International Scholarly Research Network
ISRN Machine Vision
Volume 2012, Article ID 872131, 9 pages
doi:10.5402/2012/872131

*Research Article*

# Chord-Length Shape Features for Human Activity Recognition

**Samy Sadek, Ayoub Al-Hamadi, Bernd Michaelis, and Usama Sayed**

*Institute for Electronics, Signal Processing and Communications (IESK), Otto-von-Guericke-University Magdeburg, 39106 Magdeburg, Germany*

Correspondence should be addressed to Samy Sadek, samy.bakheet@ovgu.de

Received 2 August 2012; Accepted 20 September 2012

Academic Editors: M. La Cascia, A. Prati, J. M. Tavares, and C. S. Won



Despite their high stability and compactness, chord-length shape features have received relatively little attention in the human action recognition literature. In this paper, we present a new approach for human activity recognition, based on chord-length shape features. The most interesting contribution of this paper is twofold. We first show how a compact, computationally efficient shape descriptor; the chord-length shape features are constructed using 1-D chord-length functions. Second, we unfold how to use fuzzy membership functions to partition action snippets into a number of temporal states. On two benchmark action datasets (KTH and WEIZMANN), the approach yields promising results that compare favorably with those previously reported in the literature, while maintaining real-time performance.

## 1. Introduction

Recognizing human activities in video data is a paramount, but challenging task in computer vision and image understanding. It was concluded that developing efficient approaches and algorithms for solving the problem of human action/behavior recognition would yield huge potential for a large number of potential applications, for example, human-computer interaction, video surveillance, gesture recognition, robot learning and control, and so forth. In fact, the non-rigid nature of human body and clothes in video sequences, resulting from drastic illumination changes, changing in pose, and erratic motion patterns, presents the grand challenge to human detection and action recognition [1].

In addition, while the real-time performance is a major concern in computer vision, especially for embedded computer vision systems, the majority of state-of-the-art action recognition systems often employ sophisticated feature extraction and learning techniques, creating a barrier to the real-time performance of these systems. This suggests a tradeoff between accuracy and real-time requirements. The automatic recognition and understanding of human actions in video sequences are still an underdeveloped area due to the lack of a general purpose model and most approaches proposed in the literature remain limited in their ability. For this, much research still needs to be undertaken to address the ongoing challenges. The remaining paper is structured as follows. Section 2 gives the related work. In Section 3, the chord-length functions and chord-length features are described. Section 4 details the proposed action recognition method. Experimental results corroborating the efficiency of the proposed method are presented in Section 5. Finally, Section 6 concludes and outlines some prospects for future work.

## 2. Related Literature

Over the course of the last couple of decades or so, a great deal of work has been done (and still being done) on the recognition of human activities from both still images and video sequences. Despite these years of work, the problem is still open and provides a big challenge to the researchers and more rigorous research is needed to come around it. Human action can generally be recognized using various visual cues such as motion [1, 3–5] and shape [6–10]. Scanning the literature, one notices that a significant body of work in action recognition focuses on using spatial-temporal key points and local feature descriptors [11–15]. The local features are extracted from the region around each key point

detected by the key point detection process. These features are then quantized to provide a discrete set of visual words before they are fed into the classification module. Another thread of research is concerned with analyzing patterns of motion to recognize human actions. For instance, in [3], periodic motions are detected and classified to recognize actions. In [5] the authors analyze the periodic structure of optical flow patterns for gait recognition. Alternatively, some researchers have opted to use both motion and shape cues. For example, in [16], Bobick and Davis use temporal templates, including motion-energy images and motion-history images to recognize human movement. In [17] the authors detect the similarity between video segments using a space-time correlation model. While in [18], Rodriguez et al. present a template-based approach using a Maximum Average Correlation Height (MACH) filter to capture intraclass variabilities. Jhuang et al. [19] perform actions recognition by building a neurobiological model using spatio-temporal gradient. In [20], actions are recognized by training different SVM classifiers on the local features of shape and optical flow. In parallel, a significant amount of work is targeted at modelling and understanding human motions by constructing elaborated temporal dynamic models [21–24]. Finally, there is also an attractive area of research that concentrates on using generative topic models for visual recognition based on the so-called Bag-of-Words (BoWs) model. The underlying concept of a BoW is that the video sequences are represented by counting the number of occurrences of descriptor prototypes, so-called visual words. Topic models are built and then applied to the BoW representation. Three of the most popularly used topic models are Latent Dirichlet Allocation (LDA) [25], Correlated Topic Models (CTMs) [26] and probabilistic Latent Semantic Analysis (pLSA) [27].

