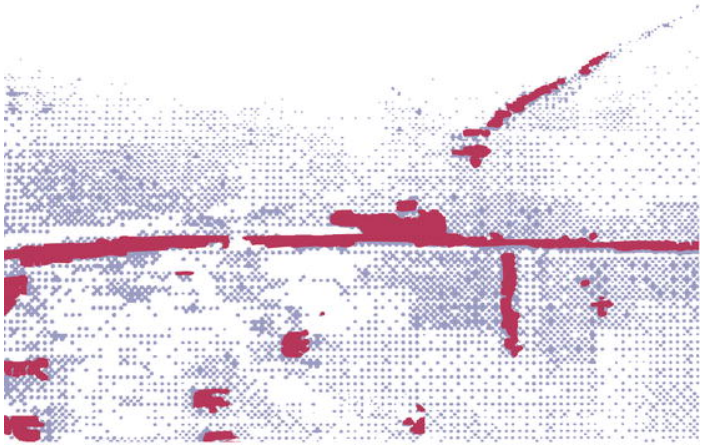
Volume 29, issue 1 1 January 2008 ISSN 0167-8655

ELSEVIER

Pattern Recognition Letters

An official publication of the
International Association for Pattern Recognition

IAPR



Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 81–89
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Recognition of human behavior by space-time silhouette characterization

Arash Mokhber *, Catherine Achard, Maurice Milgram

Institut des Systèmes Intelligents et Robotique (ISIR), Université Pierre et Marie Curie, 4 Place Jussieu, Case Courrier 252, 75252 Paris Cedex 05, France

Received 20 June 2006; received in revised form 20 February 2007
Available online 20 September 2007

Communicated by H. Wechsler

Abstract

In this study, a method for human action recognition is proposed. Only one camera is used, without calibration. Viewpoint invariance is obtained by several acquisitions of the same action. The originality of the method consists in characterizing each sequence globally, to enhance the robustness. After detection of moving areas throughout each image, a binary volume is obtained, composed by all the silhouettes of the moving person. This space-time volume is characterized by a vector of its 3D geometric moments. These moments are normalized to be invariant to the position, scale and duration of actions. Action recognition is then carried out using a nearest neighbor classifier based on Mahanalobis distance. Results are presented on a base of 1614 sequences performed by seven persons and categorized in eight actions.

Keywords: Behavior recognition; Action recognition; Motion analysis

1. Introduction

The recognition of human activity has received much attention from the computer vision community and has led to several surveys (Gavrila, 1999; Wang and Singh, 2003; Hu et al., 2004). It leads to modern applications such as video surveillance for security, human–computer interaction, entertainment systems and monitoring of patients or old people, in hospitals or in their homes. The different existing approaches can be divided in four categories: (i) 3D approaches without shape model, (ii) 3D approaches with volumetric models such as elliptical cylinders, (iii) 2D approaches with explicit shape model such as stick figures and 2D ribbons and (iv) 2D approaches without explicit shape model. Since human body is not a rigid object and may present a multitude of shapes and postures even for the same person, a robust modeling is difficult to obtain. Appearance models are therefore preferred over geometric models, in most cases. Recognizing human action can then be considered as the issue of classifying time varying data. This can be carried out by matching a request sequence with a set of labeled sequences which represent typical actions. To perform this task, actions can be characterized either globally, or as a temporal chain of local features.

1.1. Global representation of sequences

This representation has the advantage of not considering sequences as temporal objects. Therefore, one action is represented by only one feature vector, computed on the whole sequence. This point allows the use of simple similarity measurements (e.g. Mahanalobis distance), to recognize the action's label. Using this approach, Bobick and Davis (2001) characterize each action with (i) a binary motion-energy image (MEI), representing locations where motion has occurred through the sequence and (ii) a motion-history image (MHI) where intensity is a function of recency of

* Corresponding author. Tel.: +33 1 44 27 62 17; fax: +33 1 44 27 75 09.
E-mail addresses: arash.mokhber@lisif.jussieu.fr (A. Mokhber), achard@ccr.jussieu.fr (C. Achard), maum@ccr.jussieu.fr (M. Milgram).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.08.016

