Action and Simultaneous Multiple-Person Identification Using Cubic Higher-order Local Auto-Correlation

Takumi Kobayashi†
†Graduate School of
Information Science and Technology
University of Tokyo
7-3-1, Hongo, Bunkyo-ku , Tokyo, Japan
takumi@isi.imi.i.u-tokyo.ac.jp

Nobuyuki Otsu†‡
‡National Institute of
Advanced Industrial Science and Technology
(AIST)
1-1-1, Umezono, Tsukuba, Ibaraki, Japan
otsu.n@aist.go.jp

Abstract

We propose a new method – Cubic Higher-order Local Auto-Correlation (CHLAC) – to address three-way data analysis. This method is a natural extension of Higher-order Local Auto-Correlation (HLAC) [6], which deals only with two-way data. Both methods use "correlation" to summarize relative positions or motions within a local data region, and these can be calculated simply with a low computational load. Moreover, our new method (CHLAC) offers several preferable properties as well as HLAC: shift-invariance to data (rendering the method segmentation-free), additivity for data, and robustness to noise in data. In this study, we applied this method to action and simultaneous multiple-person identification from a motion-image sequence through the property of data additivity. Experimental results showed that this method performed well.

1. Introduction

Recently motion recognition in the real world, especially gait recognition, has received increased attention because of greater concerns regarding security. Security cameras are located in more and more places to monitor the actions of people, search for suspicious persons, and identify people at entrance gates. Gait recognition - identifying a person's gait from a motion-image sequence - is one form of biometric identification. In real-world applications, many persons typically appear in a motion-image sequence, and the ability to identify all of these persons at once is desirable.

Much research effort has been done on gait recognition, and almost all studies have focused on features of the human body - e.g., angles of human joints [2], human body part lengths [3], moments [8], or boundary coordinates [7] in a human body silhouette. Such heuristic approaches are based on the following premise. Persons appearing in an image must be traced, and whether an object in an image is actually a human must be determined by checking the combination of the parts. Furthermore, with such approach no more than one person can be identified at a time, so when more than one person appear in an image sequential searches for each of the individuals are needed. However, such procedural processing is complicated (with switching of the process) and is considered unstable.

Figure 1. Point-light display in Biological Motion by Johansson [4].

In this paper, we propose Cubic Higher-order Local Auto-Correlation (CHLAC) which is based on *correlation*. The idea of *correlation* is considered to be associated with *relative motion* which is seen in Biological Motion (**Fig**. 1) in [4]. In [4], a human was equipped with point-lights, and then that person's movement was observed in darkness. In other words, all point-light motion is organized spontaneously into the percept of a moving figure. This phenomenon is explained as a factor of common fate in Gestalt psychology. In this case it is considered that motion recognition requires the knowledge not of the body structure, but of the relative motion of each body part. By taking into account *correlation* or *relative motion*, we can construct such an approach that can process motion-image sequences in the same way regardless of whether the object is human.

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

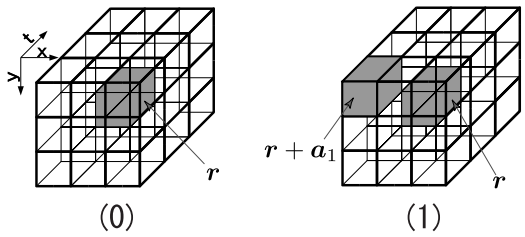

Figure 2. Example of a mask pattern: (0) $N = 0$; (1) $N = 1, \boldsymbol{a}_1 = (-1, -1, -1)^T$. Overlapping mask patterns are eliminated if mutually shifted in three-way data.

2. Cubic Higher-order Local Auto-correlation (CHLAC)

Higher-order Local Auto-Correlation (HLAC) was proposed for extracting "spatial correlation", and has been shown to work efficiently in image recognition [6]. We extend this naturally to Cubic Higher-order Local Auto-Correlation (CHLAC) to deal directly with three-way data, especially a motion-image sequence.

