A Three-Way Auto-Correlation Based Approach to Human Identification by Gait

Takumi Kobayashi Nobuyuki Otsu
National Institute of Advanced Industrial Science and Technology (AIST)
{takumi.kobayashi, otsu.n}@aist.go.jp

Abstract

We propose a scheme for gait recognition using cubic higher-order local auto-correlation (CHLAC), discriminant analysis, and k-NN decision rules. CHLAC is based on three-way (x-, y-, and time-dimensional) auto-correlations of pixels in motion images, and it effectively extracts motion features. The method has several properties preferable for recognition: shift-invariance (rendering the method segmentation-free) and robustness to noise in data. Moreover, the method is so general as to use neither a priori knowledge nor heuristics about objects such as human shapes and is applicable to any three-way data. We made the scheme more effective for gait recognition by introducing some knowledge of gait to optimise parameters in CHLAC. Our scheme was applied to the NIST gait dataset for human identification, and the result was compared to those of other methods. Our scheme outperformed the others in spite of the simple feature extraction and the simple classification rule.

1. Introduction

Motion recognition is becoming an important area in computer vision. In particular, human motion, such as gait, is expected to be a key to human identification [1]. Unlike fingerprinting, this biometric method can identify humans by observing gaits through video cameras at a distance. However, in motion recognition, especially in gait recognition, some difficult problems must be treated: segmenting, tracking, and analysing both the human shape and its changes in a time series. Much effort has so far been made to solve these problems.

Motion images have spatial and temporal information that is difficult to effectively handle all together. In usual approaches to motion image analysis, these two kinds of information are processed individually: first, each image frame is processed and usually compressed to a feature vector, and then the time series of the obtained feature vectors is analysed. For example, recent approaches to gait recognition are as follows.

Sarker *et al*. [2] used template matching of silhouettes that were roughly extracted using background subtraction. The silhouette extraction method was refined by Lee *et al*. [3] by using HMM. The template matching method was improved by Tolliver *et al*. [4] by using a variance-weighted metric and by detecting key frames in human gaits. Sundaresan *et al*. [5] applied HMM to the time series analysis of silhouettes. These methods are all based on human silhouettes and template matching (including spatio-temporal *cross*-correlation) for calculating the silhouette-based similarities. Wang *et al*. [6] extracted features from the outer contours of silhouettes without much consideration of the temporal information.

Johansson [7] performed one of the earliest psychological studies related to gait recognition, in which the experiment, called "point lights display," indicated that we can perceive human motion by the cue of only moving patterns of point lights in the dark. In terms of human identification through gaits, Cutting *et al*. [8] found that humans can recognize a particular walker by observing point lights even if familiarity cues are omitted. They also suggested that dynamic cues such as the speed, bounciness, and rhythm of the walker are more important than static cues such as the height of the walker. Note that almost exclusively dynamic cues enable us to recognize human gaits. On the other hand, Veeraraghavan *et al*. [9] compared the role of body shapes (static) with that of kinematics (dynamic) and concluded that body shape is more important than kinematics. Verges *et al*. [10] statistically showed that static parts of body shapes are important for recognition tasks. These two studies (and most of the previous work described above) used silhouette-based recognition, i.e., recognition based on static forms, but Cutting *et al*. [8] 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."

From the point of view that gait recognition is compiled from successive snapshot (shape) recognition, static cues surely play an important role, and body shapes contribute to

human identification. Dynamic cues, however, are equally or more important for identification through gaits, as Cutting *et al*. [8] pointed out (and Veeraraghavan *et al*. [9] showed that using both body shapes and kinematics outperforms using either one alone).

In the previous work [13], we proposed a method of motion recognition using the cubic higher-order local auto-correlation (CHLAC) features which compute the spatio-temporal correlations of pixels indicating movements and incorporate static and dynamic cues in a natural (unified) way. The basic idea of CHLAC is related to that in [8]: dealing equally with the spatial axes and time axis, not with compilations of snapshots. The key point is that the relations among the moving points of light connected to dynamic perception is formulated as spatio-temporal *auto-correlations* of the moving points (see Sect. 3 for details).

