Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 540375, 9 pages
doi:10.1155/2011/540375

Research Article

An Action Recognition Scheme Using Fuzzy Log-Polar Histogram and Temporal Self-Similarity

Samy Sadek,[1] Ayoub Al-Hamadi,[1] Bernd Michaelis,[1] and Usama Sayed[2]

[1] *Institute for Electronics, Signal Processing and Communications (IESK), Otto-von-Guericke University Magdeburg, 39106 Magdeburg, Germany*
[2] *Electrical Engineering Department, Assiut University, Assiut, Egypt*

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

Received 25 July 2010; Revised 26 October 2010; Accepted 8 January 2011

Academic Editor: Mark Liao

Temporal shape variations intuitively appear to provide a good cue for human activity modeling. In this paper, we lay out a novel framework for human action recognition based on fuzzy log-polar histograms and temporal self-similarities. At first, a set of reliable keypoints are extracted from a video clip (i.e., action snippet). The local descriptors characterizing the temporal shape variations of action are then obtained by using the temporal self-similarities defined on the fuzzy log-polar histograms. Finally, the SVM classifier is trained on these features to realize the action recognition model. The proposed method is validated on two popular and publicly available action datasets. The results obtained are quite encouraging and show that an accuracy comparable or superior to that of the state-of-the-art is achievable. Furthermore, the method runs in real time and thus can offer timing guarantees to real-time applications.

1. Introduction

Human action recognition has received and still receives considerable attention in the field of computer vision due to its vital importance to many video content analysis applications [1]. In spite of the voluminous existing literature on the analysis and interpretation of human motion motivated by the rise of security concerns and increased ubiquity and affordability of digital media production equipment, research on human action and event recognition is still at the embryonic stage of development. Therefore much additional work remains to be done to address the ongoing challenges. It is clear that developing good algorithms for solving the problem of action 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 nonrigid 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 [2]. 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/or learning techniques, creating a barrier to the real-time performance of these systems. This suggests that there is an inherent trade-off between recognition accuracy and computational overhead.

The rest of the paper is structured as follows. Section 2 briefly reviews the prior literature. In Section 3, the Harris scale-adaptive keypoint detector is presented. The proposed method is described in Section 4 and is experimentally validated and compared against other competing techniques in Section 5. Finally, in Section 6, the paper ends with some conclusions and ideas about future work.

2. Related Literature

For the past decade or so, many papers have been published in the literature, proposing a variety of methods for human action recognition from video. Human action can generally be recognized using various visual cues such as motion [3–6] and shape [7–11]. Scanning the literature, one notices that a large body of work in action recognition focuses on using keypoints and local feature descriptors [12–16]. The local features are extracted from the region around each keypoint. 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 [17], periodic motions are detected and classified to recognize actions. In [4] the authors analyze the periodic structure of optical flow patterns for gait recognition. Further in [18], Sadek et al. present an efficient methodology for real-time human activity based on simple statistical features. Alternatively, some other researchers have opted to use both motion and shape cues. For example in [19], Bobick and Davis use temporal templates, including motion-energy images and motion-history images to recognize human movement. In [20] the authors detect the similarity between video segments using a space-time correlation model. While in [21], Rodriguez et al. present a template-based approach using a Maximum Average Correlation Height (MACH) filter to capture intraclass variabilities, Jhuang et al. [22] perform actions recognition by building a neurobiological model using spatiotemporal gradient. In [23], 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 modeling and understanding human motions by constructing elaborated temporal dynamic models [24–27]. Finally, there is also a fertile and broadly influential area of research that uses generative topic models for modeling and recognizing action categories based on the so-called Bag-of-Words (BoW) 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 [28].

