Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Real-Time Imaging 11 (2005) 219–232

www.elsevier.com/locate/rti

Detection of cyclic human activities based on the morphological analysis of the inter-frame similarity matrix

Alexandra Branzan Albu*, Mehran Yazdi, Robert Bergevin

Computer Vision and Systems Laboratory, Department of ECE, Laval University, Québec, Canada, G1 K 7P4

Available online 31 May 2005

Abstract

This paper describes a new method for the temporal segmentation of periodic human activities from continuous real-world indoor video sequences acquired with a static camera. The proposed approach is based on the concept of inter-frame similarity matrix. Indeed, this matrix contains relevant information for the analysis of cyclic and symmetric human activities, where the motion performed during the first semi-cycle is repeated in the opposite direction during the second semi-cycle. Thus, the pattern associated with a periodic activity in the similarity matrix is rectangular and decomposable into elementary units. We propose a morphology-based approach for the detection and analysis of activity patterns. Pattern extraction is further used for the detection of the temporal boundaries of the cyclic symmetric activities. The approach for experimental evaluation is based on a statistical estimation of the ground truth segmentation and on a confidence ratio for temporal segmentations.

1. Introduction

Human motion analysis plays a central role in surveillance systems designed to observe uncontrolled environments by using a non-intrusive camera network. As outlined in several recent surveys [1,2], the real-time detection and tracking of human subjects in the scene is focused on the non-rigid motion analysis of the human body and it aims to provide robust solutions with respect to shadowing, occlusion, and pose change.

In the context of human motion analysis, the cyclic feature is mainly used as a cue for detecting activities such as walking, running, and subject identification through gait recognition. Tsai et al. [3] detect walking cycles using the spatio-temporal curvature function of trajectories corresponding to nine specific points on the human subject in motion. Their technique uses a stick-model and is designed for motion-based recognition, namely for identifying the tracked subject from his motion.

Polana and Nelson [4,5] introduce the concept of temporal texture for the detection of periodic activities such as walk, exercise, swing, and rotation. A low-level technique for recognizing repetitive human activities is proposed in [5]. This technique, based on bottom-up processing, does not require a prior segmentation of the human subject into body parts. Using the fundamental frequency of the repetitive action, a feature vector of spatio-temporal motion magnitudes is built. The experimental database in [4,5] consists of video sequences containing one activity per sequence and a cluttered background.

Seitz and Dyer [6] define the notion of period trace which allows a relaxation of the assumption that a motion should be perfectly even from one cycle to the next. The period trace is recovered using affine-invariant image matching and is useful to describe motions, like an athlete running, that are not strictly cyclic but have a cyclic component. As in [4,5], the database in [6] consists in video sequences dedicated to one specific activity.

While the study of sequences containing a single periodic activity has led to interesting results in human motion modeling and representation, there is little

*Corresponding author.
E-mail addresses: branzan@gel.ulaval.ca (A.B. Albu), yazdi@gel.ulaval.ca (M. Yazdi), bergevin@gel.ulaval.ca (R. Bergevin).

1077-2014/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.rti.2005.03.004

research about video sequences where the activity pattern changes over time. Recent work by Bobick and Davis [7] deals with the temporal segmentation of video sequences into coherent actions based on a backward-looking temporal time window. A low-level representation of motion is created from motion-energy images (MEI) and motion-history images (MHI). The temporal segmentation is performed by using three parameters corresponding to the minimum and maximum duration of an action and to the maximum number of temporal integration windows, respectively. Since the parameter values are computed with respect to a prior manual temporal segmentation of the entire database, the temporal video segmentation in [7] is very sensitive with respect to the speed of performing actions. In addition, this method is only able to handle human actions specified during the off-line training phase, which builds reference templates for every action of interest.

To address such current limitations in human activity analysis, we propose a new method for the temporal segmentation of video sequences containing both cyclic and non-cyclic activities. Our method uses the 2D inter-frame similarity plot, a concept introduced by Cutler and Davis in [8]. As shown in BenAbdelkader et al. [9], the matrix of inter-frame similarity can be successfully used for gait-based person identification. In the context of our research, we investigated the relevance of the 2D inter-frame similarity plot for the detection of cyclic and symmetric actions from video sequences containing multiple human activities. Preliminary results of our study appeared in Yazdi et al. [10]. The present paper contains a detailed description of the proposed method, as well as a complete experimental evaluation.

The rest of the paper is organized as follows. Section 2 consists of an overview of the approach. The experimental results are presented and discussed in Section 3. Section 4 contains the conclusions and describes the future work directions.

2. An overview of the proposed approach

The modular decomposition of the proposed temporal segmentation approach is shown in Fig. 1. The pre-processing phase contains three steps: background subtraction, shadow removal, and silhouette rescaling. After pre-processing, the initial sequence is transformed into a sequence of binary images containing human silhouettes in bounding boxes of standard size. The inter-frame similarity matrix computed by cross-correlation is displayed as an image. This type of image represents input data for our morphology-based approach designed for the detection and analysis of spatio-temporal patterns corresponding to periodic human activities. This approach represents a main contribution of the study reported in this paper.

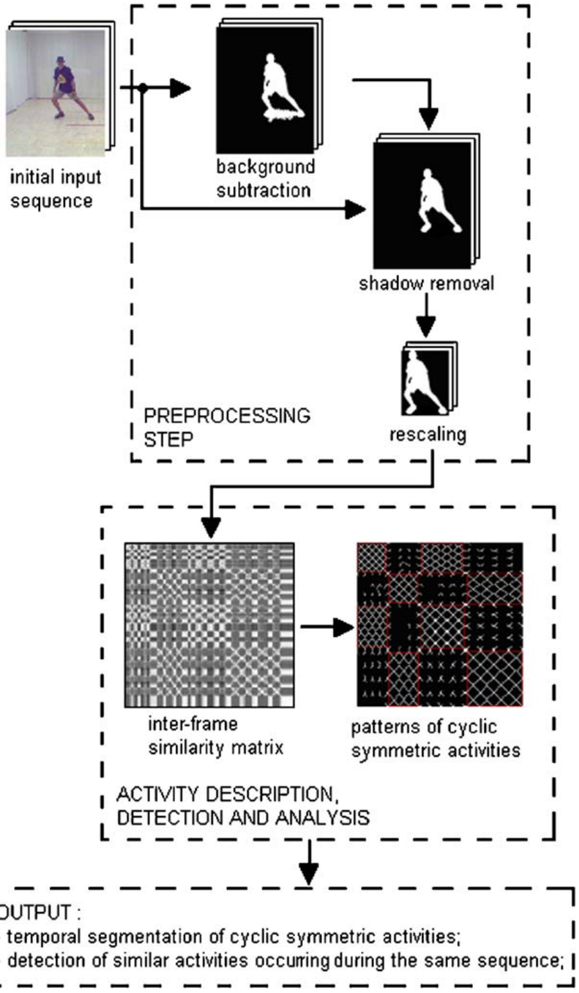
initial input sequence
background subtraction
shadow removal
rescaling
PREPROCESSING STEP
inter-frame similarity matrix
patterns of cyclic symmetric activities
ACTIVITY DESCRIPTION, DETECTION AND ANALYSIS
OUTPUT :
- temporal segmentation of cyclic symmetric activities;
- detection of similar activities occurring during the same sequence;

Fig. 1. Modular decomposition of the proposed approach.