Letting $f(\boldsymbol{r})$ be three-way data with $\boldsymbol{r} = (x, y, z)^t$, the N-th order auto-correlation function is defined as

$$x_N(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N) = \int f(\boldsymbol{r}) f(\boldsymbol{r} + \boldsymbol{a}_1) \cdots f(\boldsymbol{r} + \boldsymbol{a}_N) d\boldsymbol{r} \tag{1}$$

where $\boldsymbol{a}_i$ $(i = 1, \cdots, N)$ are displacement vectors between $\boldsymbol{r}$ and the positions to correlate with $\boldsymbol{r}$. Eq.(1) takes many different forms if we vary N and $\boldsymbol{a}_i$, but we limit $\boldsymbol{a}_i$ to a local region $3 \times 3 \times 3$ around $\boldsymbol{r}$ and N to less than or equal to 2. Note that the $3 \times 3 \times 3$ region is not necessarily limited to adjacent positions and could be taken sparsely in a local region around the reference point $\boldsymbol{r}$. A feature extracted by CHLAC corresponds to a value of $x_N(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N)$, which is determined by one set of $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N$. However, the value remains the same if the patterns of $(\boldsymbol{r}, \boldsymbol{a}_1, \cdots, \boldsymbol{a}_N)$ are identical in the point configuration because of a uniform shift. Therefore, in the CHLAC features, we eliminate such duplicate sets. In the computation, we construct mask patterns such as **Fig**. 2. These mask patterns indicate $\boldsymbol{r}$ and $\boldsymbol{a}_i$. We first multiply the values of the gray positions in the mask pattern and then sum up the resulting value for the whole region as three-way data by shifting the mask pattern. There are 279 mask patterns, and this number is reduced to 251 for binary data (whose value is 0 or 1). In this study, we converted motion-image sequences to binary data, therefore 251 mask patterns were used.

This CHLAC method extracts features from data in just one step. We can easily calculate this step because it consists only of multiplication and addition, so this is a very simple method. Furthermore, it has three preferable properties for recognition purposes.

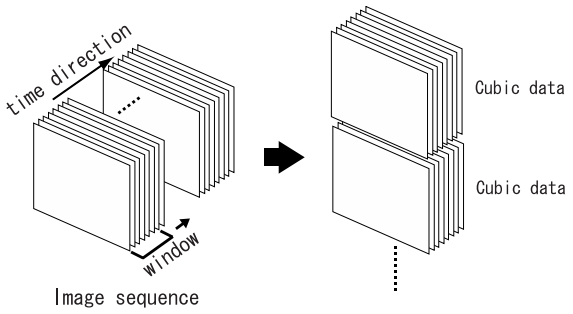

Figure 3. Cubic data

- *Additivity* in feature values because of $\boldsymbol{a}_i$ being limited to a local region and integration.
- *Shift-invariance* to data because of integration.
- *Robustness to noise* in data because correlation is robust to additive noise.

The property of *additivity* makes it possible to identify multiple persons simultaneously, and segmentation of the human figure is not necessary because of *shift-invariance*.

3. Recognition

For a motion-image sequence, we need some preprocessing before applying CHLAC feature extraction.

First, a motion-image sequence, which consists of three-way data along the x-axis and y-axis of an image frame and along the time-axis in the frame-sequence direction, is separated into many smaller units of three-way data which we call "cubic data" (**Fig**. 3). That is, by setting a time-window along the time-axis, we can regard the frames within the time-window as cubic data. Many cubic data are generated by shifting the time-window one frame at a time. Recognition can then be carried out at each time by processing the corresponding cubic data.

For recognition, motion pixels are extracted by subtracting a frame at time $t-1$ from a frame at time t and applying binarization [5] (**Fig**. 4). As a result, we obtain pixel values 1 or 0 which respectively indicate "moved " or "static".

Next, CHLAC is applied to the cubic data created through the above process. This corresponds to extracting motion correlation as a feature of human motion because CHLAC is based on the correlation of pixels in cubic data which indicate motion. In cubic or spatio-temporal correlation by CHLAC, spatial correlation is used to extract a moving human *shape* and temporal correlation is used to extract a human *motion*. In our results, we obtained a 251-dimensional feature vector from each cubic data.