## 3. Chord-Length Functions

A shape border, that is, contour, is an inalienable property of every object and can be defined as a simply connected sequence consisting of $n$ 2 d points:

$$\mathcal{C} = \{z_i = (x_i, y_i) \in \mathbb{R}^2 \mid i = 0, 1, \ldots, n-1\}, \quad (1)$$

where $z_{i+n} = z_i$ as $\mathcal{C}$ is closed. The diameter $\ell$ of the shape boundary is given as

$$\ell = \max_{i,j=0}^{N-1} \left\| z_i - z_j \right\|, \quad (2)$$

where $\|\cdot\|$ is defined as the Euclidean distance between two points $z_i$ and $z_j$. Taking as an initial point $z_i \in \mathcal{C}$, let the contour $\mathcal{C}$ be traversed anticlockwisely and partitioned into $k > 1$ arc segments of equal length, that is,

$$\widehat{z_i p_1}, \widehat{p_1 p_2}, \ldots, \widehat{p_{k-1} z_i}, \quad (3)$$

where $p_j$ is the $j$th division point and $j = 1, 2, \ldots, k-1$. Thus, we have $k-1$ chords:

$$\overline{z_i p_1}, \overline{z_i p_2}, \ldots, \overline{z_i p_{k-1}}, \quad (4)$$

and $k-1$ lengths:

$$\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{k-1}^{(i)}, \quad (5)$$

where $\lambda_j^{(i)}$ is the length of the chord $\overline{z_i p_j}$ measured as the Euclidean distance between the two points $p_j$ and $z_i$, as shown in Figure 1. Now let us assume the point $z_i$ travels along the contour, then the chord lengths $\lambda_j^{(i)}$ will vary accordingly. This implies that $\lambda_j^{(i)}$ is a function of $z_i$. Such a function is termed a chord-length function (CLF) and shortly denoted as $\lambda_j$ [31]. Therefore we obtain $k-1$ CLFs, $\lambda_1, \lambda_2, \ldots, \lambda_{k-1}$. As those functions are obtained from splitting the contour evenly and from moving the initial point $z_i$, along the contour, so that they guarantee to be invariant to translation and rotation. However, the chord length is not scale invariant, but it can be normalized to be invariant using the contour diameter $\ell$.

The CLFs apparently meet the key requirements for being a shape descriptor. Then we need to scale all the CLFs to be within the same range (e.g., $[0, 1]$). By their definition, the CLFs are derived by segmenting the contour evenly, so that it is easy to deduce that only half of the CLFs, $\lambda_1, \lambda_2, \ldots, \lambda_{k/2}$ are enough to describe the shape adequately. It is worthwhile here to point to the fact that both global and local features of a shape can be captured by using chord lengths of different levels. The local features are likely to be captured by the CLFs of the partition points closer to the initial point $z_i$, while the global features are captured by those of farther points. This is the uncanny advantage of the CLFs versus other shape descriptors.

## 4. Suggested Methodology

The framework of the proposed action recognition system is schematically illustrated in Figure 2. In the following subsections, the steps of the scheme are described in more detail.

*4.1. Preprocessing and Background Subtraction.* For the later feature extraction or classification, preprocessing could provide more meaningful features that help in improving the final recognition results. First, all the frames of each action snippet are smoothed by using Gaussian convolution. Then backgrounds are subtracted from each action snippet using a Mixture-of-Gaussians (MoG) background modeling technique. For background substraction, a GMM background model analogous to that described in [32] is used. In this model, each pixel in the scene is modeled by a mixture of $K$ Gaussian distributions. Thus the probability that a certain pixel has intensity $x_t$ at time $t$ is given by

$$p(x_t) = \sum_{i=1}^{K} w_i * \eta(x_t; \mu_i, \Sigma_i), \quad (6)$$

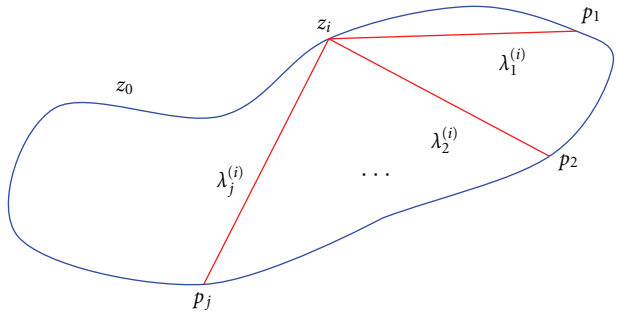


Figure 1: Chord-length function (CLF) obtained from partitioning the contour into a finite number of arcs of equal length.