motion at each pixel. A statistical model of the 7 Hu moments (Hu, 1662) is then generated over a set of MEIs and MHIs. For the recognition of actions, the Mahanalobis distance is estimated between the moment description of the query and those of the known actions. Davis (1998) also makes use of MHIs and characterize actions by multiple histograms of local motion orientations to recognize movements. As an extension, Weinland et al. (2005) introduce motion history volumes (MHV) as a free-viewpoint representation for human actions. They propose methods to align and compare MHVs of the different actions to learn and recognize basic human action classes. Ke et al. (2005) extract volumetric features on the whole video sequence optical flow field. They propose an extension of Viola and Jones method by generalizing 2D box image features to 3D spatio-temporal volumetric features. This permits event detection and action classification in video data. Efros et al. (2003) make use of motion descriptors based on optical flow measurements in spatio-temporal volumes and utilize an associated similarity measure to recognize human actions at lower resolutions. Shetman and Irani (2005) propose an extension of the two-dimensional image correlation into three-dimensional space-time video volumes correlation. With this similarity measure they are able to detect occurrences of given behavior in video sequences. As in object recognition, Dollar et al. (2005) characterize actions by detecting and using sparse informative feature points (e.g. space-time corners) in the three-dimensional (x, y, t) video data. Volumes of 3D gradient fields surrounding these feature points are used for the recognition. In Chomat and Crowley's study (1999), local spatio-temporal appearance is used for a probabilistic recognition of activities. Joint statistics of space-time filters are employed to define histograms which characterize the actions to recognize. These histograms provide the joint probability density functions required for the recognition using Bayes rule. Zelnik-Manor and Irani (2001) consider events as long term temporal objects and characterize them by spatio-temporal features at multiple temporal scales. Based on this assumption, they use a simple statistical distance measure between video sequences to isolate and cluster events within long sequences.

1.2. Sequence modeling as temporal objects

In the previously mentioned methods, actions were considered globally and not as a temporal set of images. This allows obtaining robust features and using simple distance measures to recognize actions which are represented by only one feature vector. A drawback of these methods is the difficulty to segment video sequences into action consistent parts, which is also very time-consuming. The following approaches consider sequences as temporal sets of local features. Li and Greenspan (2005) recognize and estimate the scale of time-varying human gestures by exploiting the changes in silhouette contours along spatio-temporal directions. Contours are thus parameterized and their evolution is considered as a temporal set. Dynamic time warping (DTW) and mutual information are then employed to match and recognize models. Pierobon et al. (2005) also extract features directly from 3D data (x, y, t) which makes the system insensitive to viewpoint. Frame by frame descriptions, generated from gesture sequences are then collected and compared using DTW.

Martin and Crowley (1997) propose a system for hand gesture recognition composed of three modules including tracking, posture classification and gesture recognition by a set of finite state machines. Cupillard et al. (2004) also use finite state automatons for recognizing sequential scenarios for metro surveillance. For composed scenarios, they employ Bayesian networks proposed by Hongeng et al. (2000). Other researchers prefer to use Hidden Markov Models (HMM) (Rabiner, 1990) which may be a useful tool for the recognition of patterns of variable durations. Yamato et al. (1992) developed the first HMM-based gesture recognition systems to distinguish between six tennis strokes. Sun et al. (2004) estimate motion parameters at each frame of video sequences and compute their likelihood. Then, they characterize and recognize actions using continuous HMMs. Starner et al. (1998) proposed a real-time HMM-based system for the recognition of sentence level American Sign Language (ASL) without explicit model of fingers. Ogale et al. (2005) use a different approach by representing human actions as short sequences of atomic poses. After the extraction of these atomic poses from multiview video sequences they build a probabilistic context free grammar (PCFG). The PCFG is used to analyze video sequences and recognize actions within.

The method presented in this work is based on a global characterization of sequences proposed by the same authors (Mokhber et al., 2005). The method described in this earlier study was however prone to less accurate parameter estimation, due to a bias in the database separation. The presented database has therefore been modified. Similarly, only one camera is employed, without calibration. Invariance to viewpoint is obtained by several acquisitions of the same actions. Binary silhouettes are obtained by motion detection and joined together to form a volume. This space-time shape is then utilized and characterized by its geometric 3D moments which form a feature vector for each sequence. These moments are invariant to the position, scale, and temporal duration of actions. The feature vector is then employed in a nearest neighbor framework for the recognition of actions. Significant improvement of the method and further experiments are also performed, compared to the previous approach. An extension of the system is also proposed, by employing optical flow measurements. This extension leads to slightly inferior results but permits to avoid the step of motion detection.

Another advantage of the presented approach is that it takes into account the global motion of objects through the sequences. Most of the existing methods (e.g. Bobick and Davis, 2001; Efros et al., 2003) only consider the relative movement of body parts. Authors often study with relative motion actions or employ methods where the

sequences are compensated for the global translation of the moving object (so that the global motion is not considered). Ke et al. (2005) or Shetman and Irani (2005) do not utilize such compensation but they only consider one class of event. Their methods are more similar to detection processes (which consists of detecting one pattern of motion through a sequence) than recognition ones. The approach we expose here has the ability to discriminate between different types of actions and between different global motions (e.g left to right vs. right to left). This helps to differentiate actions such as "to walk" and "to jump" which are clearly distinguishable from their global direction of motion.