The concept of regarding motion images as spatio-temporal data goes back to XYT in [11], where spatio-temporal information is used as the XT-slice. The spatio-temporal information, however, was only for detecting human contours, not for recognition. Recently, Laptev [12] dealt with the spatio-temporal concept more explicitly by detecting "interest points" (called motion events) in the spatio-temporal space of image sequences and describing these events as local feature vectors based on spatio-temporal derivatives.

In this paper, we propose a new approach to gait recognition by using a simple scheme comprising CHLAC feature extraction, discriminant analysis, and *k*-NN decision rules. In [13], CHLAC was simply applied to motion recognition, such as recognition of walking and running, where the evaluation was done by small data set and the parameter setting in CHLAC was not so much crucial. However, the difference of gaits among persons is much more fine and delicate than that of motions and we must carefully treat the parameters in the recognition scheme. Some knowledge of human gaits is introduced for parameter optimisation and integration with classifiers to make our scheme more effective for gait recognition. In an experiment using the NIST gait database [2], we compared the performance of our scheme to other algorithms and found it to be effective and superior.

2. Preprocessing

Before applying CHLAC, we preprocess input image sequences.

First, as shown in **Fig**. 1, an image sequence is regarded as three-way data using the *x*- and *y*-axes in an image frame ($X \times Y$) and the *t*-axis along the frame sequence. Motion is usually composed of characteristic (sub-) motions over certain amounts of time, such as the gait period we used. By capturing the characteristics, we set a time window that includes a constant number of frames along the time-axis.

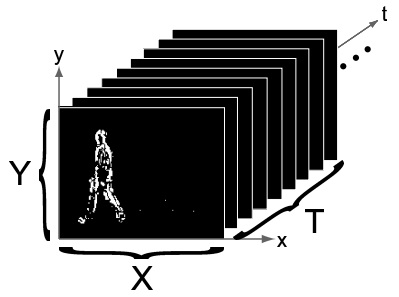

Figure 1: Cubic data showing frame motion as white pixels, which are extracted by subtracting the previous frame from the current frame and by binarizing.

The frames within a window are assigned as one unit, called "cubic data" ($X \times Y \times T$), as in **Fig**. 1. A series of cubic data 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 later. Human motion is recognized in each frame *t* by classifying the feature vector associated with the cubic data.

Second, we apply frame differencing and then automatic thresholding to binarize and detect motion pixels and to filter out both inherent noise and brightness information, such as clothing, which is irrelevant to motion information. Consequently, pixel values in each frame become 1 or 0: "moved" or "static." In **Fig**. 1, a moving human contour is visible, and the contour is sufficient for motion recognition [10]. A little isolated noise might be left in resulting frames, but need not be eliminated because CHLAC is robust to such noise (see Sect. 3.3). In this preprocessing, the frame differencing could be replaced by another method, such as silhouette extraction. The extraction of silhouettes, however, requires more complicated processing (background subtraction, etc.) while frame differencing (and binarization) is easily processed. Note that our method can use frame differencing or silhouette extraction in preprocessing, while the other methods based on template matching use only silhouette extraction as preprocessing.

3. Cubic Higher-Order Local Auto-Correlation

We now describe the details of cubic higher-order local auto-correlation (CHLAC), which was proposed in [13]. Higher-order local auto-correlation (HLAC) was proposed for extracting spatial "auto-correlations," and it was demonstrated to work effectively in image (two-way data) recognition [14]. We extended this naturally to cubic higher-order local auto-correlation to deal directly with three-way data.

In this framework, HLAC related to the static perception is considered a special case of CHLAC related to the dynamic perception.

3.1. Definition

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

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

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

where the $\boldsymbol{a}_i$ $(i = 1, \cdots, N)$ are displacement vectors from a reference point $\boldsymbol{r}$. Although Eq. (1) can take many different forms by varying N and $\boldsymbol{a}_i$, we limit $N \leq 2$ and $\boldsymbol{a}_i$ to a local region because local voxels (pixels) are considered to be highly correlated.

