Delta r = 4$. (b) Varying Δr from 1 to 6 when $T = 40$. The dashed line is the same as that of Fig. 7 (frame differencing).

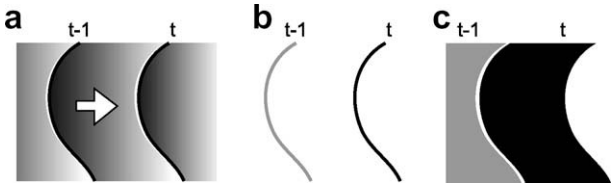

Fig. 9. Comparison of preprocessing. (a) Moving object. (b) Result of edge extraction. Only the contour edge is extracted. (c) Result of frame differencing. The moving region is extracted.

and Δt, but only certain velocities of curves, which correspond to the constant intervals, can be captured. Capturing all movements (velocities) of curves requires the use of various intervals in CHLAC. On the other hand, a moving region includes velocity information in its size. Thus, velocity information is expressed in terms of the quantity of CHLAC feature values. That is, faster movements create larger regions by frame differencing, which results in a greater number of CHLAC values of the mask patterns corresponding to the movement.

5.2. Gait recognition

Next, we applied the proposed method to gait recognition using the NIST gait dataset which is the largest one available. The original study is referred to Kobayashi and Otsu (2006). The dataset consists of 456 video sequences of 71 individuals, walking around an elliptical course (Fig. 10), with labels: Gallery for training, and Probes A–G for test. The details of this dataset are described in (Sarker et al., 2005).

The recognition scheme is almost the same as that for gesture recognition except that the parameter ranges for CHLAC are taken into account. Many different discriminant spaces are constructed for all parameter values lying in the parameter range (see Section 5.2.1), and the results for these discriminant spaces are unified as follows.

At each time t, a CHLAC feature is extracted for each parameter set, $\boldsymbol{R}_t(\Delta r, \Delta t, T)$, and k-NN decision is performed in the corresponding discriminant space $S(\Delta r, \Delta t, T)$. We repeat this k-NN decision for all discriminant spaces, and the frame t is classified by

$$\mathrm{Result}(t) = \arg\max_i \max_{\Delta r, \Delta t, T} \mathrm{kNN}_{S(\Delta r, \Delta t, T)}(\boldsymbol{R}_t(\Delta r, \Delta t, T), P_i), \tag{3}$$

$$\text{where } (\Delta t, \Delta r, T) \in \mathit{ParamRange}. \tag{4}$$

$\mathrm{kNN}_{S(\Delta r, \Delta t, T)}(\boldsymbol{x}, P_i)$ counts the number of samples belonging to the ith person P_i in the k nearest neighbors around $\boldsymbol{x}$ in the space $S(\Delta r, \Delta t, T)$. This k-NN number is some sort of the posterior likelihood of the person when using the parameter values, and by maximizing the likelihood over Δr, Δt, T and P in Eq. (3) the recognition result is more stable and accurate because the parameter values may have different discriminatory power for different people. This process also means the selection of parameters. The range of parameters in Eq. (4) are determined in the next section. Finally, the sequence is classified by accumulating the individual frame decisions.

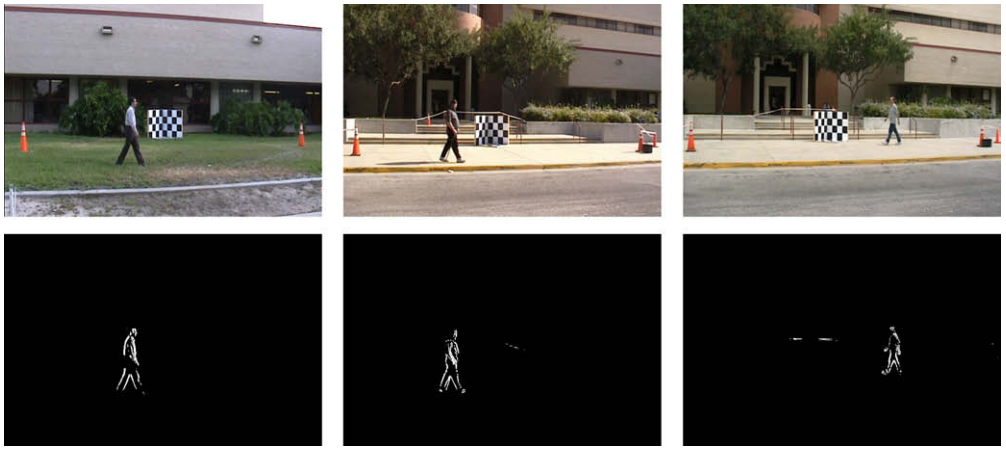

Fig. 10. Snapshots of some gait sequences and results of preprocessing.