2.1. Pre-processing

Pre-processing is always a critical step and usually consists of probabilistic methods for handling complex outdoor scenes, as proposed by Stauffer and Grimson [11]. The pre-processing approach presented in this paper consists of a sequence of three rather simple algorithms for background subtraction, shadow removal, and silhouette normalization. This sequential approach obtains good results for sequences acquired with a static camera in a typical office environment. However, the design of the main module performing the activity detection and analysis task (see Section 2.2) is flexible enough to accept data from any pre-processing module performing the three previously mentioned tasks. Therefore, the proposed temporal segmentation approach might be generalized for the analysis of outdoor scenes if used in conjunction with a more sophisticated, real-time background subtraction algorithm.

2.1.1. Background subtraction

The background subtraction, also known as the figure-ground segmentation problem, consists in classifying the pixels in each frame of the sequence as belonging to either background or foreground objects. Since the database for this study is acquired with a static camera in a typical indoor laboratory environment, the assumption of a static background is valid and allows for a reference background image to be defined.

For each frame in the sequence, the difference image between the given frame and the reference image is computed as follows:

$$D(i,j) = |R(i,j) - R_0(i,j)| + |G(i,j) - G_0(i,j)| + |B(i,j) - B_0(i,j)|, \quad (1)$$

where R, G, B and R_0, G_0, B_0 are the colour components of the considered frame and of the reference image, respectively. The difference image is then binarized using Otsu's automatic threshold selection method [12]. Isolated pixels in the binary image due to noise, specularities, or to a small amount of background motion are filtered out using a median filter of size 3 × 3. A visualization of the background subtraction step is presented in Fig. 2.

2.1.2. Shadow removal

The shadow of a human silhouette introduces morphological artefacts in the binary blob resulting from background subtraction. As shown in Fig. 2d, background subtraction does not eliminate shadows in a real-world environment with uncontrolled lighting. Without shadow removal, the inter-frame self-similarity measure is sensitive with respect to variable depth, changes in the direction of motion and environmental lighting conditions. For these reasons, a shadow removal technique is used in the pre-processing phase.

As shown in Horprasert et al. [13], the semi-transparency of the shadowing phenomenon allows the texture and colour distribution of the underlying surface to be preserved. Therefore, the shadow of a moving person exhibits a similar chromaticity but lower brightness than the corresponding background region. While our approach is based on this above-mentioned observation, it is significantly different from the technique described in [13]. Thus, we first perform background subtraction, followed by shadow removal. Moreover, the proposed approach is region-based and it does not work on a pixel-by-pixel basis. Therefore, it is more robust with respect to noise and partial occlusions. In addition, our approach works with the HSV colour space, which is considered to be more appropriate than the RGB colour space used in [13] for separating brightness from chromatic information.

A two-step shadow removal algorithm is proposed as shown in Fig. 3. While the first step performs a partition of the silhouette into subregions, the second step decides which subregions correspond to shadow and shall be removed from the foreground.

First, the foreground silhouette is partitioned into a set of sub-regions using automatic mode identification in the intensity histogram. Prior to peak detection, the histogram is smoothed with a 1 × 3 median filter. The distinct modes in the histogram are automatically separated by finding the local minima with standard signal analysis techniques, e.g. sign changes of the first derivatives. After partition, every subregion in the foreground silhouette contains only pixels with their intensity value belonging to the same histogram mode. However, there is no pairwise correspondence between histogram modes and silhouette subregions. Since a moving human might cast distinct shadows on the floor and walls, one histogram mode can be represented by several disjoint subregions.

The second step is to determine which subregions in the silhouette correspond to shadow. The chromaticity analysis is performed in the Hue-Saturation-Value (HSV) colour space. The chromatic appearance of a given region is described with a feature vector $[\mu_H, \sigma_H, \mu_S, \sigma_S]$ where $\mu_H$and $\mu_S$are the average values of the hue and saturation, and $\sigma_H$and σ_S are the standard deviations.

Since the shadow regions preserve the chromatic appearance of the background, each subregion in the

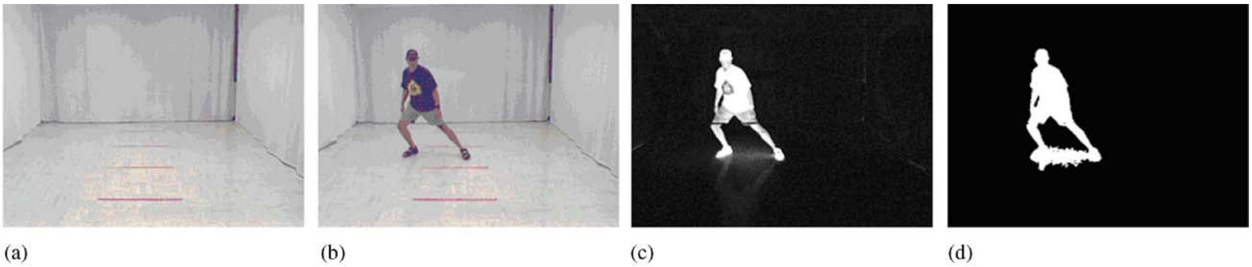

Fig. 2. Background subtraction algorithm: (a) reference image, (b) a given frame, (c) difference image, (d) final smoothed binary image.

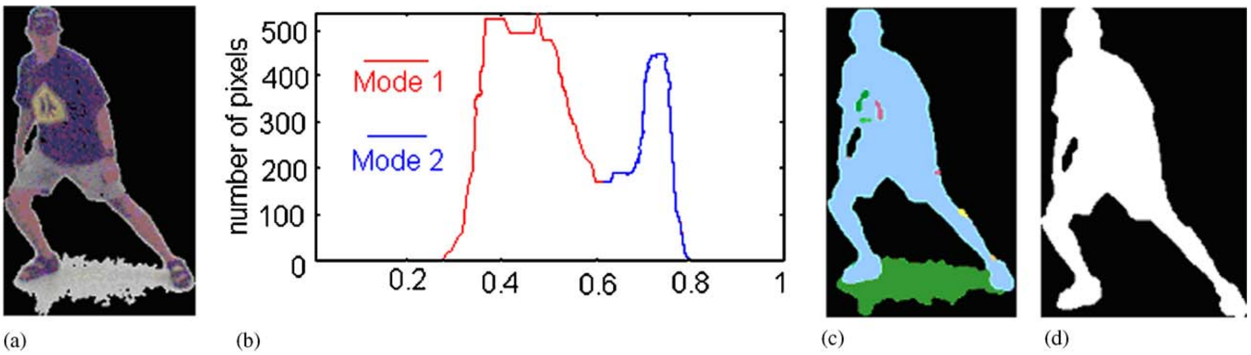

Fig. 3. Outline of the shadow removal algorithm: (a) the foreground region after background subtraction; (b) normalized bi-modal intensity histogram; (c) colour-coded foreground segmentation into several regions; (d) final binary template after shadow removal.

silhouette is compared with its corresponding subregion in the background reference image by computing the Euclidian distance between the feature vectors. The method involves a threshold which decides the acceptable level of similarity between a subregion of the silhouette and the corresponding background. Any subregion exceeding this similarity level is considered to be a shadow and it is removed from the silhouette. Fig. 3d shows the result of the proposed shadow removal algorithm for the silhouette in Fig. 3a. The threshold value is not adjustable from one frame to another and is constant for a given sequence.