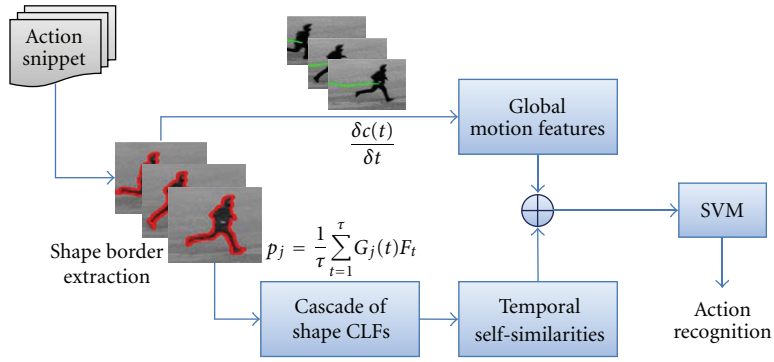


Figure 2: Block diagram of the proposed action recognizer.

where $w_i$, $\mu_i$, and $\Sigma_i$ are the weight, the mean, and the covariance of the $i$th distribution at time $t$, respectively, and $\eta$ is the Gaussian probability density function:

$$\eta(x_t;\mu,\Sigma) = \frac{1}{(2\pi)^{1/2}|\Sigma|^{1/2}} e^{(-1/2)(x_t-\mu)^T\Sigma^{-1}(x_t-\mu)}. \quad (7)$$

Therefore, unfiltered silhouettes can be produced. Finally the shape borders representing all poses of a specific action are extracted from the filtered silhouette. These preprocessing operations are summarized in Figure 3.

*4.2. Feature Extraction.* Initially, we divide a video sequence into several temporal states to compensate the time warping effects. These states are defined by vague, linguistic intervals. Gaussian membership functions are used to describe the temporal intervals,

$$\mathcal{G}_j\left(t;\varepsilon_j,\sigma,r\right) = e^{(-1/2)|(t-\varepsilon_j)/\sigma|^r}, \quad j = 1,2,\ldots,m, \quad (8)$$

where $\varepsilon_j$, $\sigma$, and $r$ are the center, width, and fuzzification factor, respectively, and $m$ is the total number of temporal states of action. Note that the membership functions defined above are chosen to be of identical shape on condition that their sum is equal to one at any instance of time, as shown in Figure 4. By using such fuzzy functions, not only can temporal information be easily extracted, the performance decline due to time warping effects can also be nullified.

*4.2.1. Chord-Length Shape Features.* As shown previously in Section 3, given a shape, $k/2$ CLFs can be defined by dividing the shape border into $k$ arcs of equal length. These functions are invariant to translation, rotation, and scaling. Though, like other shape descriptors, these descriptors are not sufficiently compact. Additionally, they depend constantly on a reference point whereby the shape border is parameterized. This dependence is simply because the contour is closed and any point on the contour can be used as a reference point, thus the CLFs might be changed. In order to avoid these problems and for convenience, the mean $\mu_j$ and variance $\sigma_j$ of the CLFs are adopted,

$$\mu_j = \frac{1}{n}\sum_{i=0}^{n-1}\lambda_j^{(i)}, \qquad \sigma_j = \frac{1}{n-1}\sum_{i=0}^{n-1}\left(\lambda_j^{(i)} - \mu_j\right)^2. \quad (9)$$

Hence, the CLF descriptor of shape can be expressed as follows:

$$F = \begin{pmatrix} \mu_1 & \sigma_1 \\ \mu_2 & \sigma_2 \\ \vdots & \vdots \\ \mu_{k/2} & \sigma_{k/2} \end{pmatrix}. \quad (10)$$

In order to obtain the CLF shape descriptor of a given action, we first obtain the CLF descriptor for all poses of this action. As each action snippet was temporally divided into a number

(a) (b) (c)

Figure 3: (a) source image, (b) silhouette, and (c) extracted shape border.