A CHLAC feature corresponds to a value of $R_N(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N)$, and we can obtain many features by varying $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N$ in the local region and using $N = 0, 1, 2$. However, in the case that the point configuration of $(\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(1)} + \boldsymbol{a}_1^{(1)}, \cdots, \boldsymbol{r}^{(1)} + \boldsymbol{a}_N^{(1)})$ matches that of $(\boldsymbol{r}^{(2)}, \boldsymbol{r}^{(2)} + \boldsymbol{a}_1^{(2)}, \cdots, \boldsymbol{r}^{(2)} + \boldsymbol{a}_N^{(2)})$ by shifting it, $R_N(\boldsymbol{a}_1^{(1)}, \cdots, \boldsymbol{a}_N^{(1)})$ takes the same value as $R_N(\boldsymbol{a}_1^{(2)}, \cdots, \boldsymbol{a}_N^{(2)})$. Therefore, we eliminate such duplicated sets for CHLAC features. The following section gives the details of computing CHLAC features.

3.2. Computation

First, we translate Eq. (1) from a continuous form to a discrete one:

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

where the components of $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N$ are limited to $\pm\Delta r$ or 0 for a_{ix} and a_{iy} and to $\pm\Delta t$ or 0 for a_{it}, and $N \leq 2$. We use Δr to denote the spatial interval along the *x*- or *y*-axis in an image frame, and Δt denotes the temporal interval along the *t*-axis in the frame sequence. The interval along the x-axis is taken identically to that along the *y*-axis because of the isotropy in the x-y plane. On the other hand, the spatial interval Δr may be different from the temporal interval Δt because the resolution of space and time may differ. The determination of these parameters will be discussed later.

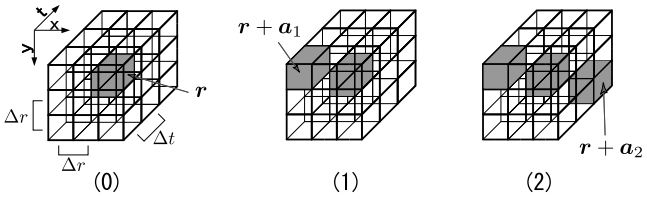

Figure 2: Examples of independent mask patterns: (0) $N = 0$; (1) $N = 1, \boldsymbol{a}_1 = (-\Delta r, -\Delta r, -\Delta t)^T$; and (2) $N = 2, \boldsymbol{a}_1 = (-\Delta r, -\Delta r, -\Delta t)^T, \boldsymbol{a}_2 = (\Delta r, \Delta r, \Delta t)^T$.

The set $(\boldsymbol{r}, \boldsymbol{r} + \boldsymbol{a}_1, \cdots, \boldsymbol{r} + \boldsymbol{a}_N)$ is represented as a local mask pattern, of which examples are shown in **Fig**. 2. In Eq. (2), we first multiply the voxel values of the gray positions in the mask pattern (correlation term), and then sum up the resulting value in the whole region of cubic data by shifting the mask pattern (integral term). For example, in the case of $\boldsymbol{a}_1 = (-\Delta r, -\Delta r, -\Delta t)^T$ and $N = 1$ as in **Fig**. 2 (1), we obtain the feature value

$$R_1(\boldsymbol{a}_1) = \sum_{x,y,t \in D_s} f(x, y, t) f(x - \Delta r, y - \Delta r, t - \Delta t).$$

Next, we describe how to construct such mask patterns. There are many mask patterns including duplicated patterns in terms of point configurations. The mask patterns that can be matched by shifting can be eliminated (**Fig.** 3): 279 independent mask patterns result. In cases where each voxel value is either 0 or 1 in the three-way data, 251 mask patterns are possible because $f(\boldsymbol{r})^2 = f(\boldsymbol{r})$ and $f(\boldsymbol{r})^3 = f(\boldsymbol{r})$, e.g.,

$$R_0 = \int f(\boldsymbol{r}) dr = \int f(\boldsymbol{r})^2 dr = R_1(\boldsymbol{0}) \quad \left[\boldsymbol{a}_1 = (0, 0, 0)^T\right].$$

The dimensions of CHLAC features correspond to the number of mask patterns. We use the latter 251 dimensional features because the voxel values in cubic data are binarized (Sect. 2).