2.1.3. *Binary blob rescaling*

Background subtraction and shadow removal result in sequences of binary blobs describing the motion of the human subject. The bounding box associated with the binary blob in motion is of variable size according to the location and posture of the subject in the field of view of the camera. To compensate for depth changes occurring between successive cycles of the same periodic activity, the bounding boxes corresponding to the binary blobs are to be rescaled to a constant size through the entire sequence. Cutler and Davis [8] use uniform height rescaling, based on the assumption that the height of a moving person should be an invariant. While this assumption works well for activities such as walking, it fails in a more generic context, involving activities such as arm swing, torso bending, etc. Thus, we propose a 2D height- and width-rescaling process using the nearest-neighbour interpolation method. Fig. 4 shows the rescaling result for a sequence of binary images corresponding to one semi-cycle of a periodic arm swing.

As shown in Fig. 5, the binary sequences resulting from the preprocessing step contain a fair amount of noise, resulting in morphological distortions of the binary blob. These distortions are due to the limited performance of the background subtraction and shadow removal algorithms and to light reflections. Such artefacts do not significantly affect the primary structure (see Fig. 7) of the inter-frame similarity matrix for periodic human activities.

2.2. *Activity detection, description, and analysis*

2.2.1. *The inter-frame similarity matrix*

Cutler and Davis [8] have introduced the notion of 2D inter-frame similarity plot for the time–frequency analysis of periodic motion. Proposing a new spatio-temporal perspective, we redefine the previous concept as follows. Given a sequence of N normalized binary frames of standard height H and width W, the inter-frame similarity matrix is $[r_{ij}]_{1\leqslant i,j\leqslant N}$, where r_{ij} is the cross-correlation of frames i and j. The similarity measure between two frames i and j is defined as

$$r_{ij} = \frac{\sum_{m=1}^{H}\sum_{n=1}^{W}(i_{mn}-\bar{i})(j_{mn}-\bar{j})}{\sqrt{\sum_{m=1}^{H}\sum_{n=1}^{W}(i_{mn}-\bar{i})^2(j_{mn}-\bar{j})^2}},$$
$$i = mean(i) \quad \text{and} \quad j = mean(j). \tag{2}$$

This similarity measure takes values from −1 to 1 according to the Cauchy–Schwarz inequality. A maximum value of 1 is reached only when comparing identical frames, while the −1 value is obtained when comparing frames where there is no overlap between the binary silhouettes.

While cross-correlation is a standard method for identifying similar grey-level images, it is less used for the comparison of binary images. In our case, inter-frame cross-correlation based on brightness information does not yield robust results with respect to specular effects due to the scene lighting pattern. Therefore, we have successfully redefined the inter-frame similarity matrix based on cross-correlation between binary images.

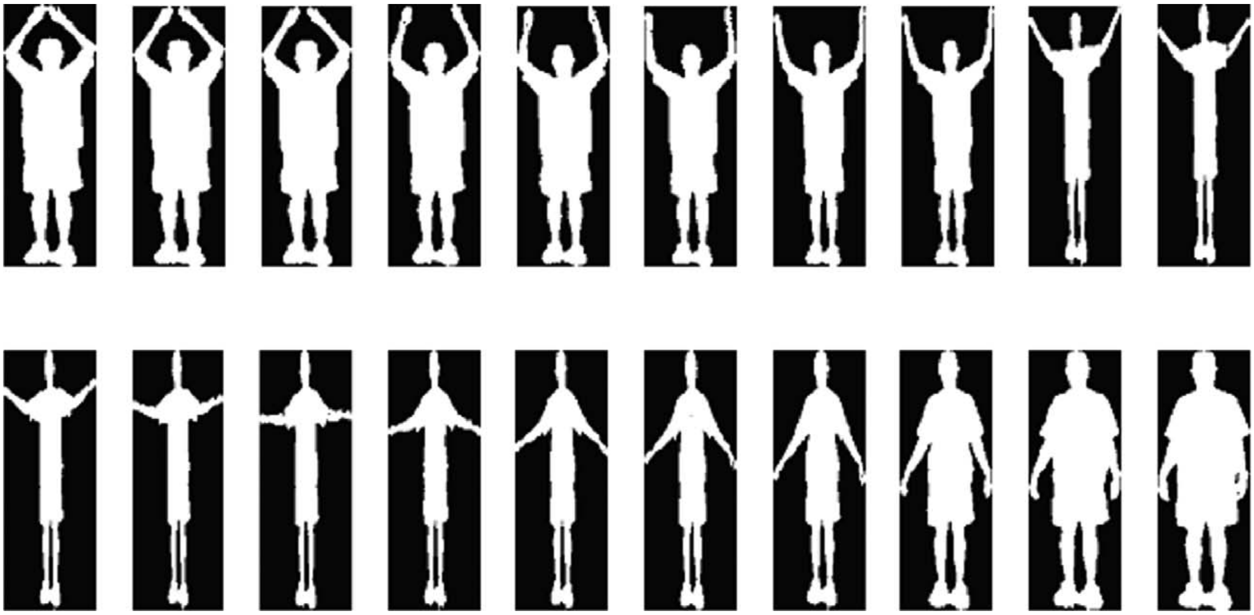

Fig. 4. One semi-cycle of a periodic arm swing.

Fig. 6a shows a 1D cross-correlation plot for a sequence containing four periodic activities. This plot is computed with respect to a reference, namely the first frame in the sequence. The peaks with values near 1 correspond to similar frames with respect to the reference. The plot shown in Fig. 6a corresponds to the first line of the inter-frame similarity matrix (see Fig. 6b) computed for the same sequence. The display of this matrix maps the $[-1,1]$ range of the correlation coefficients towards a discrete set of 256 grey levels. Since the main diagonal consists entirely of self-correlation coefficients, it represents a white line for any analysed sequence.

The inter-frame similarity matrix exhibits some relevant properties for the analysis of a particular category of human activities. Specifically, we are interested in detecting periodic and symmetric activities, where the motion performed during the first semi-cycle is repeated in the opposite direction during the second semi-cycle. In a cyclic symmetric activity, consecutive frames belonging to the first semi-cycle and similar to the reference frame (i.e. the first frame in the row) form bright segments parallel to the main diagonal (see Fig. 7a, center). In addition, bright segments orthogonal to the main diagonal represent the second semi-cycle. A periodic and symmetric activity is represented by a zigzag pattern where the primitive is a V shape corresponding to one cycle (see Fig. 7a, left). As shown in Fig. 7a, the pattern associated with a periodic activity in the inter-frame similarity matrix is rectangular, and can be further decomposed into elementary units.

2.2.2. Detection and analysis of patterns in the inter-frame similarity matrix

To obtain an accurate temporal segmentation of cyclic symmetric activities, the first goal is to isolate the activity patterns in the inter-frame similarity matrix. First, the bright regions are extracted by thresholding at 60% of the maximum brightness (see Fig. 7c). Numerous tests performed on the similarity matrices in the database of our study demonstrated that this empirical threshold value preserves a sufficient amount of relevant morphological information in the resulting binary images. It was also found that the threshold value depends on the speed of performing cyclic and symmetric actions. Indeed, a pattern in the similarity matrix corresponding to a low-speed action exhibits a higher average brightness compared to a pattern corresponding to a high-speed action. Experiments demonstrated that thresholding at 60% of the maximum brightness is effective for cyclic activities with a fundamental period ranging from 25 to 35 frames, at a 30 fps acquisition rate.

A sequence of morphological operators is used for pattern extraction. Thus, two iterative dilations with a standard 5 pixel-sized cross-shaped structuring element remove possible line disconnections due to the fixed 60% threshold. The number of iterative dilations is fixed and it was set according to the variability in the period of the analyzed actions (25–35 frames at 30 fps acquisition rate).