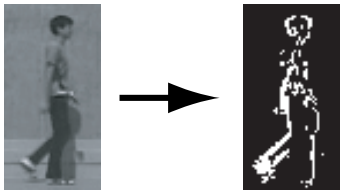

Figure 4. Extraction of motion pixels

Table 1. Results of action recognition for various time-widths, i.e., time-window widths.

time-width	10	20	30	40	50
recognition rate (%)	99.4	99.9	99.9	99.9	99.9

Table 2. Result of person identification for various action which indicates the key of person identification.

action	all	right walk	left walk	right run	left run
identification rate (%)	99.8	99.4	99.9	95.5	98.1

4. Experimental results

We have applied this method for action and simultaneous multiple-person identification from a motion-image sequence. Motion-image sequences of five persons walking and running left, right in frontal-parallel directions were captured. That is, for each person, there were four kinds of action: *leftward walk*, *leftward run*, *rightward walk*, and *rightward run*. These data were captured at 30 frames per second (fps) from digital video tape; they were encoded to MPEG images of size 352×240. The size of each person's figure in the image was about 30×80. We obtained about 2000 cubic data (**Fig**. 3). It takes 700ms to calculate the CHLAC feature associated with one frame (using Pentium IV 2.2GHz with no special optimization).

First, in a basic experiment on action recognition, there were four resulting classes (categories): rightward walk, leftward walk, rightward run, and leftward run. One-third of cubic data in each action were used for training data and the others were test data. For recognition, we applied Fisher discriminant analysis to the data. Each input data was classified into the class whose mean was nearest to the input data. Here we regarded the mean of each class as being the representation of an action. The results are shown in **Table** 1 with different time-widths.

Next, in a basic experiment on gait or person identification (five persons), we used the same process as above in selecting test and training data and in discrimination. The results are shown in **Table** 2, where identification was done

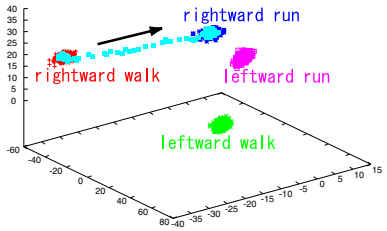

Figure 5. Fisher discriminant space and the transition from rightward walk to rightward run.

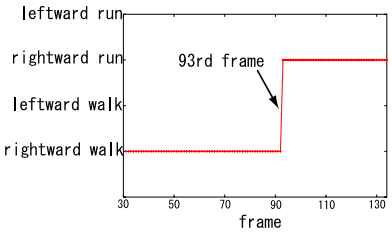

Figure 6. Result of action identification

respectively for each action and for all. The width of the time-window was fixed to 30 frames.

In the following sections, we discuss the results of more realistic experiments, where the width of the time-window was again fixed to 30 frames.

4.1. Action identification

In the real world, human action is variable rather than constant, so we need to recognize the transition from one action to another. In this experiment, we obtained 134 frames test data by uniting the rightward walk data and rightward run data and the transition point was the 77th frame; i.e., the human action changed from a rightward walk to a rightward run at the 77th frame. The transition of the action is clearly shown as the straight line between two action categories in the discriminant space (**Fig**. 5), and appeared at the 93rd frame (**Fig**. 6) since there was a 16-frame delay due to the width of the time-window in the cubic data. Initially, the cubic data consisted of only the frames of rightward walk, and then the ratio of the frames of rightward run increased once the transition occurred. This is described as follows due to the additivity of CHLAC features.

$$\boldsymbol{x} = (1-\alpha)\boldsymbol{m}_w + \alpha \boldsymbol{m}_r \qquad (2)$$

where $\boldsymbol{x}$ is the vector including the transition, $\boldsymbol{m}_w$ and $\boldsymbol{m}_r$ are the mean vectors of rightward walk and run respectively, and α is the ratio of the frames of rightward run to the total frames in the cubic data. Eq.(2) is a vector equation of a straight line, thus the transition was represented as a straight line as actually seen in **Fig**. 5.

4.2. Simultaneous multiple-person identification

It is rare that only one person appears in a camera image, and multiple persons are usually captured. In that case, most methods require that a search must be done for each person for individual identification. That means at least two processes, tracking and identification, are needed. However, more processes usually lead to greater system instability. Thus, a method that can identify multiple persons at the same time is desirable, and CHLAC makes this possible. In this experiment, we created test data by combining the second person data (rightward walk) and the third person data (leftward run) in which occlusion is not included. We used all types of action data: rightward walk, leftward walk, rightward run and leftward run, for the person identification.