5.2.1. The Parameter range for gait

There are three parameters to be determined: the spatial and temporal intervals Δr, Δt and the time width T. As seen in gesture recognition in Section 5.1, optimal parameter values may vary for each object and each motion, but they cannot be defined *a priori* without any knowledge. Therefore, we take into account some knowledge about characteristics of the human gait in order to restrict the ranges of these parameter values.

[***Spatial and temporal intervals*** Δr ***and*** Δt] The only constraint on Δr and Δt is locality. However, some knowledge about the human gait imposes further constraints on the relationship between Δr and Δt.

Suppose a human is walking from right to left in an image plane. If the image sequence is sliced at the midpoint of the height of the human, the sliced surface also forms an image plane (x-axis vs. t-axis), the so-called XT-slice (Niyogi and Adelson, 1994) (Fig. 11). The image illustrates that the trajectory of human walking can be approximated as a straight line, the gradient of which corresponds to the walking velocity. The relationship between spatial and temporal intervals is closely connected to this gradient (velocity). If $\arg(-\Delta r, \Delta t)^T$ is very different from the gradients of the human trajectories, it does not make any correlations of human positions in XT-slice, and the CHLAC feature values become close to zero. Therefore, it should be close to most gradients, i.e. the mean of the gradients (Fig. 11b). We adopt principal component analysis (PCA) for approximating each person's trajectory by a straight line and then estimate the gradient. The mean gradient over all sequences was computed as -0.49, which corresponds to $\Delta t/\Delta r = 1/2$. On the other hand, in the image frame, the size of the human body (width of the human figure) restricted Δr to the range $\Delta r \leqslant 16$.

[***Time width*** T] Since the experiment of gait recognition utilized noisy outdoor images, a larger time width T had the result that noisy frames were retained in the cubic data for a long time. Thus, T should be limited up to appropriate length. Considering that human gait is a periodic motion, CHLAC features of cubic data with a value for T close to the period would be particularly stable. We evaluate the stability of CHLAC features on the basis of the variance of features in each image sequence by varying the time width T under the assumption that the gait period is constant within each image sequence. Fig. 12 shows the variance vs. time width T. Since the variance became stationary at a time width of around 30, we set the average gait period to be 30 frames. In (Sundaresan et al., 2003; Sarker et al., 2005), it was reported that the gait period was also set at 30–40 frames. Here, it is important that the gait period is calculated based on the stability of features, i.e. the variance, without applying an object-model based analysis used in other studies: for example, making use of leg-angles. Thus, the time width $T \leqslant 30$.

As a result, the parameter ranges in Eq. (4) are determined as

$$ParamRange = \{\Delta r, \Delta t, T | \Delta t/\ \Delta r = 1/2,\ \Delta r \leqslant 16,\ T \leqslant 30\}. \qquad (5)$$

The discriminant spaces are constructed for every parameter value satisfying the constraint in Eq. (5), and then k-NN decisions are performed in these spaces. These constraints are not particularly heuristic because they are definitely derived from the data by introducing a little knowledge about the human gait. These constraints make it possible to extract CHLAC features more effectively for the human gait, and by combining these constraints with the decision rules in Eq. (3) our scheme becomes much more efficient.

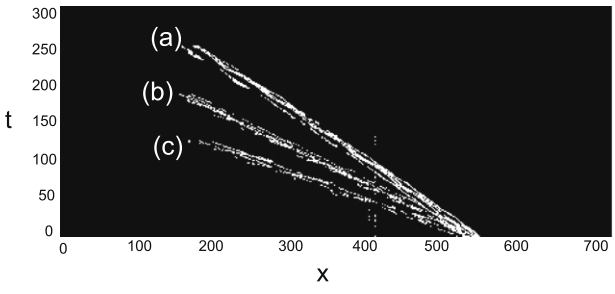

Fig. 11. Trajectories of different walking humans in an XT-slice. (a) Slowest walk, (b) middle speed walk and (c) fastest walk.