2. Motion detection

The first step of the activity recognition process consists of detecting moving pixels throughout sequences. This can be avoided by the use of optical flow measurement (cf. Section 6.3) which can also be considered as space-time sequential shapes. For the detection, the current image is compared at any given time to a reference image that is continuously updated. It is also necessary to remove shadows that are eventually present in the scene. To authorize multi-modal backgrounds, the history of each pixel of the reference image is modeled by a mixture of K Gaussian distributions (Porikli and Tuzel, 2003; Stauffer and Grimson, 1999). The probability of observing the value of the current pixel $\mathbf{X}_t$ is then given by

$$P(\mathbf{X}_t) = \sum_{i=1}^{K} w_{i,t} \cdot N(\mathbf{X}_t, \boldsymbol{\mu}_{i,t}, \boldsymbol{\Sigma}_{i,t}) \tag{1}$$

where for ith Gaussian at time t, $w_{i,t}$ is the weight of the Gaussian, $\boldsymbol{\mu}_{i,t}$ is its mean vector and $\boldsymbol{\Sigma}_{i,t}$ its covariance matrix. $N(\,)$ is the Gaussian probability density function defined as followed:

$$N(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right] \tag{2}$$

where n is the dimension of the vector. In this study n is equal to 3 because we chose to work with color images with three channels (RGB). Initialization of the Gaussian mixture is carried out by the K-means algorithm on the first images of the sequence where it is assumed that no movement occurs. Each pixel of the background is modeled by

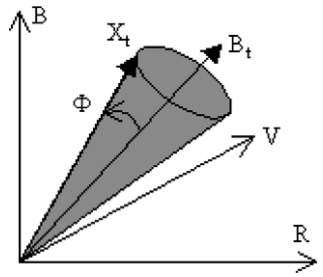


Fig. 1. Shadow is defined as a cone.

$K = 2$ Gaussians. It appears that this is a reasonable compromise between the computing time and the quality of results. For each new pixel X_t, its nearest Gaussian is searched. If the distance between this Gaussian and the current pixel is less than a threshold value, the latter is assigned to the background. Otherwise, it is classified as a pixel belonging to a moving object. To consider lighting changes during the process of acquisition, the pixels labeled as background are used to update the reference image and thus the Gaussian they are closest to:

$$\begin{aligned} \boldsymbol{\mu}_t &= (1-\alpha)\boldsymbol{\mu}_{t-1} + \alpha\mathbf{X}_t \\ \boldsymbol{\Sigma}_t &= (1-\alpha)\boldsymbol{\Sigma}_{t-1} + \alpha(\mathbf{X}_t - \boldsymbol{\mu}_t)(\mathbf{X}_t - \boldsymbol{\mu}_t)^{\mathrm{T}} \end{aligned} \tag{3}$$

where α is empirically set to 0.1. This method leads to reasonably good detection results. However, shadows are often detected as a moving object. As a result, the shapes of the detected silhouettes are significantly deteriorated and disturb the algorithm of action recognition. A second stage is employed to address this issue. In this work it is assumed that shadows decrease the brightness of pixels but do not affect their color, as proposed by Porikli and Tuzel (2003). Thus, the angle Φ between the color vector of the current pixel X_t and that of the corresponding background pixel B_t, (mean of the nearest Gaussian), is an effective parameter to detect shadows. Note that if Φ is below a threshold value, and the brightness of the current pixel is lower than the brightness of the background, it is assumed that the pixel corresponds to shadow. Therefore, shadow is defined as a cone around the color vector corresponding to the background, as shown in Fig. 1. At the end of the process, only pixels detected as moving by the mixture of Gaussian and which do not correspond to shadows are preserved. Several morphological operations end this stage and lead to a binary map of moving pixels, for each image.

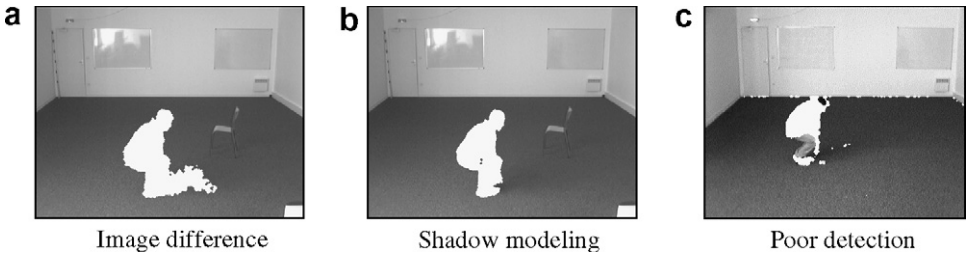


Fig. 2. Typical results for poor and good motion detection.

As can be seen in Fig. 2a and b, fairly good detection results are obtained. However, as presented in Fig. 2c, for some images of these sequences, the detection is not as clear. This is due to the close similarities between background colors and those of the moving person. Nonetheless, the space-time characterization of these binary images, presented in Section 3, is robust enough to lead to quite acceptable action recognition results.

It may be observed through the displayed figures that the background image is relatively simple. However, for more complicated backgrounds, the detection of moving areas would be as efficient as here. The only condition for obtaining acceptable results, is that the background color should be different than the color of the moving object (complicated backgrounds are therefore even more suitable for this process).