3.3. Desirable properties

This CHLAC method extracts spatio-temporal features from three-way data in only one step, which differs from the traditional approaches requiring two steps: shape feature extraction and temporal feature extraction. The CHLAC features are easily calculated because they consist only of multiplication and addition, so this is a simple and concise method. Furthermore, it has the following three desirable properties for recognition.

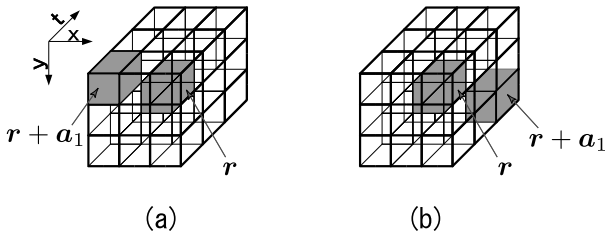

Figure 3: Example of duplicate mask patterns: (a) $N = 1, \boldsymbol{a}_1 = (-\Delta r, -\Delta r, -\Delta t)^T$; (b) $N = 1, \boldsymbol{a}_1 = (\Delta r, \Delta r, \Delta t)^T$. The mask pattern in (a) corresponds to that in (b) shifted by $(\Delta r, \Delta r, \Delta t)^T$.

- *Shift-invariance* to data: This is because the features are based on an integral (summation). Note that the shift-invariance renders the method *segmentation-free*.
- *Additivity* for data: Suppose that regions A and B are disjoint ($A \cap B = \phi$); then, the feature value of this data is given as

$$R_{\text{whole}} = \int_{\boldsymbol{r}\in(A\cup B)_s} g(\boldsymbol{r})d\boldsymbol{r} \approx \int_{\boldsymbol{r}\in A_s} g(\boldsymbol{r})d\boldsymbol{r} + \int_{\boldsymbol{r}\in B_s} g(\boldsymbol{r})d\boldsymbol{r} = R_A + R_B,$$

  where $g(\boldsymbol{r}) = f(\boldsymbol{r})f(\boldsymbol{r}+\boldsymbol{a}_1)\cdots f(\boldsymbol{r}+\boldsymbol{a}_N)$. This holds because auto-correlations are almost limited to each region (A or B) due to their locality. This property makes it possible to simultaneously identify multiple objects [13].

- *Robustness to noise* in data: The auto-correlation is robust to additive noise, as shown in the following. Let s_i be signal and and n_i be random noises with means of 0 and variances of σ^2 at the i-th voxel; then, assuming that $s_i \gg n_i$,

$$\mathbf{E}(s_i + n_i)(s_j + n_j) = \mathbf{E}(s_i s_j + \sigma^2\delta_{ij}) \fallingdotseq \mathbf{E}s_i s_j,$$

  where $\mathbf{E}$ is the expectation, and δ_{ij} is the Kronecker delta. In addition, noise such as isolated points hardly affects CHLAC feature values because the portion of such noise is usually much smaller than the portion of the object; furthermore, the correlations between the noise point and surrounding points are mostly zero.

4. Recognition Scheme

After CHLAC feature extraction, we apply discriminant analysis and k-NN decision rules for gait recognition as follows.