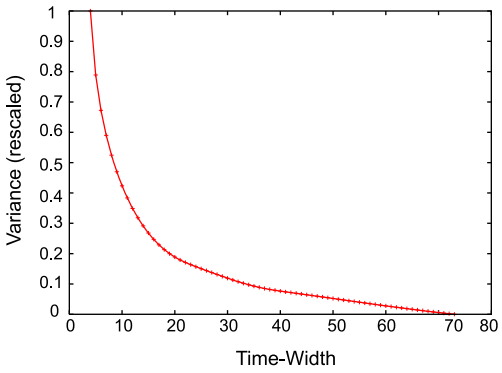

Fig. 12. Variance vs. time width. The variance is unbiased, rescaled to [0,1] and averaged over all image sequences. It becomes stationary at a time width of around 30.

5.2.2. Recognition results

The identification results as compared to those resulting from the use of other methods (Sundaresan et al., 2003; Sarker et al., 2005; Tolliver and Collins, 2003; Lee et al., 2003; Wang et al., 2003) are shown in Fig. 13. The identification rate of our method is also presented in Table 1a. The results show that our scheme outperforms the others in all probes. It is noted that the identification results of Probes D–G are worse than those of Probes A–C for all methods. This is caused by differences in ground surface conditions: The Gallery and Probes A–C are "grass," while Probes D–G are "concrete." The surfaces may slightly affect the gait period and the preprocessing results. Thus, Probes D–G whose surfaces are different from that of the Gallery, pose challenging problems, but the performance of our method is nevertheless much better than any other methods even for these probes. This is because CHLAC is robust with respect to the results of preprocessing as discussed next.

Table 1 shows our method's dependency on the quality of preprocessed data: namely the effect of noise in background and in the human region. The term "bbox" means that the human region (bounding box) is extracted and pixels in the other region are set to 0 (noiseless) after binarization to suppress the amount of background noise. The term "half-threshold" means binarization with

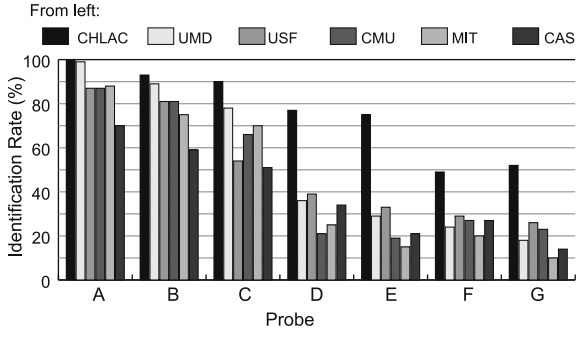

Fig. 13. The identification rate (%) for each probe compared with those of the other methods: UMD (Sundaresan et al., 2003), USF (Sarker et al., 2005), CMU (Tolliver and Collins, 2003), MIT (Lee et al., 2003), CAS (Wang et al., 2003) (these are top-rank results). See each paper for the detailed identification rate, or (Sarker et al., 2005) for the collective results. Detailed results for the proposed method (CHLAC) are shown in Table 1a.

Table 1
The identification rate (%) for the proposed method under various conditions. Details are in the text.

Probe	Half-threshold		Automatic-threshold
	bbox		Non-bbox
A	100	100	99
B	93	90	90
C	90	90	83
D	77	67	61
E	75	70	61
F	49	39	40
G	52	45	45
	(a)	(b)	(c)

the half value suggested by automatic-thresholding for increasing the amount of noise and the thickness of a human contour. Comparing Table 1b and c, the proposed method is slightly affected by background noise, but from (a) and (b) it can be seen that information about a human contour is more helpful than noise. Thus, CHLAC can cope with noisy data (Section 4.3), which makes it effective for use outdoors.

6. Conclusion

We have presented a feature extraction method, cubic higher-order local auto-correlation (CHLAC), for three-way data, particularly for motion images. The method is based on three-dimensional (spatio-temporal) auto-correlations of pixels, which are closely related to geometric meanings: namely gradients and curvatures. Particularly, for motion images, it extracts not only the gradients and the curvatures of shape but also the velocity and the acceleration of the motion simultaneously. Thus, neither a specific model of the objects nor time series analysis is required, unlike traditional approaches to motion analysis. It is also noteworthy that CHLAC is robust with respect to noise in data and is a *segmentation-free* method. In addition, its computational cost is so low that the method can be applied in real time.