3. Characterization of space-time silhouettes

Features representative of the sequence are extracted from all the binary images obtained by the detection process. Our initial idea was to represent each binary silhouette by its geometrical two-dimensional moments. These moments were normalized to obtain invariance in translation and scaling. However, an important piece of information appears to be missing after the normalization. This is due to the fact that actions which represent a person walking from the left to the right, from the right to the left, or approaching the camera have similar feature vectors. For example, when a person approaches the camera the change in height that characterizes this action disappears during normalization. To avoid this problem and obtain robust features, we chose to work with global "space-time volumes" composed by the binary silhouettes extracted from each sequence (i.e. all moving points of the sequence). To form these volumes, all the binary images obtained by motion detection for one action are concatenated together in chronological order (according to the temporal axis). Fig. 3 presents a three-dimensional view of such a volume for the action "to crouch down".

These volumes are characterized by their three-dimensional geometric moments.

Let $\{x, y, t\}$ be the set of points belonging to the binary "space-time volume" $V(x, y, t)$, where x and y represent the space coordinates and t, the temporal coordinate. The moment of order $(p + q + r)$ of this volume is determined by:

$$A_{pqr} = E\{x^p y^q t^r\} = \frac{\int\int\int V(x,y,t) x^p y^q t^r \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t}{\int\int\int V(x,y,t)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t} \quad (4)$$

where $E\{x\}$ represents the expectation of x. In order to work with features invariant in translation, the central moments are considered, as follows:

$$Ac_{pqr} = E\{(x - A_{100})^p (y - A_{010})^q (t - A_{001})^r\} \quad (5)$$

These moments must also be invariant to the scale to preserve invariance with the distance of action or with the size of people. A direct normalization on the different axes, by dividing each component by the corresponding standard deviation is not desirable because it leads to an important loss of information, that is, the shape of the binary silhouettes appears to be spherized. Hence, an identical normalization is carried out on the first two axes, while the third (time) is normalized, separately. The normalization performed by preserving the ratio of width-to-height of the binary silhouettes is thus obtained by the following relation:

$$M_{pqr} = E\left\{ \left(\frac{x - A_{100}}{Ac_{200}^{1/4} Ac_{020}^{1/4}} \right)^p \left(\frac{y - A_{010}}{Ac_{200}^{1/4} Ac_{020}^{1/4}} \right)^q \left(\frac{t - A_{001}}{Ac_{002}^{1/2}} \right)^r \right\} \quad (6)$$

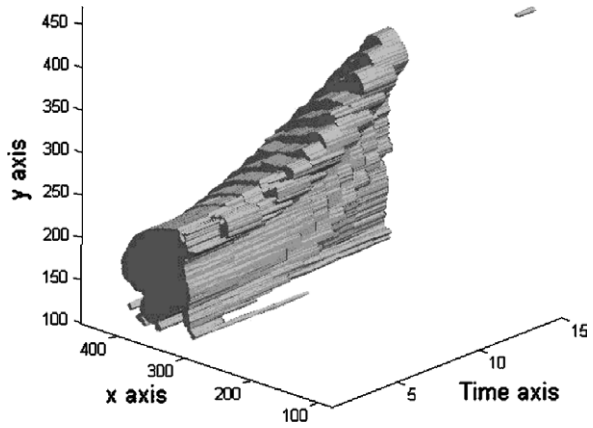


Fig. 3. Space-time binary volume for a typical action.

4. Non-binary silhouettes

Since the space-time volume $V(x, y, t)$ has only binary values, all space-time points used to compute the expectation (i.e. the moment) have the same weight. Using this method every motion-detected point has the same importance for the moment estimation. This includes details in the silhouette and false detected points due to noise or poor motion detection. The purpose of the present study is to consider the global movement of a person's body. For example, when a person sits down, his silhouette becomes smaller and closer to the ground. It is not desirable for us to give importance to the relative motion of small body parts such as hands or even arms and legs. It would therefore be interesting to advantage the points that are located near the silhouette center (i.e. those that characterize more global motion) over those that are closer to the boundaries. For this purpose new space-time volumes $V_2(x, y, t)$ and $V_3(x, y, t)$ are constructed, which are non-binary. Based on V, V_2 is computed by assigning to every pixel, its distance to the nearest background pixel in the same frame. If the pixel already corresponds to background, it has zero-value. V_3 is built by assigning to each pixel its distance to the nearest background pixel in the whole volume. Fig. 4b and c present the result of such operations. The values of these new space-time volumes may be used to compute the vector of moments. Accordingly, the weight of space-time points is no longer uniform in the expectation estimation. It is proportional to the value

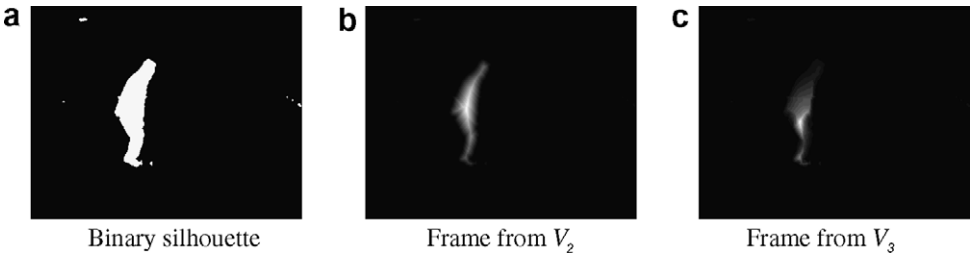


Fig. 4. Non-binary silhouettes obtained by distance computing.

of V_2 or V_3 at any considered point. In Section 6.3, the geometric moments are also estimated over the norm and squared norm of optical flow volumes and results are compared.