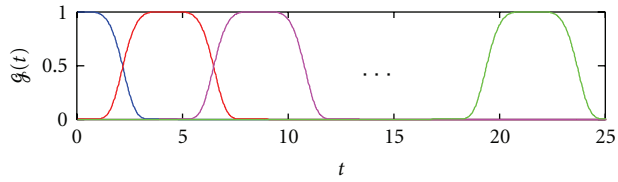


Figure 4: Gaussian membership functions used to represent the temporal intervals, with $\varepsilon_j = \{0, 4, 8, \ldots\}$, $\sigma = 2$ and $r = 5$.

of fuzzy states representing poses of the action, thus the CLF descriptor of an action pose is obtained by

$$p_j = \frac{1}{\tau}\sum_{t=1}^{\tau}\vartheta_j(t)F_t, \qquad j = 1, 2, \ldots, m, \tag{11}$$

where $F_t$ and $\tau$ are the CLFs shape descriptor at time $t$ and the length of temporal state, respectively. Accordingly the final CLFs descriptor of the action can be constructed by concatenating all the CLFs shape descriptors of its temporal poses. The resulting feature vectors (i.e., CLFs descriptors) are then normalized to the integral value of unity. The normalized feature vectors obtained can be exploited as shape descriptors for classification and matching. Generally, many approaches in computer vision directly combine such normalized vectors to obtain the resultant feature vector per video clip, which in turn can be classified by any machine learning algorithm (SVM, ANN, NB, decision trees, etc.). In contrast, in this work, we aim to enrich these vectors by the self-similarity analysis. This is paramount to improve the ability to discriminate between temporal variations of different human actions.

*4.2.2. Temporal Self-Similarities Construction.* For comparing the similarity between two vectors, one can adopt several metrics (Euclidean metric, Cosine metric, Mahalanobis metric, etc.). Whilst such metrics might have some intrinsic merit, they have some limitations to be used with our approach because we might care more about the overall shape of expression profiles rather than the actual magnitudes, which is of main concern in applications such as action recognition. Therefore, we use a different similarity metric in which the trends and relative changes are considered. Such metric is based on Pearson Linear Correlation (PLC),

$$s(u, v) = \frac{1 - \rho(u, v)}{2}, \tag{12}$$

where $\rho(u, v)$ is the PLC between the two vectors $u$ and $v$ that is, defined as

$$\rho(u, v) = \frac{\sum_{i=1}^{m}(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\sum_{i=1}^{m}(u_i - \overline{u})^2 \sum_{i=1}^{m}(v_i - \overline{v})^2}}. \tag{13}$$

The means $\overline{u}$ and $\overline{v}$ of $u$ and $v$ are given by

$$\overline{u} = \frac{1}{m}\sum_{i=1}^{m}u_i, \qquad \overline{v} = \frac{1}{m}\sum_{i=1}^{m}v_i. \tag{14}$$

Given a set of feature vectors $P = \{p_1, p_2, \ldots, p_m\}$ that represent $m$ poses (or temporal states) of an action, the temporal self-similarity matrix of the action is given as

$$S = \left[s_{ij}\right]_{i,j=1}^{m} = \begin{pmatrix} 0 & s_{12} & \cdots & s_{1m} \\ s_{21} & 0 & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & 0 \end{pmatrix}, \tag{15}$$

where $s_{ij} = s(p_i, p_j), i, j = 1, 2, \ldots, m$. The main diagonal elements are zero because $s(p_i, p_i) = 0$. Meanwhile, because $s_{ij} = s_{ji}$, $S$ is a symmetric matrix. It is important to point out that the self-similarities matrix achieves the goal of reducing the dimensionality of the feature space from $m \times k$, to $m(m-1)/2$, without losing the relevant temporal information. For the present work, various values of $m$ were tried but $m = 5$ was found to give the best results.

*4.3. Fusing Motion Features with Shape Features.* Global features of motion have proven to be advantageous in many applications of object recognition. This encourage us to extend the idea and fuse motion features and CLF features to form the final SVM model. The motion features extracted here are based on calculating the center of of gravity Figure 5 (i.e. shape centroid) that delivers the center of motion and is given by

$$\vec{v}(t) = \frac{\delta\vec{c}(t)}{\delta t}, \tag{16}$$

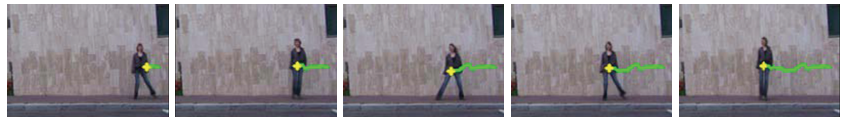

Figure 5: Center of gravity (CoF) delivering the center of motion for a "siding" action form WEIZMANN dataset [2].