Next, a shrinking operator based on conditional erosion reshapes the dilated image edges into one-pixel thin sets of linear segments without disconnecting the

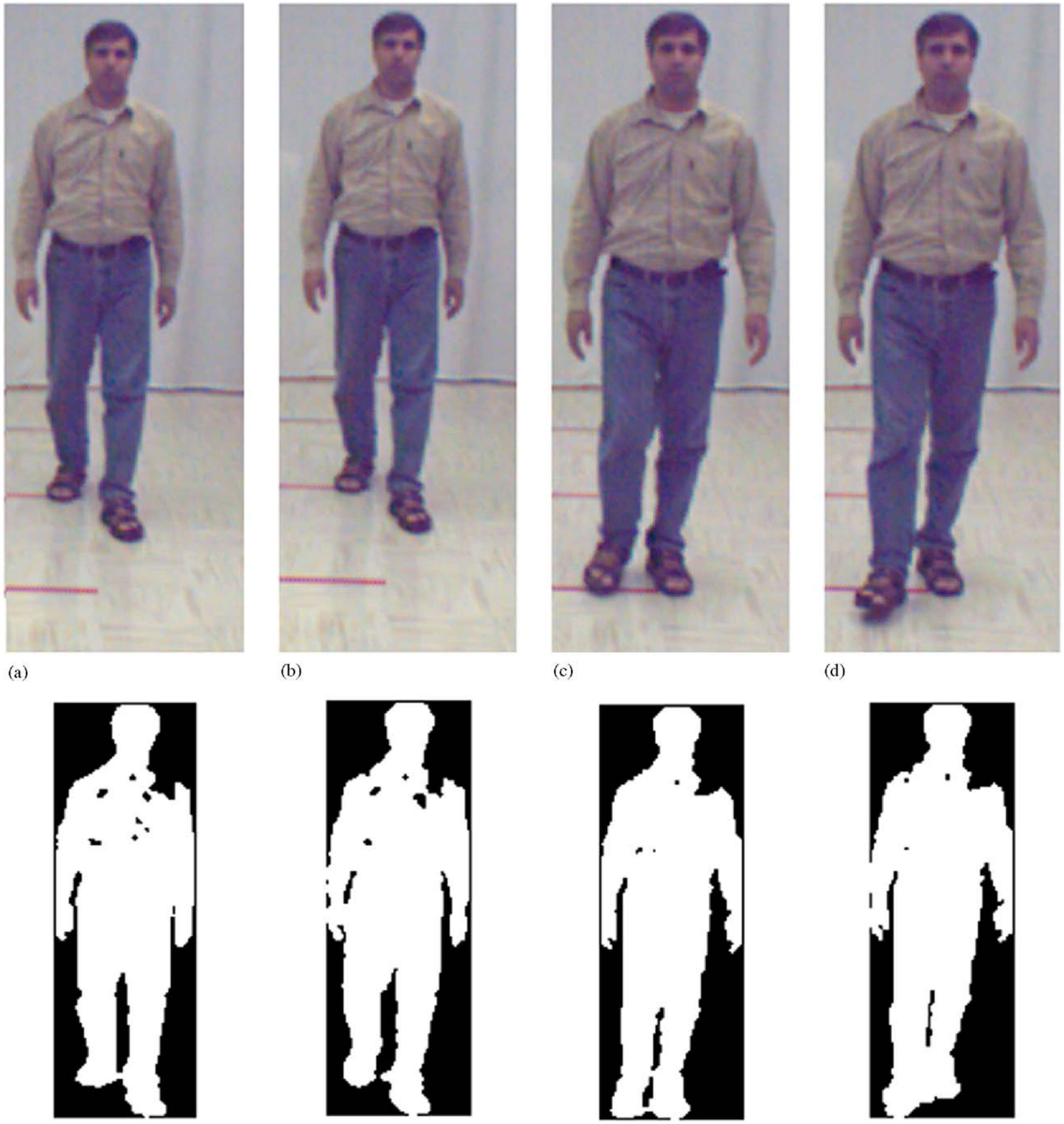

Fig. 5. Morphological distortions introduced by the pre-processing step.

lines. Finally, image cleaning is performed by removing isolated pixels. The final result of morphological processing (see Fig. 7d) contains separable patterns corresponding to cyclic symmetric activities respectively.

As shown in Fig. 7a, an ideal cyclic and symmetric activity is represented by a regular pattern which can be further decomposed into rectangular elementary closed contours. The number of elementary closed contours is related to the number of cycles in the activity. To extract the activity pattern, we implement a region growing technique which fills the spaces enclosed by the elementary contours and produces elementary regions (see Fig. 7e). Global patterns in a pairwise correspondence to cyclic symmetric activities are obtained by merging adjacent elementary regions. A typical example of segmenting the inter-frame similarity matrix into activity patterns is shown in Fig. 7f.

Once the pattern extraction phase is completed, the amount of motion information captured within these

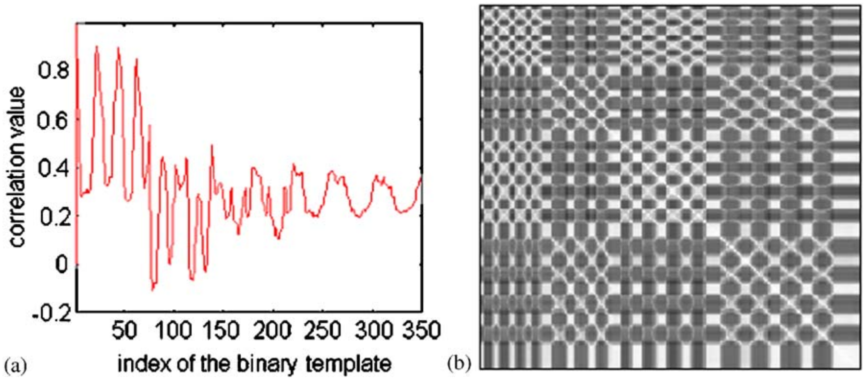

Fig. 6. (a) Sequential 1D cross-correlation plot for a sequence containing four periodic activities with respect to the reference frame no.1; (b) inter-frame similarity matrix computed for the same sequence.

patterns is assessed. The analysis of activity patterns is based on the following observations:

(a) Every cyclic and symmetric activity in the video sequence has a corresponding pattern aligned on the main diagonal of the inter-frame similarity matrix. Therefore, the upper-left and lower-right corners of the bounding box enclosing the pattern correspond to the first and last frames of the activity, respectively. This observation is fundamental for the accurate temporal detection of cyclic symmetric activities in a video sequence. Experimental results on temporal segmentation will be presented and discussed in the next section.

(b) Counting the elementary regions in the activity pattern allows for the computation of the number of complete cycles in the activity. As shown in Fig. 7a, a perfect four-cycle symmetric activity is represented by a pattern containing 24 elementary rectangular regions. The previous statement remains true for activities where the motion speed varies slightly from one cycle to the next.

(c) Patterns not aligned on the main diagonal correspond to similar activities performed at different moments during the same sequence. Fig. 8a shows the inter-frame similarity matrix for a sequence where a human subject alternates two cyclic symmetric activities, namely arm waving and squatting. The detected activity patterns shown in Fig. 8b form a specific configuration. Thus, for arm waving, we detect two main patterns aligned on the main diagonal and two additional patterns centred at the intersection of the Cartesian axes drawn from the main patterns' centres. The same reasoning applies for the patterns associated to squatting. This specific configuration of the activity patterns may allow for the detection of similar activities performed during the same sequence.