5. Presentation of the sequence database

A sequence database comprising eight actions is considered:

(1) "to crouch down",
(2) "to stand up",
(3) "to sit down",
(4) "to sit up",
(5) "to walk",
(6) "to bend down",
(7) "to get up from bending", and
(8) "to jump".

Various viewpoints were acquired for each action: front, 45°, 90°, −45° and −90°. Each action was executed by seven people, and repeated 230 times on average. The database comprises 1614 sequences. Presented below, are some examples of images of the database representing various actions and silhouettes of actors (Fig. 5).

The scale of actions (i.e. the distance with respect to the camera) may also change from one sequence to the other. Fig. 6 presents the binary images obtained when a same person performs the same action at two different scales.

For each action, a vector of features composed of the 14 moments of 2nd and 3rd order is considered:

$$O = \{M_{200}, M_{011}, M_{101}, M_{110}, M_{300}, M_{030}, M_{003}, M_{210}, M_{201}, M_{120}, M_{021}, M_{102}, M_{012}, M_{111}\}$$

Fig. 5. Sample images from the sequence database.

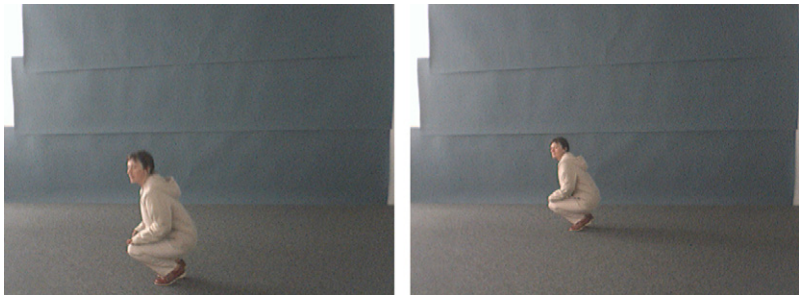


Fig. 6. A person performing the same action at two different scales.

Note that the moment M_{020} is not calculated. This is due to the normalization which makes M_{020} inversely proportional to M_{200}. In addition, the moment M_{002} is always equal to 1.

6. Recognition results

For the recognition, the sequence database is divided into two disjointed sets: (1) a reference database, and (2) a test database. To test invariance of the method compared to people's morphology, the reference database is made up of actions carried out by six people. The sequences achieved by the last person were assigned to the test database. The recognition is done by searching for the vector of features corresponding to the query action the nearest vector in the reference database, using the Mahalanobis distance. The action to be recognized is then assigned to the class of this nearest vector.

6.1. Recognition with binary volumes

Raw binary volumes are initially used to compute the vector of moments and recognize actions. Presented in Table 1, are the seven recognition rates obtained by placing each of the seven persons in the test database one by one. The average recognition rate on the seven persons is also presented.

The average recognition rates, on the eight actions vary from 77.1% to 97.2%, depending on the person. Thus, one may conclude that actions are well recognized. Note that the person present in the test database is not at any time present in the reference database. This shows that the characterization is relatively invariant to the silhouette of the person. The worst recognition rate (77.1%) is obtained for person 7. This is not surprising because this person presents a particular binary silhouette due to her clothing, as shown in Fig. 7. This person wears a long skirt (and it is the only person with a skirt in the base). In spite of this characteristic, the recognition rate is still acceptable, which demonstrates that the global characterization of actions is robust. An extension of the number of actors in the base is envisaged in order to improve classification results. In Table 2, the confusion matrix obtained by averaging the seven confusion matrices corresponding to the different people, is presented. The most poorly recognized action is action 4 (''to sit up'') sometimes confused with action 7 (''to get up from bending'') which is a nearly similar action. Other actions such as (''to crouch down'' and ''to walk'') are significantly well recognized (with recognition rates of 97.2% and 98.4%, respectively). The recognition rate for the remaining actions is acceptable.

6.2. Recognition with ''distance'' volumes

The distance to background is computed on the obtained silhouettes, thus leading to non-binary space-time volumes. The vector of features is computed on each volume and used for the recognition. Tables 3 and 4 present the average recognition rates per person and the average confusion matrix that are obtained by computing the distance to the background on the same frames (V_2). As can be seen, the results are slightly improved. The average recognition rate on the seven persons is 90.0% (w.r.t. 89.5% for the previous method). It is assumed that the improve-