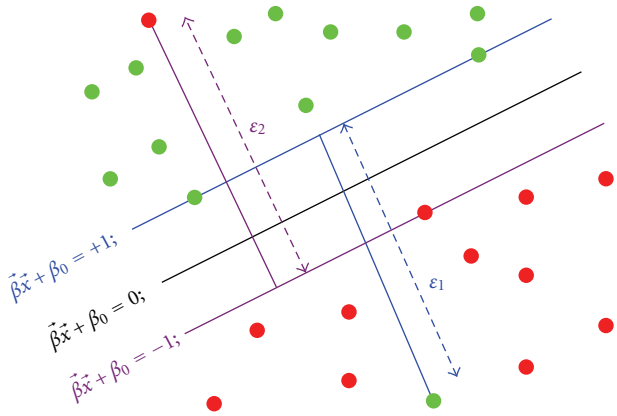


Figure 6: Generalized optimal separating hyperplane.

where the spatial coordinates of $\vec{c}(t)$ are given by

$$c_x = \frac{1}{6\lambda}\sum_{i=1}^{n}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i),$$
$$c_y = \frac{1}{6\lambda}\sum_{i=1}^{n}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i), \tag{17}$$

where $\lambda = (1/2)|\sum_{i=1}^{n}(x_i y_{i+1} - x_{i+1} y_i)|$. Such features have profound implications, not only about the type of motion (e.g., translational or oscillatory), but also about the rate of motion (i.e., velocity). With these features, it would be able to distinguish, for example, between an action where motion occurs over a relatively large area (e.g., running) and an action localized in a smaller region, where only small parts of the body are in motion (e.g., boxing). It is worthwhile mentioning that fusing motion information with local features was very beneficial for the current action recognition task, and thereby a dramatic improvement in recognition accuracy was achieved.

*4.4. Action Classification Using SVM.* In this section, we formulate the action recognition task as a multiclass learning problem, where there is one class for each action, and the goal is to assign an action to an individual in each video sequence. There are various supervised learning algorithms by which an action recognizer can be trained.

Support Vector Machines (SVMs) are used in our framework due to their outstanding generalization capability and reputation of a highly accurate paradigm. SVMs [38] are based on the Structure Risk Minimization principle from computational theory and are a solution to data overfitting in neural networks. Originally, SVMs were designed to handle dichotomic classes in a higher dimensional space where a maximal separating hyperplane is created. On each side of this hyperplane, two parallel hyperplanes are conducted. Then SVM attempts to find the separating hyperplane that maximizes the distance between the two parallel hyperplanes. Intuitively, a good separation is achieved by the hyperplane having the largest distance (see Figure 6). Hence, the larger the margin is, the lower the generalization error of the classifier will be. More formally, let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in R^d, y_i \in \{-1, +1\}\}$ be a training dataset; Coretes and Vapnik stated in their paper [38] that this problem is best addressed by allowing some examples to violate the margin constraints. These potential violations are formulated using some positive slack variables $\xi_i$ and a penalty parameter $C \geq 0$ that penalize the margin violations. Thus the optimal separating hyperplane is determined by solving the following primal quadratic programming (QP) problem:

$$\min_{\boldsymbol{\beta},\beta_0} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_i \xi_i$$
$$\text{subject to} \quad (y_i(\langle \mathbf{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \geq 1 - \xi_i \forall i) \wedge (\xi_i \geq 0 \forall i). \tag{18}$$

Geometrically, $\boldsymbol{\beta} \in R^d$ is a vector going through the center and perpendicular to the separating hyperplane. The offset parameter $\beta_0$ is added to allow the margin to increase and not to force the hyperplane to pass through the origin that restricts the solution. For computational purposes it is more convenient to solve SVM in its dual formulation. This can be accomplished by forming the Lagrangian and then optimizing over the Lagrange multiplier $\boldsymbol{\alpha}$. The resulting decision function has weight vector $\boldsymbol{\beta} = \sum_i \alpha_i \mathbf{x}_i y_i, 0 \leq \alpha_i \leq C$. The instances $\mathbf{x}_i$ with $\alpha_i > 0$ are called *support vectors*, as they uniquely define the maximum margin hyperplane.

In this approach, several classes of actions are created. Several one-versus-all SVM classifiers are trained using the features extracted from action snippets in the training dataset. The updiagonal elements of the temporal similarity matrix representing the shape features are first transformed into plain vectors based on the element scan order. The motion feature are then concatenated with the shape features to generate the final hybrid feature vectors. The dimension of final feature vector is $(m(m-1)/2)+m = m(m+1)/2$. Finally, the final feature vectors are fed into the SVM classifiers for the final decision.