3. Experimental results

The database for this study contains real-world video sequences acquired with a monocular camera in an indoor office environment at a frame rate of 30 frames/s. The frame size is 480 × 640 pixels, while the length of the video sequences varies between 170 and 670 frames. Our technique was tested on two types of cyclic and symmetric human activities: (A) controlled motion, such as swinging, squatting, bending etc.; (B) uncontrolled, natural motion, such as walking on quasi-linear paths with different orientations with respect to the camera.

3.1. Temporal segmentation of sequences containing human activities of type A

Activities belonging to type A produce clearly defined patterns in the inter-frame similarity matrix, mainly due to the absence of partial self-occlusion. Eight video sequences in the database correspond to scenarios in which one human subject performs type A activities, such as cyclic aerobic exercises (arm swinging, arm waving, leg bending, and combinations of arm and leg motions) in alternation with walking or standing. Fig. 9a contains a collection of relevant frame samples belonging to a video sequence where four cyclic and symmetric actions occur. The inter-frame similarity matrix corresponding to this sequence is shown in Fig. 9b.

We aim at an accurate temporal extraction of the cyclic and symmetric human activities in each input sequence. The proposed approach successfully detected every cyclic symmetric activity in the database, as well as its temporal boundaries. To evaluate our temporal segmentation approach, ground truth segmentation is to be estimated, since there is a non-negligible inter-observer variability in human segmentation. A statistical estimation of the ground truth was built from ten human segmentations, which independently marked the activity boundaries in the database sequences. The

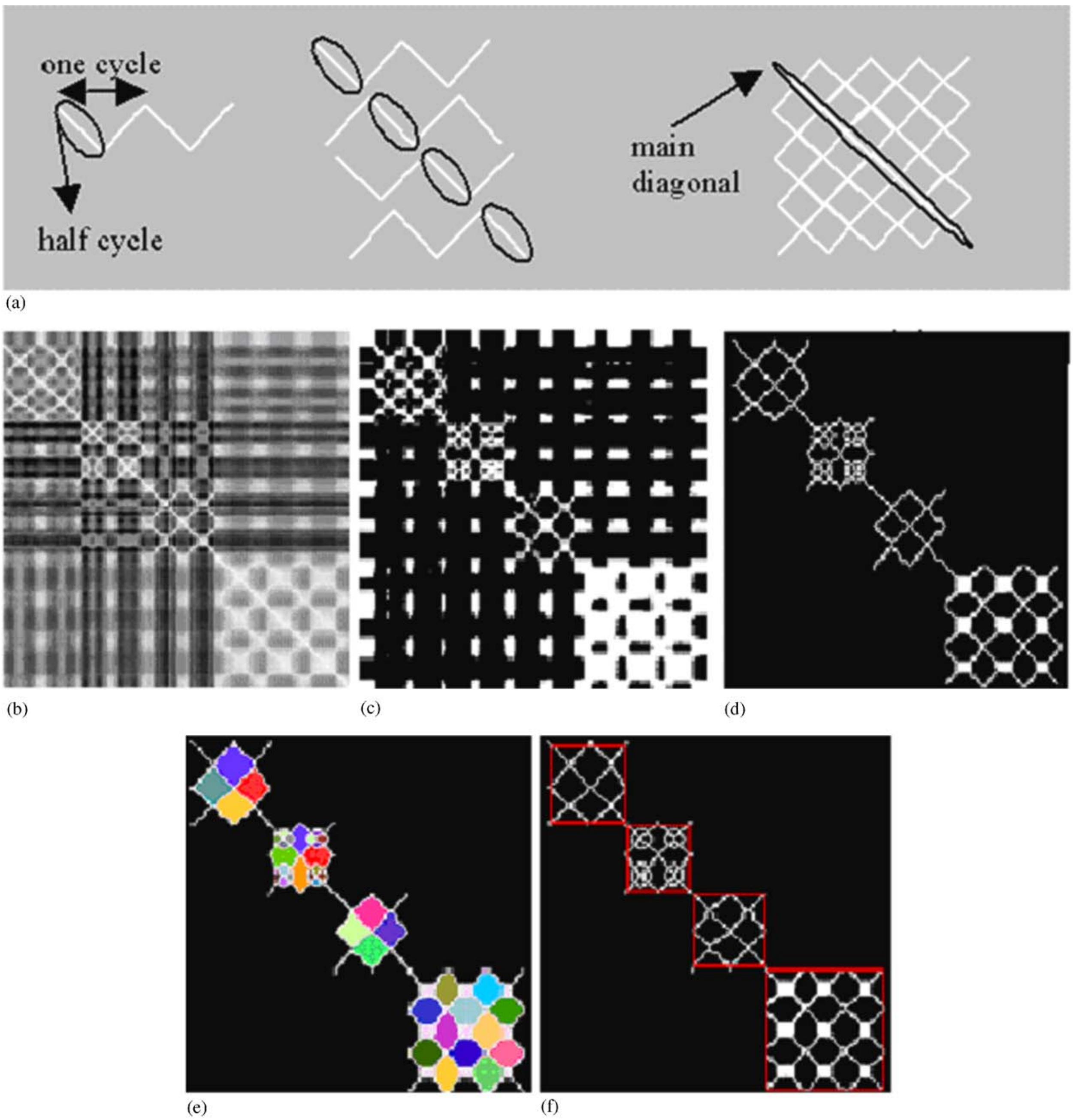

Fig. 7. (a) Pattern forming in the similarity matrix; (b) similarity matrix of a sequence containing four cyclic and symmetric activities; (c) binary image after thresholding; (d) morphological processing; (e) pattern extraction with region growing; (f) bounding boxes for patterns centred on the main diagonal.

distribution of human segmentations for a given sequence is outlined as a histogram of action boundaries versus the frame index. The histograms of action boundaries detected by the human observers for the sequence in Fig. 9a is shown in Fig. 10a. The left and right histogram maxima provide a statistical ground truth estimate for the boundaries of every detected activity.

Next, the performance of our approach with respect to the estimated ground truth (EGT) segmentation is evaluated. In order to provide a quantitative evaluation for a given automatic temporal segmentation S, a *confidence ratio* is defined as follows:

$$C(i, S) = \begin{cases} N(i)/10 & \text{if } S(i) = \text{true}, \\ 1 - N(i)/10 & \text{otherwise}, \end{cases} \tag{3}$$

where $S(i)$ is true if and only if frame i is detected by the evaluated automatic segmentation S as part of the activity, $N(i)$ is the number of human observers that have also detected frame i as belonging to the activity.

The normalization coefficient is set to 10 since 10 human manual segmentations were used in the estimation. The confidence ratio C takes values in the interval [0, 1]. Low values of C(i, S) mean that few observers have made the same decision for frame i as S. Conversely, high values of C(i, S) mean that the majority of observers agree with S. The previously defined evaluation measure allows for comparing the automatic segmentation performed by the proposed algorithm with the EGT segmentation. Fig. 10b shows the plot of the confidence ratio corresponding to the four activities shown in Fig. 9a.

Since the confidence ratio C performs a frame-by-frame evaluation, it conveys interesting information about local errors and about the global robustness of the segmentation as well. Thus, for both considered segmentations (automatic and EGT), there are transitory phases at the beginning and the end of each action. Since the transitory phases are highly similar for both segmentations in the case of every considered action (see Fig. 10b), one may conclude that our algorithm yields an excellent performance for the segmentation of cyclic and symmetric human activities. More results for the evaluation of the proposed method are summarized in Table 1.