Table 1
Recognition rates using binary volumes

Person	1	2	3	4	5	6	7	Average
Rate	89.9	90.2	82.7	97.2	92.1	95.2	77.1	89.5

Table 2
Average confusion matrix using binary volumes

	1	2	3	4	5	6	7	8
1	97.2	0.0	0.0	0.0	0.0	2.8	0.0	0.0
2	0.0	90.5	0.0	0.0	0.0	0.0	9.5	0.0
3	4.8	0.0	84.1	0.0	0.0	9.4	1.2	0.5
4	0.0	3.3	0.0	76.1	0.0	0.0	17.1	3.6
5	0.0	0.0	0.0	0.0	98.4	0.9	0.2	0.4
6	11.1	0.0	0.0	0.0	0.0	88.3	0.0	0.5
7	0.0	10.7	0.0	0.7	0.0	0.0	88.2	0.3
8	0.0	8.6	0.0	0.0	0.0	1.7	0.9	88.8

Table 3
Recognition rates using same frame distance

Person	1	2	3	4	5	6	7	Average
Rate	92.6	88.3	88.1	93.1	89.8	96.2	77.1	90.0

Fig. 7. Detection of a particular silhouette.

Table 4
Average confusion matrix using same frame distance

	1	2	3	4	5	6	7	8
1	94.6	0.0	0.0	0.0	0.0	5.4	0.0	0.0
2	0.0	93.6	0.0	0.5	0.0	0.0	6.0	0.0
3	9.9	0.0	77.3	0.0	0.0	10.6	0.6	1.6
4	0.0	7.9	0.0	76.3	0.0	0.0	15.3	0.5
5	0.0	0.0	0.2	0.0	99.4	0.0	0.0	0.4
6	6.8	0.0	0.7	0.0	0.0	92.5	0.0	0.0
7	0.0	10.1	0.0	3.5	0.0	0.0	86.4	0.0
8	0.9	5.0	0.0	0.0	0.0	0.0	0.0	94.1

ment is indeed due to the noise reduction and the characterization of more global motions provided by this new space-time volume. Person 7 still yields the lower rate, and confusion stands between the same actions, which is not surprising. The results obtained by employing the distance to background on the whole volume (V_3) are also similar but less enhanced (Tables 5 and 6). As shown in Fig. 4c and b, V_3 is less accurate than V_2. The latter gives a strong weight to the center of silhouettes and thus advantages the motion of the centroid of the person. On the other hand V_3 gives more importance to the 3D center of space-time volumes and do not emphasize the center of silhouettes at each frame. Nonetheless, the average recognition rate stands at 89.7%.

Table 5
Recognition rates using whole volume distance

Person	1	2	3	4	5	6	7	Average
Rate	95.8	85.9	85.7	94.4	88.2	97.1	74.0	89.7

Table 6
Average confusion matrix using whole volume distance

	1	2	3	4	5	6	7	8
1	96.7	0.0	0.0	0.0	0.0	3.3	0.0	0.0
2	0.0	92.4	0.0	0.5	0.0	0.0	7.1	0.0
3	8.6	0.0	82.3	0.0	0.0	6.9	0.6	1.6
4	0.0	5.5	0.0	76.0	0.0	0.0	17.9	0.6
5	0.3	0.0	0.2	0.0	98.6	0.0	0.0	0.9
6	8.9	0.7	0.0	0.0	0.0	90.5	0.0	0.0
7	0.0	10.0	0.0	2.8	0.0	0.0	86.9	0.3
8	0.0	8.1	0.0	0.0	0.0	3.1	0.0	88.8

6.3. Recognition with optical flow volumes

Another solution for the action recognition problem consists of employing optical flow measurements. This method has the advantage of avoiding the step of motion detection. Optical flow (Lucas and Kanade, 1981) is therefore computed over every sequence, which characterizes motion speed and direction at every pixel of every frame. Results will be compared to those obtained with binary silhouettes of motion detection.

As moments have to be calculated over positive scalar functions the norm V and squared norm V^2 of flow vectors are utilized. If V_x and V_y are the horizontal and vertical flow channels at pixel (x, y) and time t, V is defined by:

$$V(x,y,t) = \left(V_x^2 + V_y^2\right)^{1/2}(x,y,t) \tag{7}$$

and is used as space-time volume to compute the feature moments. In this method, higher weight is given to pixels that move faster. Fig. 8 presents the map of V and V^2 at key frames of action "to crouch down".

Tables 7 and 8 present the recognition rates. The results are deteriorated compared to those obtained with binary or distance volumes. The average recognition rate is 85.8% when employing V as weighting function and 85.9% when employing V^2. A difference that should be noted on these last results is that person 7 presents similar results as the other persons. The characterization is therefore not as similar to binary silhouettes as the "distance to background" characterization was. In addition, since faster pixels have more important weights, the centers of the silhouettes are not privileged in the moment estimation. By avoiding the step of motion detection, this method however has the