## 5. Experiments and Results

In this section the experiments we conducted to assess the performance of the proposed approach are described and some of their results are presented. And also in order to

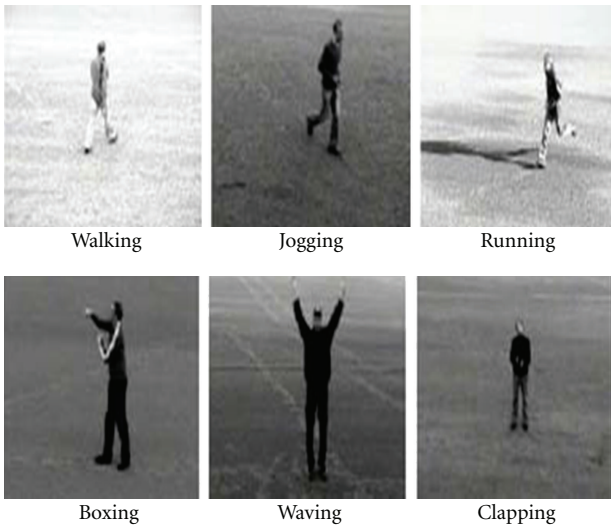


Figure 7: Sample frames for actions in the KTH action dataset used in the evaluation process.

demonstrate the effectiveness of the proposed method, the obtained results are compared with those reported in the current literature. Two main experiments were carried out to evaluate this approach. The first one was carried out on the the publicly benchmark KTH action dataset [39], while the second one was conducted on the popular Weizmann action dataset [2].

*5.1. Experiment 1.* The KTH action dataset contains six types of human actions (i.e., walking, jogging, running, boxing, hand waving, and hand clapping), performed repeatedly by 25 individuals under four different scenarios including outdoors (*s1*), outdoors with scale variation (*s2*), outdoors with different clothes (*s3*), and indoors (*s4*). Typical example frames of six action categories in the KTH dataset can be seen in Figure 7. In order to prepare the experiments and to provide an unbiased estimation of the generalization abilities of the classification process, the sequences for each action were partitioned into two independent subsets, that is, a training set and a test set. More specifically, a set of sequences (72% of all sequences) performed by 18 subjects were used for training and other sequences (the remaining 28%) performed by other 7 subjects were set aside as a test set. SVMs with Gaussian radial basis function (RBF) kernel are trained on the training set, while the evaluation of the recognition performance is performed on the test set. The confusion matrix that shows the recognition results achieved on the KTH action dataset is given in Table 1, while the comparison of the obtained results with those obtained by other methods available in the literature is shown in Table 2.

As follows from the figures tabulated in Table 1, most actions are correctly classified. Furthermore there is a high distinction between arm actions and leg actions. Most of the mistakes where confusions occur are between "jogging" and "running" actions and between "boxing" and "clapping" actions. This is intuitively plausible due to the fact of high similarity between each pair of these actions. From the comparison given by Table 2, it turns out that our method performs competitively with other state-of-the-art methods and its results compare favorably with previously published results.

Table 1: Confusion matrix for the KTH action dataset.

| Action | walking | running | jogging | boxing | waving | clapping |
|---|---|---|---|---|---|---|
| walking | 0.94 | 0.01 | 0.05 | 0.00 | 0.00 | 0.00 |
| running | 0.00 | 0.96 | 0.04 | 0.00 | 0.00 | 0.00 |
| jogging | 0.04 | 0.08 | 0.88 | 0.00 | 0.00 | 0.00 |
| boxing | 0.00 | 0.00 | 0.00 | 0.94 | 0.02 | 0.04 |
| waving | 0.00 | 0.00 | 0.00 | 0.02 | 0.93 | 0.05 |
| clapping | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.96 |

Table 2: Comparison with the state-of-the-art on the KTH action dataset.

| Method | Accuracy |
|---|---|
| Our method | 93.5% |
| Liu and Shah [13] | 92.8% |
| Wang and Mori [28] | 92.5% |
| Jhuang et al. [19] | 91.7% |
| Rodriguez et al. [18] | 88.6% |
| Rapantzikos et al. [29] | 88.3% |
| Dollár et al. [11] | 81.2% |
| Ke et al. [30] | 63.0% |