Learning

Input: All training image sequences
For $(\Delta r, \Delta t, T) \in ParamRange$
Do
1. **Defining Cubic Data** by using T
2. **Frame Differencing and Binarization**
3. **CHLAC Feature Extraction** using $\Delta r, \Delta t$
4. **Discriminant Analysis**
   4.1 Applying DA for All features to create $S(\Delta r, \Delta t, T)$
   4.2 Mapping all features into $S(\Delta r, \Delta t, T)$

Done

Figure 4: Learning phase

In the learning phase, CHLAC features of a certain parameter set, $R(\Delta r, \Delta t, T)$, are extracted from all cubic data of the whole image sequence of a training set. Fisher Discriminant Analysis (DA) is applied to these features, and then the features belonging to each person are clustered in the discriminant space $S(\Delta r, \Delta t, T)$ that is preferable for the recognition. Many different discriminant spaces are constructed for all parameter sets that lie in the parameter range (see Sect. 5.2). This learning phase is summarized in **Fig**. 4.

In the recognition phase, the classifier is based on a k-NN decision rule (say, $k = 10$). At each time t, a CHLAC feature is extracted for each parameter set, $R_t(\Delta r, \Delta t, T)$, and the k-NN decision is made in the corresponding discriminant space, $S(\Delta r, \Delta t, T)$. We repeat this k-NN decision for all discriminant spaces constructed in the learning phase. The frame at t is classified as follows:

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

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

$\text{kNN}_{S(\Delta r,\Delta t,T)}(\boldsymbol{x}, P_i)$ counts the number of training samples belonging to i-th person, P_i, in the k-nearest neighbors of $\boldsymbol{x}$ in the space $S(\Delta r, \Delta t, T)$. This k-NN number is regarded as the confidence of the person on the parameter set, and by searching the maximum confidence over $\Delta r, \Delta t, T$, and P in Eq. (3), the recognition result is more stable and accurate because the parameter sets may have different discriminant power for each person. The constraints in (4) are determined in Sect. 5.2.

Recognition

Input: Test image sequence (T_s image frames)
$\text{conf}_p \leftarrow \mathbf{0}$
For $t \leq T_s$
Do
  For $(\Delta r, \Delta t, T) \in ParamRange$
  Do
    1. **Defining Cubic Data**
       Image frames ($t \sim t + T - 1$) as cubic data
    2. **Frame Differencing and Binarization**
    3. **CHLAC Feature Extraction** using $\Delta r, \Delta t$
    4. **Mapping** the feature into $S(\Delta r, \Delta t, T)$
  Done
  k-NN Classifier
    1. Calculate confidence (i.e. Result_F) by Eq. (3)
    2. $\text{conf}_p[\text{Result}_F]$++
Done
$\text{Result} \leftarrow \arg\max_i \text{conf}_p[i]$

Figure 5: Recognition phase

Finally, the sequence is identified as a person as

$$\text{Result} = \arg\max_i \sum_{\text{Rresult}_F(t) \in P_i} 1. \qquad (5)$$

Namely, in the image sequence, the person that the most number of frames support is the final classification/identification result, which makes it possible to avoid the effect of imprecise recognition results derived from noisy cubic data. Recognition phase is summarized in **Fig**. 5.

5. Experiment

5.1. Gait data

To evaluate the performance of our scheme, we used the NIST gait dataset, which is the largest such dataset available. It consists of 456 video sequences of 71 individuals (persons) walking around an elliptical course, with labels: Gallery (for training) and probes A through G (for testing). The details are given by Sarker *et al*. [2].

5.2. Optimising parameters in CHLAC

Three parameters must be determined: the spatial and temporal intervals $\Delta r, \Delta t$, and T, which cannot be optimally defined without knowledge and can take any values. Therefore, we take into account some knowledge about the characteristics of human gaits to restrict the range of these parameters.

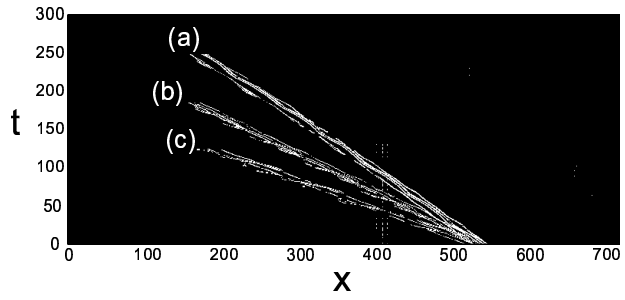

Figure 6: Trajectories of different humans walking in an XT-slice. (a): slowest, (b): middle speed, and (c): fastest walks.