3. Scale-Adaptive Keypoint Detection

Harris keypoint detector [29] still retains its superior performance to that of many competitors [30]. However Harris detector is originally not scaleinvariant. The reliable Harris detector can be adapted to be invariant to scale changes by joining the original Harris detector with automatic scale selection. In this case, the second moment matrix quantifying the scale-adaptive detector is given by

$$\mu(\cdot;\sigma_i,\sigma_d) = \sigma_d^2 g(\cdot;\sigma_i) * \begin{pmatrix} L_x^2(\cdot;\sigma_d) & L_xL_y(\cdot;\sigma_d) \\ L_yL_x(\cdot;\sigma_d) & L_y^2(\cdot;\sigma_d) \end{pmatrix}, \quad (1)$$

where σ_i and σ_d are the integration and differentiation scale, respectively, and L_x and L_y are the derivatives of the *scale-space* representation $L(\cdot;\sigma_d)$ of the image with respect to x and y directions, respectively. The local derivatives are computed using Gaussian kernels of size σ_d. The $L(x,y;\sigma_d)$ is constructed by convolving the image with a Gaussian kernel of size σ_d. In [31], several differential operators were compared, and the experiments showed that the Laplacian of Gaussians (LoG) finds the highest percentage of correct characteristic scales

$$|\mathrm{LoG}(\cdot;\sigma_d)| = \sigma_d^2 \left| L_{xx}(\cdot;\sigma_d) + L_{yy}(\cdot;\sigma_d) \right|. \quad (2)$$

The eigenvalues of the matrix $\mu(\cdot;\sigma_i,\sigma_d)$ characterize the *cornerness* ς of a point in a given image. The sufficiently large values of the eigenvalues indicate the presence of a corner at a point. The larger the values, the stronger the corner. As an alternative way, the cornerness of a point is examined by

$$\varsigma = \det(\mu(\cdot;\sigma_i,\sigma_d)) - \alpha\,\mathrm{trace}^2(\mu(\cdot;\sigma_i,\sigma_d)), \quad (3)$$

where α is a tunable parameter. Note that computing the cornerness by (3) is computationally less expensive and numerically stable than that of the eigenvalues. The parameter α and the ratio σ_d/σ_i were experimentally set to 0.05 and 0.7, respectively. Corners are generally located at positive local maxima in a 3×3 neighborhood. It may be reasonable to get rid of unstable and weak maxima points, therefore only the maxima points of values greater than predetermined threshold are eligible to be nominated for being corners. The nominated points are then checked for whether their LoG response achieves local maxima over scales. Only the points satisfying the criteria of local maxima are keypoints.

4. Suggested Recognition Method

In this section, our method developed for recognizing human actions in video sequences, which applies fuzzy logic in action modeling, is introduced. A schematic block diagram of such an action recognizer is depicted in Figure 1. As seen from the block diagram, for each action snippet, the keypoints are first detected by the scale-adapted detector described in Section 3. To make the method more robust against time warping effects, action snippets are temporally split into a number of overlapping states defined by Gaussian membership functions. Local features are then extracted based on fuzzy log-polar histograms and temporal self-similarities. Since the global features tend to be conceivably relevant and advantageous to the current task, the final features, so-called hybrid features, fed into classifiers are constructed using both local and global features. Along next subsections further details are provided concerning the implementation aspects.

4.1. Preprocessing and Keypoint Detection. For later successful feature extraction and classification, it is important to preprocess all video sequences to remove noisy, erroneous, and incomplete data and to prepare the representative features that are suitable for knowledge generation. To wipe off noise and weaken image distortion, all frames of each action snippet are first smoothed by Gaussian convolution with a kernel of size 3×3 and variance $\sigma = 0.5$. Then the scale-invariant keypoints are detected using the scale-adapted detector previously described in Section 3. The