*5.2. Experiment 2.* The Weizmann action dataset was first provided by Blank et al. [2] in 2005, which contains a total of 90 video clips (i.e., 5098 frames) performed by 9 individuals. Each video clip contains one person performing an action. There are 10 categories of actions involved in the dataset, namely, *walking*, *running*, *jumping*, *jumping in place*, *bending*, *jacking*, *skipping*, *galloping sideways*, *one hand waving,* and *two-hand-waving*. Typically, all the clips in the dataset are sampled at 25Hz and last about 2 seconds with image frame size of $180 \times 144$. A sample frame for each action in the Weizmann dataset is illustrated in Figure 8. In order to provide an unbiased estimate of the generalization abilities of the proposed method, we have used the leave-one-out cross-validation (LOOCV) technique in the validation process. As the name suggests, this involves using a group of sequences from a single subject in the original dataset as the testing data and the remaining sequences as the training data. This is repeated such that each group of sequences in the dataset is used once as the validation. More specifically, the sequences of 8 subjects were used for training and the sequences of the remaining subject were used for validation data. Again, as with the first experiment, SVMs with Gaussian RBF kernel are trained on the training set, while the evaluation of the recognition performance is performed on the test set.

The recognition results obtained by the proposed method are summarized in a confusion matrix in Table 3, where correct responses define the main diagonal. From the figures in the matrix, a number of points can be drawn.

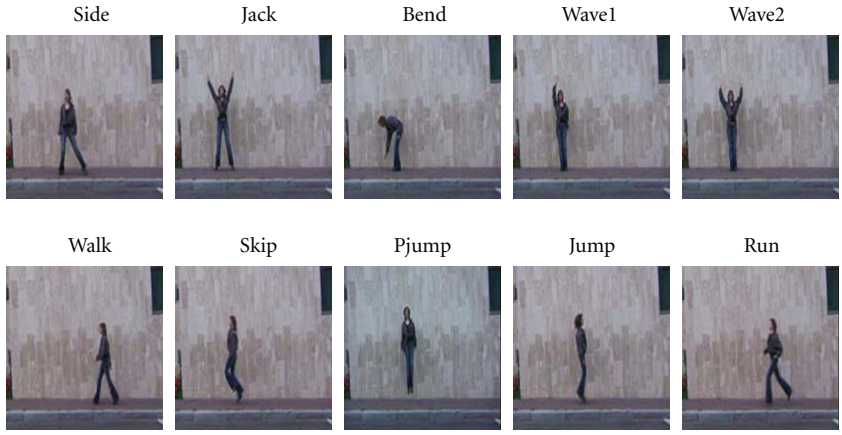


Figure 8: A sample frame for each action in the WEIZMANN action dataset [2].

Table 3: Confusion matrix for the WEIZMANN dataset.

| Action | wave2 | wave1 | walk | skip | side | run | pjump | jump | jack | bend |
|---|---|---|---|---|---|---|---|---|---|---|
| wave2 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| wave1 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| walk | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| skip | 0.00 | 0.00 | 0.00 | 0.89 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 |
| side | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| run | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| pjump | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| jump | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.11 | 0.00 | 0.89 | 0.00 | 0.00 |
| jack | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| bend | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

Table 4: Comparison with other state-of-the-art on the WEIZMANN action dataset.

| Method | Accuracy |
|---|---|
| Our method | 97.8% |
| Fathi and Mori [33] | 100% |
| Bregonzio et al. [34] | 96.6% |
| Zhang et al. [35] | 92.8% |
| Niebles et al. [36] | 90.0% |
| Dollár et al. [11] | 85.2% |
| Klaser et al. [37] | 84.3% |

The majority of actions are correctly classified. An average recognition rate of 97.8% is achieved with our proposed method. What is more, there is a clear distinction between arm actions and leg actions. The mistakes where confusions occur are only between *skip* and *jump* actions and between *jump* and *run* actions. This intuitively seems to be reasonable due to the fact of high closeness or similarity among the actions in each pair of these actions. In order to quantify the effectiveness of the proposed method, the results obtained are compared qualitatively with those obtained previously by other investigators. The outcome of this comparison is presented in Table 4. In light of this comparison, we can see that the proposed method is competitive with the state-of-the-art methods. It is important to mention that all the methods [11, 34–37] that we have compared our method with, except the method proposed in [33], have used similar experimental setups, so that the comparison seems to be meaningful and most fair. A final remark that we want to make here is that this approach is able to work at about 28 fps (using a 2.8 GHz Intel dual core machine with 4 GB of RAM). Therefore, it can offer timing guarantees to real-time applications and embedded systems.