5.2.1 Spatial and temporal intervals, Δr and Δt

The only constraint on Δr and Δt is locality; however, many combinations remain. Some knowledge of human gaits further constrains the relationship between Δr and Δt.

Suppose we have a fronto-parallel view of a human walk. If the image sequence is sliced horizontally at the middle of the human shape, the sliced surface also forms an image plane (in the $x-t$ plane): a so-called XT-slice [11] (**Fig**. 6). This shows that the trajectory of human walking can be approximated as a straight line, of which the gradient denotes the walking velocity. The relationship between the spatial and temporal intervals is closely connected to this gradient (velocity). If $\arg(-\Delta r, \Delta t)^T$ is far from the gradients of the human trajectories, almost of the CHLAC feature values are close to zero because no human is at the time and place $(-\Delta r, \Delta t)^T$ from the current human position in the XT-slice. Therefore, $\arg(-\Delta r, \Delta t)^T = -\Delta t/\Delta r$ should be close to most of the gradients, that is, the mean of the gradients (**Fig**. 6 (b)). We adopted principal component analysis (PCA) to approximate each person's trajectory by a straight line and then estimate the gradient. After applying PCA to the dot patterns (x, t) composing the trajectory in the image sequence, the eigenvector associated with the maximum eigenvalue represents the gradient vector of the human trajectory in an XT-slice. From the first eigenvectors of all image sequences, the mean gradient of the trajectories is calculated. In practice, however, preprocessing causes some noise in XT-slices, which makes the estimation imprecise. Thus, we use the contribution rate, $\eta_1 = \lambda_1 / \sum_i \lambda_i$, to evaluate the appropriateness of the straight-line approximation. The eigenvectors of which the contribution rates are less than a threshold (say, 0.99) are discarded, and the mean gradient is estimated by averaging only good (reliable) eigenvectors. The mean gradient over persons was computed as -0.49, which showed that $\Delta t/\Delta r = 1/2$.

On the other hand, in the image frame, knowledge about the human body (the width of human figures) restricted Δr

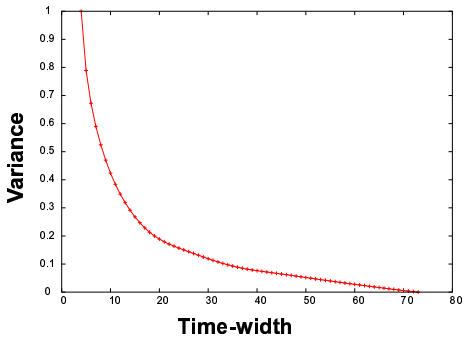

Figure 7: Variance vs. time-width. The variance is averaged through all image sequences. It becomes enough small around 30 frames.

to 16 or less.

5.2.2 Time-width, T

As the time-width T of cubic data increases, more information on cyclic gaits is obtained, and the CHLAC features become stable. However, if T is too large, the effect of unreliable frames that include much noise could remain for a long time, i.e., much cubic data could include the noisy frames. Because of this trade-off, T should have a limited length. If we assume that human gaits are periodic motions, the CHLAC features of cubic data whose T is close to the period would be ideally stable. To investigate the stability of CHLAC features, we checked the variance of features in each image sequence for various time-widths T under the assumption that the gait period is constant within each image sequence. The variance vs. time-width T is shown in **Fig**. 7. The variance became enough small around 30 frames, so we set this as the average gait period. In [5] and [2], the gait period was also 30-40 frames. Here, note that the gait period was calculated using the stability of features (i.e., variance) without applying an object-model (such as the angle of legs) based analysis used in the other studies. Thus, the time-width, T, was determined as 30 frames or less. We did not assume that the gait periods of all persons are 30 frames, but roughly set the upper bound of T using data. The same is true of Δr.