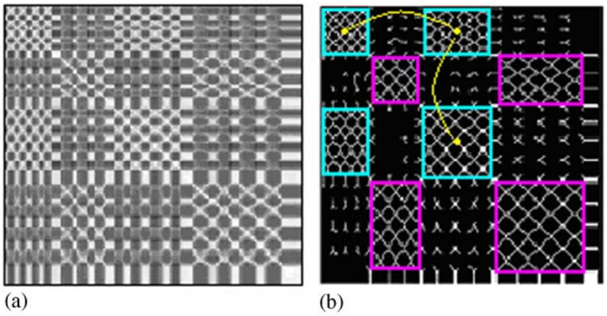

Fig. 8. (a) Inter-frame similarity matrix for a sequence containing two cyclic symmetric activities (arm waving and squatting) performed in alternation. (b) Pattern configuration: squatting patterns are in blue, while arm waving patterns are in pink. Centres of some squatting patterns are shown in yellow. Yellow curves show the relationship between the centres on the main diagonal and those found elsewhere in the image.

The performance evaluation over the experimental database leads to the following conclusions:

(a) The global average boundary detection error is 4.76 frames, which is rather encouraging at a 30 frames/s rate;
(b) Maximal errors (14 frames) were obtained for activities finishing with an incomplete cycle. While the human visual system is able to accurately detect an activity where the last cycle is incomplete, our algorithm does not consider an incomplete cycle as part of the activity.

3.2. Temporal segmentation of walking sequences (activities of type B)

Sequences containing uncontrolled, natural human activities (type B) are also present in the experimental database for this study. This paper reports the results obtained from the analysis of two such typical sequences, where human subjects walk along a piecewise linear trajectory path. Each sequence contains a different human subject following approximately the same trajectory, and inter-subject variability did not affect the performance of the temporal segmentation approach in a significant manner.

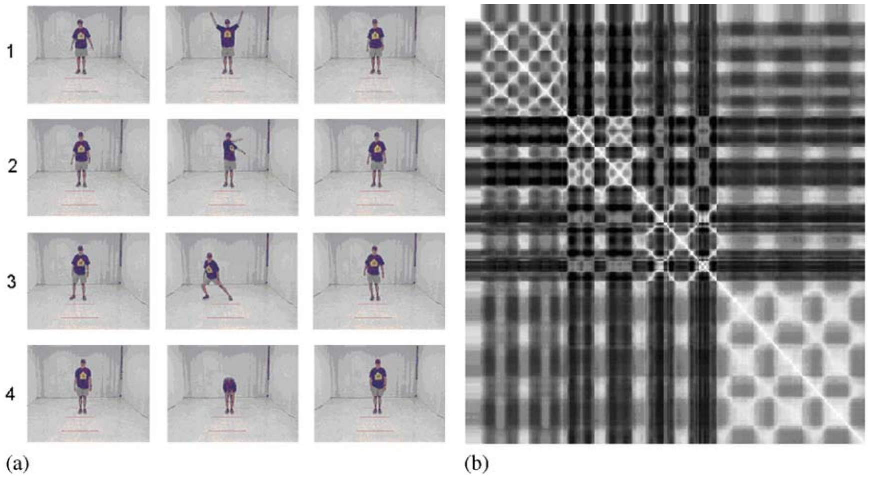

Fig. 9. (a) Relevant frame samples in a sequence containing four cyclic and symmetric activities: (1) arm waving; (2) arm rotation; (3) leg flexing; (4) torso bending; (b) inter-frame similarity matrix for the sequence.

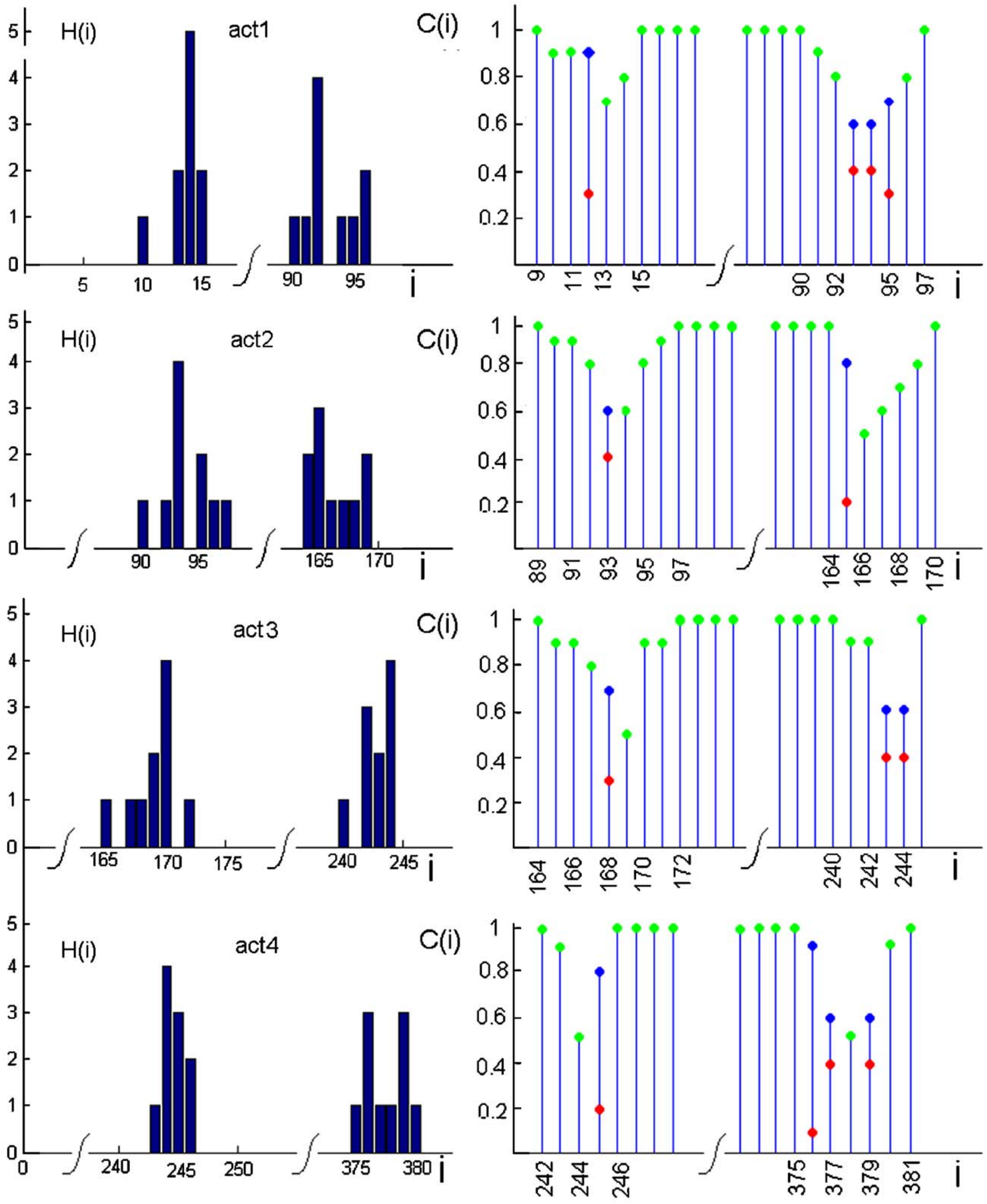

Fig. 10. (a) Histograms of the human boundary detections (*i* is the frame index) for the activities shown in Fig. 9a; (b) confidence ratio versus frame index for automatic and EGT segmentations (🟢 (green) = confounded auto and EGT; 🔴 (red) = EGT; 🔵 (blue) = auto).