## 6. Conclusion and Future Work

In this paper, we have introduced an approach for human activity recognition based on CLF shape features. On two benchmark action datasets, the results achieved by the approach have demonstrated that it leads to significant improvements in recognizing accuracy and efficiency and maintains competitiveness with existing state-of-the-art approaches. However, it would also be advantageous to explore the empirical validation of the approach on more realistic datasets presenting many technical challenges in data handling, such as object articulation, occlusion, and significant background clutter. These issues are crucial and thus will be more thoroughly investigated within the scope of future work.

## Acknowledgments

This work is supported by Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by DFG and BMBF Bernstein-Group (FKZ: 01GQ0702). The authors would also like to thank the anonymous reviewers for their constructive comments and insightful suggestions made on an earlier version of the paper that greatly contributed to improving the quality of this work.

## References

[1] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "An efficient method for real-time activity recognition," in *Proceedings of the International Conference of Soft Computing and Pattern Recognition (SoCPaR '10)*, pp. 69–74, Paris, France, December 2010.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1395–1402, October 2005.

[3] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, 2000.

[4] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, vol. 2, pp. 726–733, October 2003.

[5] L. Little and J. E. Boyd, "Recognizing people by their gait: The shape of motion," *Journal of Computer Vision*, vol. 1, no. 2, pp. 1–32, 1998.

[6] W. L. Lu, K. Okuma, and J. J. Little, "Tracking and recognizing actions of multiple hockey players using the boosted particle filter," *Image and Vision Computing*, vol. 27, no. 1-2, pp. 189–205, 2009.

[7] S. Sadek, A. AI-Hamadi, M. Elmezain, B. Michaelis, and U. Sayed, "Human activity recognition via temporal moment invariants," in *Proceedings of the 10th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT '10)*, pp. 79–84, Luxor, Egypt, December 2010.

[8] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "Human activity recognition: a scheme using multiple cues," in *Proceedings of the International Symposium on Visual Computing (ISVC '10)*, vol. 1, pp. 574–583, Las Vegas, Nev, USA, November 2010.

[9] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "A statistical framework for real-time traffic accident recognition," *Journal of Signal and Information Processing*, vol. 1, pp. 70–81, 2010.

[10] C. Thurau and V. Hlavac, "Pose primitive based human action recognition in videos or still images," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, 2008.

[11] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS '05)*, pp. 65–72, October 2005.

[12] I. Laptev and P. Pérez, "Retrieving actions in movies," in *Proceedings of the 11th International Conference on Computer Vision (ICCV '07)*, 2007.

[13] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

[14] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "Towards robust human action retrieval in video," in *Proceedings of the British Machine Vision Conference (BMVC '10)*, Aberystwyth, UK, September 2010.

[15] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity," *Eurasip Journal on Advances in Signal Processing*, vol. 2011, Article ID 540375, 2011.

[16] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[17] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 405–412, June 2005.

[18] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

[19] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, October 2007.

[20] K. Schindler and L. Van Gool, "Action Snippets: how many frames does human action recognition require?" in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

[21] X. Feng and P. Perona, "Human action recognition by sequence of movelet codewords," in *Proceedings of the 1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT '02)*, pp. 717–721, 2002.

[22] N. Ikizler and D. Forsyth, "Searching video for complex activities with finite state models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, June 2007.

[23] B. Laxton, J. Lim, and D. Kriegmant, "Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, 2007.

[24] N. Oliver, A. Garg, and E. Horvitz, "Layered representations for learning and inferring office activity from multiple sensory channels," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 163–180, 2004.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[26] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems*, vol. 18, pp. 147–154, 2006.

[27] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pp. 50–57, 1999.

[28] Y. Wang and G. Mori, "Max-Margin hidden conditional random fields for human action recognition," in *Proceedings*

*of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 872–879, June 2009.

[29] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1454–1461, June 2009.

[30] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 166–173, October 2005.

[31] D. Zhang and G. Lu, "A comparative study of fourier descriptors for shape representation and retrieval," in *Proceedings of the 5th Asian Confenence on Computer Vision (ACCV '02)*, 2002.

[32] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, vol. 2, pp. 246–252, June 1999.

[33] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

[34] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1948–1955, June 2009.

[35] Z. Zhang, Y. Hu, S. Chan, and L. T. Chia, "Motion context: a new representation for human action recognition," in *Proceedings of the 10th European Conference on Computer Vision (ECCV '08)*, vol. 4, pp. 817–829, 2008.

[36] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[37] A. Klaser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proceedings of the British Machine Vision Conference (BMVC '08)*, 2008.

[38] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[39] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, vol. 3, pp. 32–36, August 2004.