As a result, the parameter range, i.e., the constraints of the parameters in (4), is determined as

$$\Delta t/\Delta r = 1/2, \quad \Delta r \leq 16, \quad T \leq 30. \tag{6}$$

Discriminant spaces are constructed for every parameter satisfying the constraints in (6), and k-NN decisions are made in these spaces. These constraints are not so heuristic and not so strong because they are truly derived from the data (training set) by introducing a little knowledge about human gaits. They make it possible to extract CHLAC features more effectively for human gaits, and by combining this knowledge with the decision rules in Eq. (3) and (5), our scheme becomes much more efficient.

5.3. Results

The identification results compared with those of the methods in [5] [2] [4] [3] and [6] are shown in **Fig**. 8. The identification rate of our scheme is also given in **Table** 1, column (a). Our scheme outperformed the others in all probes. The identification results of probes D to G are worse than those of probes A to C for all methods. This is caused by the difference of walking surfaces: A to C are on grass, D to G are on concrete, and Gallery is on grass. The surface may affect gait periods, preprocessing, and recognition. Thus, probes D to G, whose surfaces are different from that of Gallery, are difficult and challenging problems. However, our scheme performs much better than any other method even in these probes because CHLAC is robust to the results of preprocessing, which is shown as follows.

Table 1 shows our method's dependence on the quality of preprocessed data: noise in the background and in human regions. The term "bbox" means that the human region (bounding box) is extracted, and pixels in the other regions are set to 0 (noiseless) after binarization, which controls (suppresses) the amount of background noise. The term "half-threshold" means binarization with half of the value suggested by automatic thresholding, which controls the amount of noise and thickness of human contours at the same time. "Half-threshold" increases noise but makes human contours thicker, which is the opposite of "automatic-threshold." If we compare columns (b) and (c) in **Table** 1, our method is slightly affected by background noise, but if we compare columns (a) and (b), we see that using information on human contours overcomes noise.

Here, we discuss the reason that our scheme is so effective. A preprocessed frame contains only dot patterns of a human contour, and a frame sequence contains the manifold formed by successive human contours (dot patterns) in three dimensions (x, y, t). This manifold includes all information about the person's movement. It consists of global and local characteristics that correspond to the motion speed and gait, respectively. Cubic higher-order *local* auto-correlation extracts these *local* characteristics. CHLAC is not only derived from the concept of correlation but is also closely connected to gradients and curvatures (local characteristics) in the particular case of binary (1 or 0) data. The gradient of the manifold is approximated by the configuration of every two neighbouring points, i.e., which direction the next point is shifted in. In a similar way, the curvature is characterized by the configuration of every three neighbouring points. These configurations of two or three points are directly described by the first and second order mask patterns,

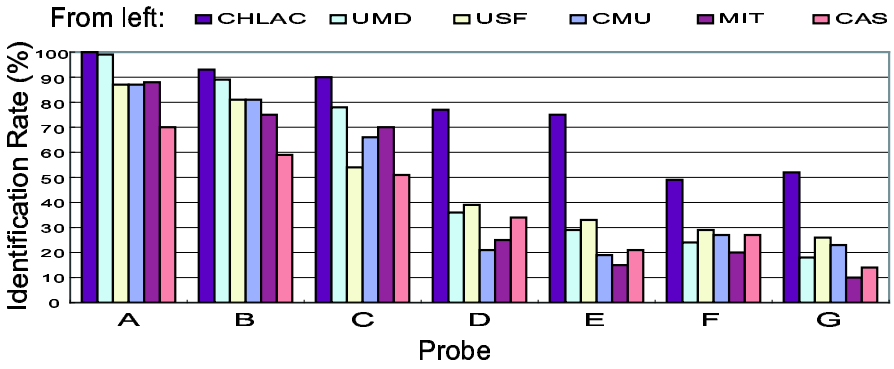

Figure 8: The identification rate (%) for each probe using the following methods: CHLAC, UMD [5], USF [2], CMU [4], MIT [3], and CAS [6] (these are the top-ranked results). See each paper for detailed identification rates, or [2] for collective results. The details of our method (CHLAC) are given in **Table** 1, column (a).