Table 7
Recognition rates using optical flow norm

Person	1	2	3	4	5	6	7	Average
Rate	86.9	78.2	89.3	93.8	80.3	87.6	76.0	85.8

Table 8
Recognition rates using optical flow squared norm

Person	1	2	3	4	5	6	7	Average
Rate	90.2	75.2	89.0	91.3	81.1	84.8	87.5	85.9

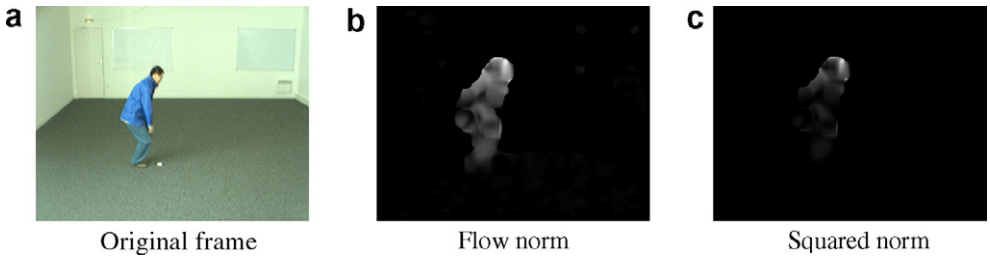


Fig. 8. Optical flow vector norm and squared norm images. (a) Original frame; (b) flow norm; (c) squared norm.

counterpart of requiring that only one person is present in the scene.

7. Discussion

As mentioned before, an advantage of the global representation of sequences is to avoid temporal representation. Actions are therefore represented by only one vector, which permits to use simple measurements to determine the similarity between actions and recognize them. Furthermore, one can presume that this global characterization is even more robust than temporal object modeling of sequences. To convince ourselves of this point, results obtained with a semi-global characterization are considered. Here the extracted features are computed on "space-time micro-volumes" composed of several (but not all) successive frames of the sequences. The feature vectors are therefore extracted on a sliding temporal window, which comprise the binary masks of moving points detected through N frames. This allows representing a sequence by a temporal succession of semi-global feature vectors. Hidden Markov Models are then employed to learn and recognize the actions. A study is performed on the evolution of the recognition rate according to the length N of the temporal window and the number of states used for the HMMs. The number of states varies from 1 to 6 and the length of the temporal window varies from 2 to 17 frames. Fig. 9 presents the results and demonstrates that the highest window length associated to the lowest number of states lead to the best results. This shows that the characterization becomes more robust when the sequences are considered more globally. The results obtained by this approach (89.8%) are similar to the recognition rates presented in this work. Thus, a semi-global characterization may be a valuable alternative for action recognition.

Another issue that should be pointed out is that results become more accurate when more global motion is considered. The method based on the distance of pixels to the background gives rise to a non binary function that characterizes the shape to be recognized. Blank et al. (2005) utilize a similar approach to recognize actions. They assign for every internal point of the silhouette a value that reflects the mean time required for a random walk to hit the boundaries. This function (as well as the one proposed here) has level sets that represent smoother versions of the bounding contour of the silhouette. To recognize actions, Blank et al. also use global moment features of this function and obtain significantly improved results. However, their database is only comprised of binary masks that are compensated for translation of the center of mass and all silhouettes have the same scale. Therefore, they only consider the motion of the body parts relative to the torso. As we mentioned before, our method does not make use of such compensations but instead keeps the information about the change in position and scale from one frame to another. This allows th recognition of the direction of motion (i.e. left or right orientation) and whether a person is moving towards the camera or far from it.

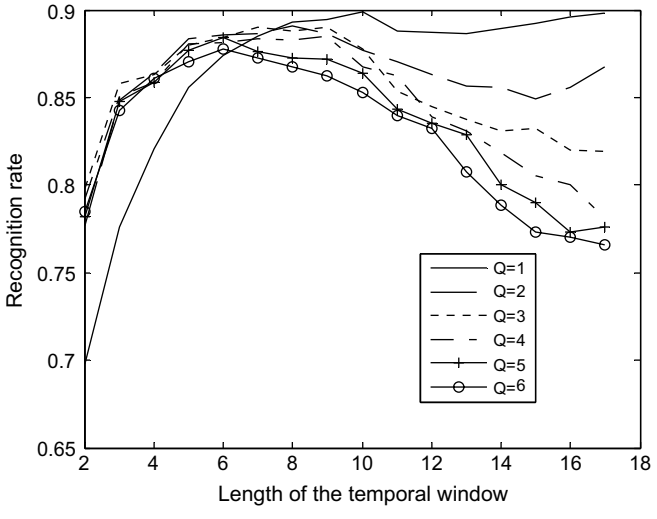


Fig. 9. Recognition rate according to the length of the temporal window and the number of states in HMMs.

Results obtained with optical flow measurements show that the speed of body parts is also an effective cue for the recognition of actions, but is not as well suited as the variation of the shape and position of the moving object when considering more global motions.

8. Summary and conclusions

In this work, a general method to recognize actions of everyday life is proposed. The approach has the ability to distinguish between different classes of actions. It does not compensate for the global motion so that the recognition of some actions is facilitated. Motion detection is initially performed on each image by modeling each pixel by a mixture of Gaussians and removing shadows. The 3D volume constructed for each sequence from the binary images resulting from detection, is characterized by its 3D geometrical moments. Those are normalized in order to obtain invariance to the position and scale of actions, to the morphology of people executing them and to the duration of actions. Invariance to viewpoint is obtained by several acquisitions of the same action at different angles. Moments are also calculated over non-binary volumes in which the value of each pixel depends on the distance to the background. Using this method, a recognition rate of 90.0% is obtained on a database of 1614 sequences, divided in eight actions and carried out by seven persons. Optical flow measurements are made and used as well for the feature estimation. They provide the less improved results, compared to the other methods.