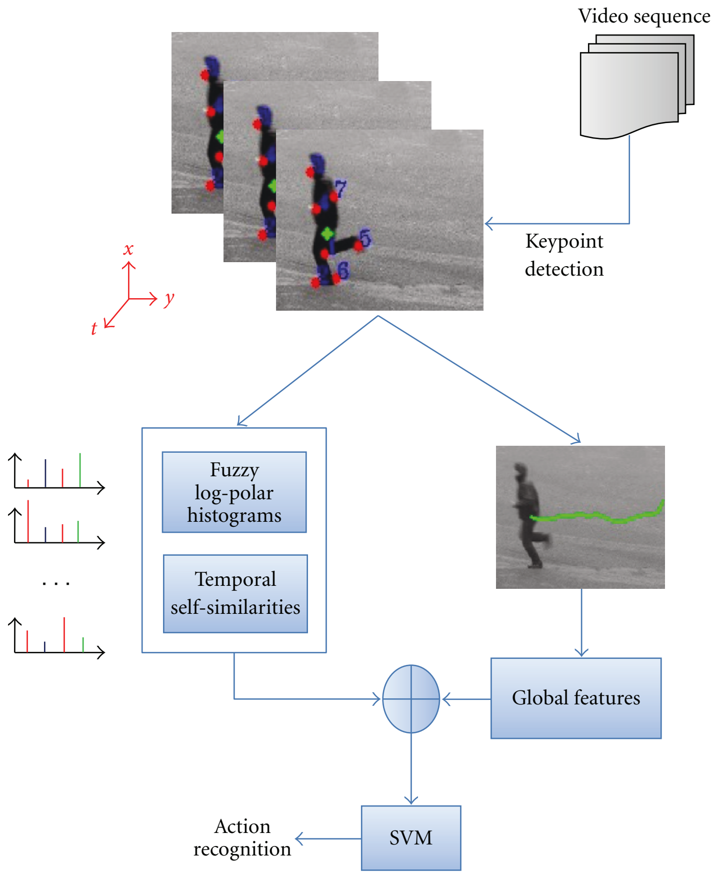

Figure 1: Block diagram of our fuzzy action recognizer.

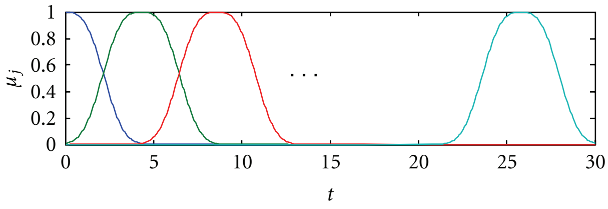

Figure 2: Gaussian membership functions used to represent the temporal intervals, with $\varepsilon_j = \{0, 4, 8, \ldots\}$, $\sigma = 2$, and $m = 3$.

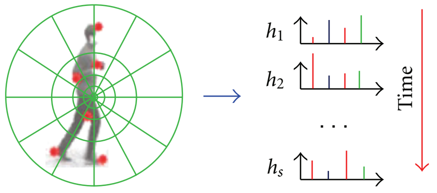

Figure 3: Fuzzy log-polar histograms representing the spatio-temporal shape contextual information of action snippet.

obtained keypoints are filtered so that under a certain amount of additive noise, only stable and more localized keypoints are retained. This is carried out in two steps. First, low contrast keypoints are discarded, and second isolated keypoints not satisfying the spatial constraints of feature point are excluded.

4.2. Local Feature Extraction. Feature extraction forms the cornerstone of any action recognition procedure, but is also the most challenging and time-consuming part. The next subsections describe in more detail how such features are defined and extracted.

4.2.1. Fuzzy Log-Polar Histograms. First, we temporally partition an action snippet into several segments. These segments are defined by linguistic intervals. Gaussian functions are used to describe these intervals, which are given by

$$\mu_j\left(t;\varepsilon_j,\sigma,m\right) = e^{-(1/2)|(t-\varepsilon_j)/\sigma|^m}, \quad j = 1, 2, \ldots, s, \tag{4}$$

where ε_j, σ, and m are the center, width, and fuzzification factor, respectively, while s is the total number of temporal segments. 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 2. It is thus seen that by using such fuzzy functions, not only can local temporal features be extracted precisely, the performance decline resulting from time warping effects can also be reduced or eliminated. To extract now the local features of the shape representing action at an instance of time, our own temporal localized shape context is defined, inspired by the basic idea of shape context. Compared with the shape context [32], our localized shape context differs in meaningful ways. The idea behind a modified shape context is based on computing rich descriptors for fewer keypoints. The shape descriptors presented here calculate the log-polar histograms on condition that they are invariant to simple transforms like scaling, rotation, and translation. The histograms are normalized for all affine transforms as well. Furthermore the shape context is reasonably extended by combining local descriptors with fuzzy memberships functions and temporal self-similarities paradigms. Human action is generally composed of a sequence of poses over time. Reasonable estimate of a pose can be constructed using a small set of keypoints. Ideally, such points are distinctive, persist across minor variation of shapes, robust to occlusion, and do not require segmentation. Let B be the set of sampled keypoints $\{(x_i, y_i)\}_{i=1}^{n}$ representing an action at an instance of time t_i, then for each keypoint p_i, the log-polar coordinates ρ_i and η_i are given by