Table 1: The identification rates (%) for various conditions of our method. Details are in text.

Probe	bbox		non-bbox
	half-threshold	automatic-threshold	
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)

respectively: each mask pattern denotes the direction of the gradient or the curvature (see **Fig**. 2). Furthermore, the gradient and curvature can be regarded as the velocity and acceleration of an individual point by considering the time-axis and can also be understood as the characteristics of the human shape in the x- and y-axes. In addition, CHLAC roughly extracts global characteristics by an integral of local characteristics. Thus, CHLAC can effectively extract the characteristics of human gaits.

Note that CHLAC is applicable to three-dimensional geometrical (x, y, z) data and to any other form of three-way data.

6. Conclusion

We have proposed a novel scheme for human identification by gaits in image sequences. The scheme consists of feature extraction using cubic higher-order local auto-correlation (CHLAC), discriminant analysis, and *k*-NN decisions. While traditional silhouette-based approaches require at least two steps of shape and time series analysis, CHLAC enables directly extracting spatio-temporal features as the spatio-temporal auto-correlations of motion voxels in an image sequence. The range of parameters of spatial and temporal intervals and the time-width in CHLAC was derived from data by considering the characteristics of human gaits, and then utilized effectively for optimal recognition. It is noted that the parameter range was adaptively and analytically determined from data, not by hand.

Our experiments using the NIST gait dataset showed that our scheme is greatly superior to other methods, especially on more challenging problems (probes D to G).

CHLAC is robust to noise in data and is applicable as a *segmentation-free* method for various motion recognition tasks other than gait recognition. Moreover, the geometric meaning of CHLAC, such as gradients and curvatures, makes this method applicable to three-dimensional geometrical analysis, such as object recognition.

References

[1] Nixon, M., Carter, J. Advances in automatic gait recognition. In: International Conference on Automatic Face and Gesture Recognition, 2004.

[2] Sarker, S., Philips, P., Liu, Z., Vega, I.R., Grother, P., Bowyer, K. The humanid gait challenge problem: Data sets, performance, and analysis. Pattern Analysis and Machine Intelligence, Vol. 27, pp. 162–177, 2005

[3] Lee, L., Dalley, G., Tieu, K. Learning pedestrian models for silhouette reinforcement. In: International Conference on Computer Vision, 2003.

[4] Tolliver, D., Collins, T. Gait shape estimation for identification. In: International Conference on Audio- and Video-Based Biometric Person Authentication, 2003.

[5] Sundaresan, A., Chowdhury, A., Chellappa, R. A hidden markov model based framework for recognition of humans from gait sequences. In: International Conference on Image Processing, 2003.

[6] Wang, L., Tan, T., Ning, H., Hu, W. Silhouette analysis-based gait recognition for human identification. Pattern Analysis and Machine Intelligence, Vol. 25, pp. 1505–1518, 2003

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

[8] Cutting, J., Kozlowski, L. Recognizing friends by their walk: Gait perception without familiarity cues. Bulletin of the Psychonomic Society, Vol. 9, pp. 353–356, 1977.

[9] Veeraraghavan, A., Roy-Chowdhury, A., Chellappa, R. Matching shape sequences in video with applications in human movement analysis. Pattern Analysis and Machine Intelligence, Vol. 12, pp. 1896–1909, 2005.

[10] Veres, G., Gordon, L., Carter, J., Nixon, M. What image information is important in silhouette-based gait recognitnion? In: International Conference on Computer Vision, 2004.

[11] Niyogi, S., Adelson, E. Analyzing and recognizing walking figures in xyt. In: IEEE Conference on Computer Vision and Pattern Recognition, 1994.

[12] Laptev, I. On space-time interest points. International Journal of Computer Vision, Vol. 64, pp. 107–123, 2005

[13] Kobayashi, T., Otsu, N. Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation. In: International Conference on Pattern Recognition, 2004. anonymous

[14] Otsu, N., Kurita, T. A new scheme for practical flexible and intelligent vision systems. In: IAPR Workshop on Computer Vision, 1988.