A parallel study is performed, using semi-global features estimated on a sliding sub-window of sequences. By varying the parameters of this method, it is observed that the recognition rates increase as the features are considered more globally on the sequence. One can therefore conclude that a global or semi-global representation of sequences is

more robust than approaches which consider actions as temporal objects. An extension of the number of actors in the database is envisioned to be more robust to the silhouette of the person or his clothing and improve classification results. Furthermore, increasing the number of examples may lead to a finer modeling of each class.

References

Blank, M., Gorelick, L., Shetman, E., Irani, M., Basri, R., 2005. In: Proc. of the IEEE Int. Conf. on Computer Vision, Beijing, China.

Bobick, A.F., Davis, J.W., 2001. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 23, 257–267.

Chomat, O., Crowley, J.L., 1999. Probabilistic recognition of activity using local appearance. In: Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition, Colorado, USA.

Cupillard, F., Avanzi, A., Brémont, F., Thonnat, M., 2004. Video understanding for metro surveillance. In: Proc. of the IEEE Int. Conf. on Networking, Sensing and Control, Taipei, Taiwan.

Davis, J.W., 1998. Recognizing movement using motion histograms. Technical Report No. 487, MIT Media Laboratory Perceptual Computing Section.

Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features. In: Proc. of the IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China.

Efros, A.A., Berg, A.C., Mori, G., Malik, J., 2003. Recognizing action at a distance. In: Proc. of the IEEE Int. Conf. on Computer Vision, Nice, France.

Gavrila, D.M., 1999. The visual analysis of human movement: A survey. Comput. Vis. Image Understand. 73, 82–98.

Hongeng, S., Bremont, F., Nevatia, R., 2000. Bayesian framework for video surveillance application. In: Proc. of the Int. Conf. on Computer Vision, Barcelona, Spain.

Hu, M.K., 1662. Visual pattern recognition by moment invariants. IRE Trans. Inform. Theor. 8, 179–187.

Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybernet. 34, 334–352.

Ke, Y., Sukthankar, R., Herbert, M., 2005. Efficient visual event detection using volumetric features. In: Proc. of the IEEE Int. Conf. on Computer Vision, Beijing, China.

Li, H., Greenspan, M., 2005. Multi-scale gesture recognition from time varying contours. In: Proc. of the IEEE Int. Conf. on Computer Vision, Beijing, China.

Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: Proc. of the DARPA Image Understanding Workshop, Washington, DC, USA.

Martin, J., Crowley, J.L., 1997. An appearance based approach to gesture recognition. In: Proc. of the Int. Conf. on Image Analysis and Processing, Florence, Italy.

Mokhber, A., Achard, C., Milgram, M., Qu, X., 2005. Action recognition with global features. In: Proc. of the IEEE Workshop on Human Computer Interaction, Beijing, China.

Ogale, A.S., Karapurkar, A., Aloimonos, Y., 2005. View invariant modeling and recognition of human actions using grammars. In: Proc. of the IEEE Int. Conf. on Computer Vision, Beijing, China.

Pierobon, M., Marcon, M., Sarti, A., Tubaro, S., 2005. Clustering of human actions using invariant body shape descriptor and dynamic time warping. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance, Como, Italy.

Porikli, F., Tuzel, O., 2003. Human body tracking by adaptive background models and mean-shift analysis. In: Proc. of the IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, Nice, France.

Rabiner, L., 1990. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Readings in Speech Recognition. Morgan Kaufmann Publishers Inc., pp. 267–296.

Shetman, E., Irani, M., 2005. Space-time behavior based correlation. In: Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA. pp. 267–296.

Starner, T., Weaver, J., Pentland, A., 1998. Real time American sign language recognition from video using HMMs. IEEE Trans. Pattern Anal. Mach. Intell. 12, 1371–1375.

Stauffer, C., Grimson, W., 1999. Adaptive background mixture models for real-time tracking. In: Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition, Ft. Collins, CO, USA. pp. 246–252.

Sun X., Chen C., Manjunath, B.S., 2004. In: Proc. of the Int. Conf. on Pattern Recognition, Cambridge, United Kingdom.

Wang, J.J., Singh, S., 2003. Video analysis of human dynamics – A survey. Real-time Imaging J. 9, 320–345.

Weinland, D., Ronfard, R., Boyer, E., 2005. In: Proc. of the IEEE Int. Conf. on Computer Vision, Beijing, China.

Yamato, J., Ohya, J., Ishii, K., 1992. Recognizing human action in time-sequential images using Hidden Markov Models. In: Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition, Los Alamitos. pp. 379–385.

Zelnik-Manor, L., Irani, M., 2001. Event based analysis of video. In: Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition, Hawaii, USA. pp. 123–130.