The trajectory is comprised of the following segments: (a) perpendicular motion to the camera axis from right to left; (b) the subject moves towards the camera at a 45° angle; (c) perpendicular motion to the camera axis, from right to left; (d) the subject moves away from the camera at a 45° angle; (e) the subject walks towards the camera along the camera axis. The choice of this particular trajectory enables us to identify different walking patterns with respect to the changing point of view and distance to the camera. Due to self-occlusion and artefacts in the pre-processing phase, the walking direction plays a central role in the description of the motion pattern. Fig. 11 contains relevant frame samples corresponding to each linear segment of the trajectory.

The analysis of the inter-frame similarity matrix of the walking sequences led to somehow surprising results with respect to previous work on this topic [5]. Indeed, the walking pattern is cyclic and symmetric only when the person is approaching the camera along the camera axis. Several possible explanations are following:

(a) In previous studies [8], the walking pattern was acquired and analysed using a treadmill. Natural walking is obviously different from treadmill walking, since the speed is not controlled by the motor that provides constant power to the drive belt.

(b) The amount of self-occlusion depends on the walking direction. Due to self-occlusion, walking at a 45° angle with respect to the camera (see Fig. 11b,d) is a cyclic, but not a symmetric motion.

(c) Self-occlusion is symmetric for a person walking perpendicular to the camera axis (see Fig. 11a,c), and therefore it does not interfere with the symmetry of the walking cycle. However, a symmetric pattern was not obtained for this direction. The main reason for this result is the continuously changing angle between the camera axis and the walking subject.

Fig. 12 illustrates the acquisition set-up as well as the maximum angle variation for the perpendicular walking

Table 1
Comparison of the automatic activity segmentation with the corresponding EGT segmentation

Activity	EGT boundaries start frame-end frame (no. of sequence)	Automated boundary detection	Overlap error (in frames)
Waving-1	14–92 (1)	14–92	0
Waving-2	4–114 (2)	6–113	2
Waving-3	3–94 (3)	6–94	3
Waving-4	9–81 (4)	9–81	0
Arm swing	93–165 (1)	94–161	4
Waving and leg lifting	7–77 (5)	9–76	2
Squat-1	10–270 (6)	10–269	1
Squat-2	126–235 (2)	125–234	1
One leg flexing	170–244 (1)	168–245	2
Leg splitting	147–218 (8)	148–218	1
Lateral head rotation	5–97 (7)	18–87	13
Head bending	123–229 (7)	137–220	14
Upfront torso bending	244–376 (1)	248–372	4
Left side torso bending	135–220 (6)	134–217	3
Right side torso bending	270–344 (6)	271–339	5

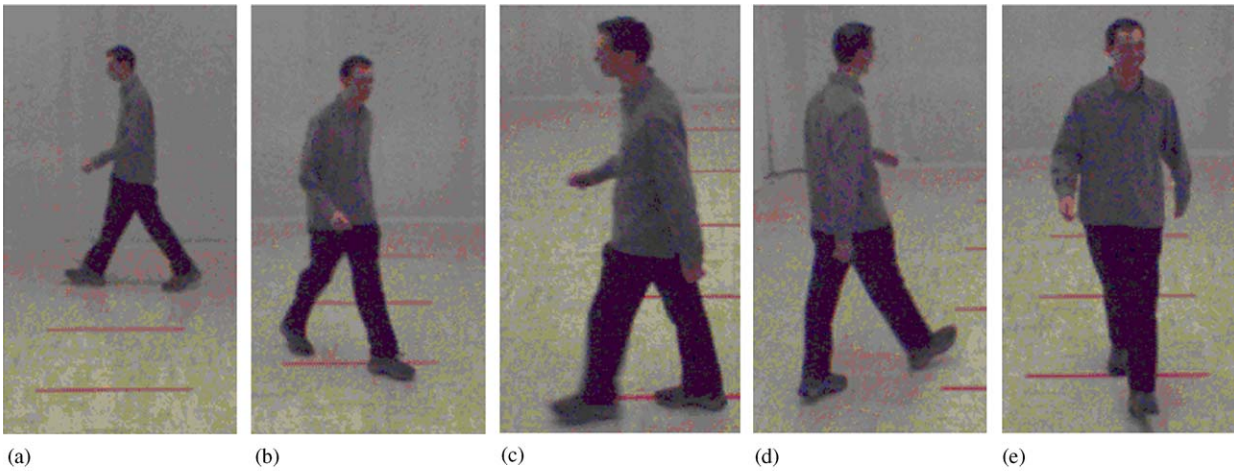

Fig. 11. (a)–(e) Key-frames from a walking sequence along a piecewise linear trajectory.

direction in Fig. 11a. Since the relative angle variation is $\Delta\alpha = \pm 19.63\%$, the initial assumption of a linear walking trajectory is not perfectly verified and therefore the motion pattern is not symmetric. For the third part of the piecewise linear trajectory shown in Fig. 11c, the relative angle variation is even stronger.

The only cyclic and symmetric pattern detected in the two analysed walking sequences corresponds to the frontal walking towards the camera (see Fig. 11e). This result is coherent with the human perception of frontal walking, since there is no self-occlusion in this view, and the walking cycle is symmetric with respect to the alternate limb motion. The detection process of the rectangular pattern corresponding to the frontal walking is detailed in Fig. 13.

The spatio-temporal pattern corresponding to the frontal walking was detected with high accuracy, as shown in Table 2 containing the validation of the automatic segmentation with respect to the EGT segmentation. The EGT segmentation was computed from ten human manual segmentations using the same protocol as in Section 3.1.

3.3. Computation speed

The proposed approach for temporal segmentation was implemented on a 2.66 GHz Pentium IV personal computer with 1024 MB RAM. The total time required for the temporal segmentation of each sequence in the experimental database is shown in Table 3.

The frame size after normalization is uniquely determined for every sequence, since the normalized size of the bounding box is chosen to be the size of the bounding box in the first frame of the given sequence. Thus, the computation time depends both on the length of the sequence and on the normalized frame size. Most of the computation time is spent on the generation of the inter-frame similarity matrix. Ongoing work focuses on algorithmic optimization, as well as on the implementation of a fixed-size (300 frames) sliding temporal window which will allow the real-time implementation of our approach. After optimization, a five- to ten-fold reduction is expected in the maximal computation time, which will result in a reasonable delay in the response of a real-time system.

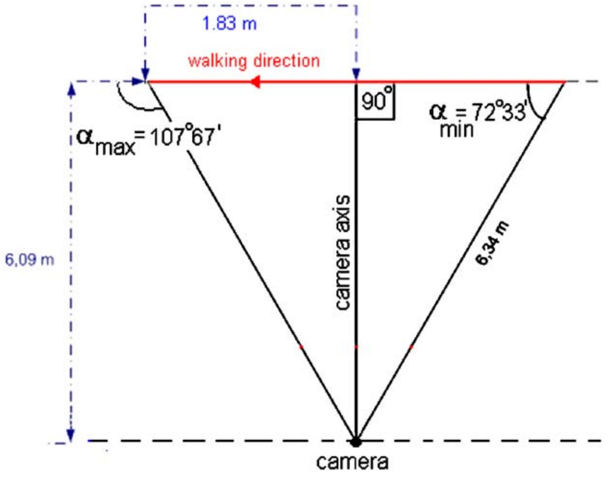

Fig. 12. The geometry of the acquisition set-up. Angle measures for extreme positions of the subject walking perpendicular to the camera axis.