The two experimental results for gesture and gait recognition showed the effectiveness of the CHLAC method for motion analysis. The objects to be recognized are different in these experiments; one is concerned with the motion itself regardless of performers and the other is concerned with the performer of the motion via the gait. The CHLAC feature extraction method can be applied to such different tasks by combining it with a subsequent multivariate analysis suited to the particular tasks, e.g. discriminant analysis in this paper.

It should be remarked that the CHLAC method is so general as to be applicable to three-dimensional geometrical analysis, such as object recognition.

Appendix A. CHLAC mask patterns

All mask patterns are shown in Table A.1. M_{-1}, M_0, M_{+1} indicate layers in a mask pattern (Fig. A.1a) and a–i are positions in each layer (Fig. A.1b). The mask pattern of $N = 0$ is only No. 1, and those of $N = 1$ are Nos. 2–14, together with No. 252. The others are of $N = 2$.

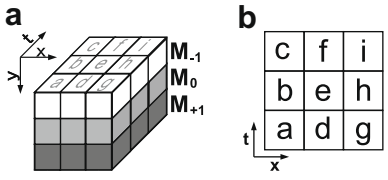

Fig. A.1. Mask layers (a) and position labels in each layer (b).

Table A.1
CHLAC mask patterns

No.	Mask pattern		
	M_{-1}	M_0	M_{+1}
1	×	e	×
2	×	e	a
3	×	e	b
4	×	e	c
5	×	e	d
6	×	e	e
7	×	e	f
8	×	e	g
9	×	e	h
10	×	e	i
11	×	e,a	×
12	×	e,b	×
13	×	e,c	×
14	×	e,d	×
15	a,b	e	×
16	a,c	e	×
17	a,d	e	×
18	a,e	e	×
19	a,f	e	×
20	a,g	e	×
21	a,h	e	×
22	a,i	e	×
23	a	a,e	×
24	a	b,e	×
25	a	c,e	×
26	a	d,e	×
27	a	e,f	×
28	a	e,g	×
29	a	e,h	×
30	a	e,i	×
31	a	e	a
32	a	e	b
33	a	e	c
34	a	e	d
35	a	e	e
36	a	e	f
37	a	e	g
38	a	e	h
39	a	e	i
40	b,c	e	×
41	b,d	e	×
42	b,e	e	×
43	b,f	e	×
44	b,g	e	×
45	b,h	e	×
46	b,i	e	×
47	b	a,e	×
48	b	b,e	×
49	b	c,e	×
50	b	d,e	×
51	b	e,g	×
52	b	e,h	×
53	b	e,i	×
54	b	e	a
55	b	e	b
56	b	e	c
57	b	e	d
58	b	e	e
59	b	e	f
60	b	e	g
61	b	e	h
62	b	e	i
63	c,d	e	×
64	c,e	e	×
65	c,f	e	×
66	c,g	e	×
67	c,h	e	×
68	c,i	e	×
69	c	a,e	×
70	c	b,e	×
71	c	c,e	×
72	c	d,e	×
73	c	e,g	×
74	c	e,h	×

Please cite this article in press as: Kobayashi, T., Otsu, N., Three-way auto-correlation approach to motion recognition, Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.09.006

Table A.1 (*continued*)

No.	Mask pattern		
	M_{-1}	M_0	M_{+1}
75	c	e,i	×
76	c	e	a
77	c	e	b
78	c	e	c
79	c	e	d
80	c	e	e
81	c	e	f
82	c	e	g
83	c	e	h
84	c	e	i
85	d,e	e	×
86	d,f	e	×
87	d,g	e	×
88	d,h	e	×
89	d,i	e	×
90	d	a,e	×
91	d	b,e	×
92	d	c,e	×
93	d	d,e	×
94	d	e,f	×
95	d	e,i	×
96	d	e	a
97	d	e	b
98	d	e	c
99	d	e	d
100	d	e	e
101	d	e	f
102	d	e	g
103	d	e	h
104	d	e	i
105	e,f	e	×
106	e,g	e	×
107	e,h	e	×
108	e,i	e	×
109	e	a,e	×
110	e	b,e	×
111	e	c,e	×
112	e	d,e	×
113	e	e	a
114	e	e	b
115	e	e	c
116	e	e	d
117	e	e	e
118	e	e	f
119	e	e	g
120	e	e	h
121	e	e	i
122	f,g	e	×
123	f,h	e	×
124	f,i	e	×
125	f	a,e	×
126	f	b,e	×
127	f	c,e	×
128	f	d,e	×
129	f	e,g	×
130	f	e	a
131	f	e	b
132	f	e	c
133	f	e	d
134	f	e	e
135	f	e	f
136	f	e	g
137	f	e	h
138	f	e	i
139	g,h	e	×
140	g,i	e	×
141	g	a,e	×
142	g	b,e	×
143	g	c,e	×
144	g	d,e	×
145	g	e,f	×
146	g	e,i	×
147	g	e	a
148	g	e	b
149	g	e	c