Because of the property of additivity and shift-invariance, the CHLAC feature of multiple persons can be factorized as

$$\boldsymbol{x} = \alpha_1 \boldsymbol{f}_1 + \alpha_2 \boldsymbol{f}_2 + \alpha_3 \boldsymbol{f}_3 + \alpha_4 \boldsymbol{f}_4 + \alpha_5 \boldsymbol{f}_5 + \boldsymbol{\epsilon} \qquad (3)$$

where $\boldsymbol{x}$ is a CHLAC feature vector of multiple persons, $\boldsymbol{f}_i$ $(i = 1, \cdots, 5)$ is a CHLAC feature mean vector of the ith person, α_i is 1 or 0 indicating whether or not the ith person is shown in the image, and $\boldsymbol{\epsilon}$ is a residual error vector. Then, the least-square solution of Eq.(3) is given by

$$(\alpha_1 \cdots \alpha_5)^t = (\boldsymbol{F}^t \boldsymbol{F})^{-1} \boldsymbol{F}^t \boldsymbol{x} \qquad (4)$$

where $\boldsymbol{F}$ is a matrix $[\boldsymbol{f}_1 \cdots \boldsymbol{f}_5]$. In Eq.(4), the only condition is that $\boldsymbol{F}$ has full column rank 5, which is usually satisfied for original 251-dimensional vector $\boldsymbol{x}$. In the identification, the values of $\alpha_1, \cdots, \alpha_5$ are rounded off, so that multiple persons can be identified at the same time. We computed the identification rate as the ratio of the number of correct frames to the number of total frames. The result was *0%* for the original vector. This result implies that the feature vectors are not well clustered in original CHLAC feature space. Thus, it is suggested to use Fisher discriminant analysis to improve clustering. However, its ordinary use results in 4-dimensional mapped vectors and does not satisfy the full-rank condition in Eq.(4). Therefore, to avoid this problem by increasing the dimensionality, we extended Fisher discriminant analysis, employing also the eigenvectors associated with the eigenvalue of zero. The experimental result was *100%* for the extended Fisher vector, since the feature vectors were completely clustered in the extended Fisher discriminant space (**Fig**. 7).

In reality, occlusion of persons may lead to identification error in the very moment of the occlusion, however there are several methods to cope with the problem, such as utilization of histories.

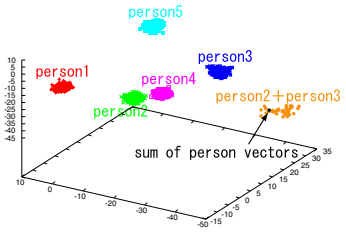

Figure 7. Three-dimensional subspace within the extended Fisher discriminant space

5. Conclusion

We have developed a new method - Cubic Higher-order Local Auto-Correlation (CHLAC) - as a natural extension of HLAC [6]. We have shown here that the CHLAC method is not so dependent on the time-window width and is effective even for realistic problems: action and simultaneous multiple-person identification. Of particular importance, this method makes it possible to identify multiple persons at the same time, which can hardly be done with previously reported methods. CHLAC is, however, sensitive to scale variance, but there is possibility to make CHLAC scale-invariance along the lines of [1]. The CHLAC method is so general that it is widely applicable to various other kinds of three-way data analysis and recognition in addition to that of motion-image sequences.

References

[1] S. Akaho. Translation, scale and rotation invariant features based on higher-order autocorrelations. *Bulletin of the Electrotechnical Laboratory*, 57(10):973–981, 1993.

[2] A. Ali and J. Aggarwal. Segmentation and recognition of continuous human activity. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 28–35, 2001.

[3] A. Bobick and A.Y.Johnson. Gait recognition using static,activity-specific parameters. In *IEEE Computer Vision and Pattern Recognition Conference*, pages 423–435, 2001.

[4] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[5] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. on System, Man and Cybernetics*, 9(1):62–66, 1979.

[6] N. Otsu and T. Kurita. A new scheme for practical flexible and intelligent vision systems. In *IAPR Workshop on Computer Vision*, pages 431–435, 1988.

[7] E. Tassone, G. West, and S. Venkatesh. Temporal pdms for gait classification. In *International Conference on Pattern Recognition*, pages II: 1065–1068, 2002.

[8] L. Wang, W. Hu, and T. Tan. A new attempt to gait-based human identification. In *International Conference on Pattern Recognition*, pages I: 423–430, 2002.