4. Conclusions and future work

This paper deals with segmenting cyclic symmetric human actions from continuous real-world indoor video sequences acquired with a static camera. Specifically, we perform the accurate detection of temporal boundaries for activities such as aerobic exercises and frontal walking. We redefine the concept of inter-frame similarity matrix introduced in [8] and propose a new morphology-based method for extracting relevant motion information from this spatio-temporal template. We have tested our approach on a variety of periodic and cyclic human activities, and provided robust statistical ground truth estimation for the validation of our results. The quantitative evaluation of the proposed approach is based on a new measure, called the confidence ratio, which allows for a precise performance assessment. This confidence ratio will be appropriate for future comparisons of our approach with other temporal segmentation methods.

Ongoing work in our project deals with implementing the concept of inter-frame similarity matrix in real-time. Specifically, we are testing a sliding window of a given frame length to detect any cyclic and symmetric human activities that have just occurred during the currently analysed video sequence. Moreover, we are searching for ways to relax the symmetry constraint, which will allow the analysis of walking or running from several views.

Our research group is also currently working on the integration of robust background subtraction and shadow removal techniques based on Gaussian Mixture Models [14] with our proposed temporal segmentation method.

One of the main practical contributions of our approach is the accurate detection of frontal walking. The extraction of the spatio-temporal pattern corresponding to a human subject walking towards the camera may be embedded in real-time surveillance applications. Since face detection is feasible in this particular view, the real-time detection of frontal walking may be used as a trigger for a face identification system.

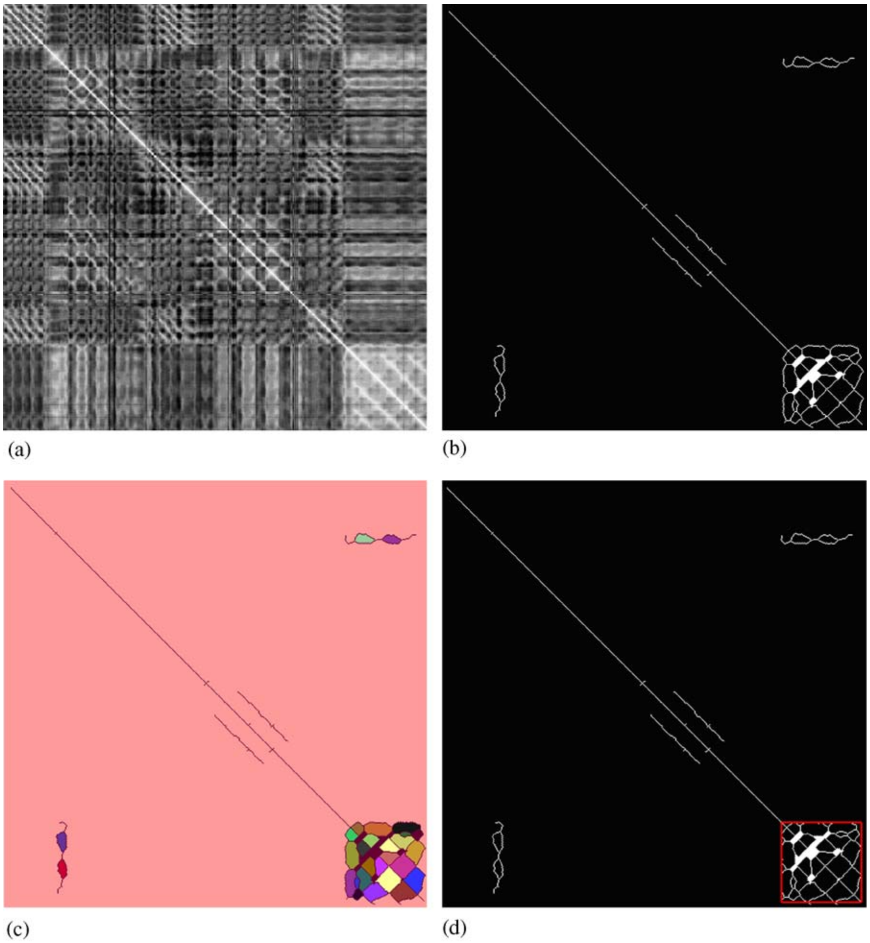

Fig. 13. (a) Inter-frame similarity matrix for the sequence shown in Fig. 11; (b) binary result of thresholding; (c) morphological processing and region growing; (d) red bounding box for the extracted pattern.

Table 2
Comparison of the automatic activity segmentation with the corresponding EGT segmentation

Activity	EGT boundaries start frame–end frame	Automatic boundary detection	Overlap error (in frames)
Walking sequences			
Frontal walking-1	552–660	553–660	1
Frontal walking-2	530–644	533–644	3

Table 3
Computation time for the sequences in the experimental database

Sequence no. (type)	Total number of frames	Frame size after normalization	Computation time (s)
1. (A)	174	73 × 177	5.92
2. (A)	105	77 × 184	2.27
3. (A)	444	75 × 217	45.12
4. (A)	227	79 × 246	13.54
5. (A)	273	38 × 114	7.84
6. (A)	360	67 × 192	25.53
7. (A)	237	216 × 187	25.97
8. (A)	311	105 × 299	37.38
9. (B)	511	40 × 149	34.97
10. (B)	471	68 × 160	38.55

Acknowledgements

Research reported in this paper was supported by two Discovery Grants from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Gavrila MD. The visual analysis of human movement: a survey. Computer Vision and Image Understanding 1999;73(1):82–98.

[2] Moeslund TB, Granum E. A survey of computer vision-based human motion capture. Computer Vision and Image Understanding 2001;81(3):231–68.

[3] Tsai P, Shah M, Keiter K, Kasparis T. Cyclic motion detection for motion-based recognition. Pattern Recognition 1994;27(12): 1591–603.

[4] Polana R, Nelson R. Detecting activities. Journal of Visual Communication and Image Representation 1994;5(2):172–80.

[5] Polana R, Nelson R. Detection and recognition of periodic, non-rigid motion. International Journal of Computer Vision 1997;23(3):261–82.

[6] Seitz SM, Dyer CR. View-invariant analysis of cyclic motion. International Journal of Computer Vision 1997;9(3):1–23.

[7] Bobick AF, Davis JV. The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001;23(3):257–67.

[8] Cutler R, Davis LS. Robust real-time periodic motion detection, analysis, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;22(8): 781–96.

[9] BenAbdelkader C, Cutler R, Davis LS. Motion-based recognition of people in eigengait space. Proceedings of IEEE international conference on automatic face and gesture recognition. vol. 1; 2004, p. 352–7.

[10] Yazdi M, Branzan-Albu A, Bergevin R. Morphological analysis of spatio-temporal patterns for the segmentation of cyclic human activities. Proceedings of the 2004 IEEE international conference on pattern recognition (ICPR2004) vol. 4, 2004. 23–26 August Cambridge, UK, p. 240–243.

[11] Stauffer C, Grimson WE. Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;22(8):747–57.

[12] Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 1979;9(1):62–6.

[13] Horprasert TR, Harwood D, Davis L. A statistical approach for real-time robust background subtraction and shadow detection. In: Proceedings of IEEE ICCV'99 frame rate workshop, Corfu, Greece, 1999.

[14] Martel-Brisson N, Zaccarin A. Moving cast shadow detection from a gaussian mixture shadow model. Proceedings of IEEE international conference on computer vision and pattern recognition CVPR 2005, San Diego, CA, USA, 20–25 June 2005.