Table A.1 (*continued*)

No.	Mask pattern		
	M_{-1}	M_0	M_{+1}
150	g	e	d
151	g	e	e
152	g	e	f
153	g	e	g
154	g	e	h
155	g	e	i
156	h,i	e	×
157	h	a,e	×
158	h	b,e	×
159	h	c,e	×
160	h	d,e	×
161	h	e	a
162	h	e	b
163	h	e	c
164	h	e	d
165	h	e	e
166	h	e	f
167	h	e	g
168	h	e	h
169	h	e	i
170	i	a,e	×
171	i	b,e	×
172	i	c,e	×
173	i	d,e	×
174	i	e,g	×
175	i	e	a
176	i	e	b
177	i	e	c
178	i	e	d
179	i	e	e
180	i	e	f
181	i	e	g
182	i	e	h
183	i	e	i
184	×	a,b,e	×
185	×	a,c,e	×
186	×	a,d,e	×
187	×	a,e,f	×
188	×	a,e,g	×
189	×	a,e,h	×
190	×	a,e,i	×
191	×	a,e	c
192	×	a,e	f
193	×	a,e	g
194	×	a,e	h
195	×	a,e	i
196	×	b,c,e	×
197	×	b,d,e	×
198	×	b,e,g	×
199	×	b,e,h	×
200	×	b,e,i	×
201	×	b,e	g
202	×	b,e	h
203	×	b,e	i
204	×	c,d,e	×
205	×	c,e,g	×
206	×	c,e,h	×
207	×	c,e,i	×
208	×	c,e	a
209	×	c,e	d
210	×	c,e	g
211	×	c,e	h
212	×	c,e	i
213	×	d,e,f	×
214	×	d,e,i	×
215	×	d,e	c
216	×	d,e	f
217	×	d,e	i
218	×	e,f,g	×
219	×	e,f	a
220	×	e,f	d
221	×	e,f	g
222	×	e,g,i	×
223	×	e,g	a
224	×	e,g	b

(*continued on next page*)

Please cite this article in press as: Kobayashi, T., Otsu, N., Three-way auto-correlation approach to motion recognition, Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.09.006

Table A.1 (*continued*)

No.	Mask pattern		
	M_{-1}	M_0	M_{+1}
225	×	e,g	c
226	×	e,g	f
227	×	e,g	i
228	×	e,h	a
229	×	e,h	b
230	×	e,h	c
231	×	e,i	a
232	×	e,i	b
233	×	e,i	c
234	×	e,i	d
235	×	e,i	g
236	×	e	a,c
237	×	e	a,f
238	×	e	a,g
239	×	e	a,h
240	×	e	a,i
241	×	e	b,g
242	×	e	b,h
243	×	e	b,i
244	×	e	c,d
245	×	e	c,g
246	×	e	c,h
247	×	e	c,i
248	×	e	d,f
249	×	e	d,i
250	×	e	f,g
251	×	e	g,i
252	×	e,e	×
253	×	e,e,e	×
254	a	e,e	×
255	b	e,e	×
256	c	e,e	×
257	d	e,e	×
258	e	e,e	×
259	f	e,e	×
260	g	e,e	×
261	h	e,e	×
262	i	e,e	×
263	×	e,e,a	×
264	×	e,e,b	×
265	×	e,e,c	×
266	×	e,e,d	×
267	×	e,e	i
268	×	e,e	h
269	×	e,e	g
270	×	e,e	f
271	×	e,e	e
272	×	e,e	d
273	×	e,e	c
274	×	e,e	b
275	×	e,e	a
276	×	e,e,i	×
277	×	e,e,h	×
278	×	e,e,g	×
279	×	e,e,f	×

All 279 patterns are for real valued data and, especially, No. 1–251 patterns are for binary data, i.e. the patterns used in this study.

References

Cutting, J., Kozlowski, L., 1977. Recognizing friends by their walk: Gait perception without familiarity cues. Bull. Psychonomic Soc. 9, 353–356.

Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features. In: VS-PETS, pp. 65–72.

Hayamizu, S., Hasegawa, O., Itou, K., Sakaue, K., Tanaka, K., Nagaya, S., Nakazawa, M., Endoh, T., Togawa, F., Sakamoto, K., Yamamoto, K., 1996. RWC multimodal database for interactions by integration of spoken language and visual information. In: Internat. Conf. on Spoken Language Processing, pp. 2171–2174.

Ishihara, T., Otsu, N., 2004. Gesture recognition using auto-regressive coefficients of higher-order local auto-correlation features. In: Internat. Conf. on Automatic Face and Gesture Recognition, pp. 583–588.

Iwata, K., Satoh, Y., Kobayashi, T., Yoda, I., Otsu, N., 2007. Application of the unusual motion detection using CHLAC to the video surveillance. In: Internat. Conf. on Neural Information Processing.

Jhuang, H., Serre, T., Wolf, L., Poggio, T., 2007. A biologically inspired system for action recognition. In: Internat. Conf. on Computer Vision.

Johansson, G., 1973. Visual perception of biological motion and a model for its analysis. Percept. Psychophys. 14, 201–211.

Kim, T.K., Wong, S.F., Cipolla, R., 2007. Tensor canonical correlation analysis for action classification. In: Internat. Conf. on Computer Vision and Pattern Recognition, pp. 1–8.

Kobayashi, T., Otsu, N., 2004. Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation. In: Internat. Conf. on Pattern Recognition, pp. 741–744.

Kobayashi, T., Otsu, N., 2006. A three-way auto-correlation based approach to human identification by gait. In: IEEE Workshop on Visual Surveillance, pp. 185–192.

Laptev, I., 2005. On space-time interest points. Internat. J. Comput. Vision 64, 107–123.

Lee, L., Dalley, G., Tieu, K., 2003. Learning pedestrian models for silhouette refinement. In: Internat. Conf. on Computer Vision, pp. 663–670.

Nixon, M., Carter, J., 2006. Automatic recognition by gait. Proc. IEEE 94, 2013–2024.

Niyogi, S., Adelson, E., 1994. Analyzing and recognizing walking figures in xyt. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 469–474.

Otsu, N., 1979. A threshold selection method from gray-level histograms. Systems Man Cybernet. 9, 62–82.

Otsu, N., Kurita, T., 1988. A new scheme for practical flexible and intelligent vision systems. In: IAPR Workshop on Computer Vision, pp. 431–435.

Raytchev, B., Hasegawa, O., Otsu, N., 2000. User-independent online gesture recognition by relative motion extraction. Pattern Recognition Lett. 21, 69–82.

Sarker, S., Philips, P., Liu, Z., Vega, I.R., Grother, P., Bowyer, K., 2005. The humanID gait challenge problem: Data sets, performance, and analysis. Pattern Anal. Machine Intell. 27, 162–177.

Sivic, J., Zisserman, A., 2003. Video google: A text retrieval approach to object matching in videos. In: Internat. Conf. on Computer Vision, pp. 1470–1477.

Sundaresan, A., Chowdhury, A., Chellappa, R., 2003. A hidden markov model based framework for recognition of humans from gait sequences. In: Internat. Conf. on Image Processing, pp. 93–96.

Tolliver, D., Collins, T., 2003. Gait shape estimation for identification. In: Internat. Conf. on Audio- and Video-based Biometric Person Authentication, pp. 734–742.

Veres, G., Gordon, L., Carter, J., Nixon, M., 2004. What image information is important in silhouette-based gait recognition? In: Internat. Conf. on Computer Vision and Pattern Recognition, pp. 776–782.

Wang, L., Tan, T., Ning, H., Hu, W., 2003. Silhouette analysis-based gait recognition for human identification. Pattern Anal. Machine Intell. 25, 1505–1518.

Wilson, A.D., Bobick, A.F., 1999. Parametric hidden markov models for gesture recognition. Pattern Anal. Machine Intell. 21, 884–900.

Please cite this article in press as: Kobayashi, T., Otsu, N., Three-way auto-correlation approach to motion recognition, Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.09.006