$$\begin{aligned} \rho_i &= \log\left(\sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}\right), \\ \eta_i &= \arctan\left(\frac{y_i - y_c}{x_i - x_c}\right), \quad i = 1, 2, \dots, n, \end{aligned} \tag{5}$$

where (x_c, y_c) is the center of mass of B, which is invariant to image translation, scaling, and rotation. For this the angle η_i is computed with respect to a horizontal line passing through the center of mass. Now, to calculate the modified version of shape context, a log-polar histogram is overlaid on the shape as shown in Figure 3. Thus the histogram representing the shape context of action is constructed for each temporal phase j by

$$h_j(k_1, k_2) = \sum_{\substack{\rho_i \in \operatorname{bin}(k_1), \\ \eta_i \in \operatorname{bin}(k_2)}} \mu_j(t_i), \quad j = 1, 2, \dots, s. \tag{6}$$

By applying a simple linear transformation on the indices k_1 and k_2, the 2D histograms are converted into 1D as follows:

$$h_j(k) = h_j(k_1 d\eta + k_2), \quad k = 0, 1, \dots, d\rho d\eta - 1. \tag{7}$$

The resulting 1D histograms are then normalized to achieve robustness to scale variations. The normalized histograms obtained can be used as shape contextual information for classification and matching. Many approaches in various computer vision applications directly combine these histograms to get one histogram per video and classify it using any classification algorithm. In contrast, in this paper, we aim to enrich these histograms with self-similarity analysis after using a suitable distance function to measure similarity (more precisely dissimilarity) between each pair of these histograms. This is of most importance to accurately discriminate between temporal variations of different actions.

4.2.2. Temporal Self-Similarities of Action Snippet. Video analysis is seldom carried out directly on row video data. Instead feature vectors extracted from small portions of video (i.e., frames) are used. Thus the similarity between two video segments is measured by the similarity between their corresponding feature vectors. For comparing the similarity between two vectors, one can use several metrics such as Euclidean metric, Cosine metric, and Mahalanobis metric, and so forth. Whilst such metrics may have some intrinsic merits, they have some limitations to be used with our approach because we might care more about identifying the spatial locations of significant changes over time rather than the actual magnitudes, which is of main concern in applications such as action recognition. Therefore, we propose a new similarity (or more precisely, dissimilarity) metric in which the spatial changes are considered. Such metric is defined as

$$\rho(\vec{\mu}, \vec{v}) = \arg\max_{k}\left(\frac{(u_k - v_k)^2}{u_k + v_k}\right) \tag{8}$$

which can be easily normalized to unity, if desired. To reveal the inner structure of human action in video clip, second statistical moments (i.e., mean and variance) might seem to be not quite appropriate. Instead self-similarity analysis is of immense relevance to this task, which adapts this approach. Formally speaking, given a sequence of fuzzy histograms $H = (h_1, h_2, \dots, h_m)$ that represent m time slices of an action snippet, then the temporal self-similarity matrix is defined by

$$S = \left[s_{ij}\right]_{i,j=1}^{m} = \begin{pmatrix} 0 & s_{12} & \cdots & s_{1m} \\ s_{21} & 0 & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & 0 \end{pmatrix}, \tag{9}$$

where $s_{ij} = \rho(h_i, h_j)$, $i, j = 1, 2, \dots, m$. The main diagonal elements are zero because $s(h_i, h_i) = 0 \ \forall i$. Meanwhile, because $s_{ij} = s_{ji}$, S is a symmetric matrix.

4.3. Fusing Global Features and Local Features. It emerges from the discussion in the previous subsections that the features extracted using fuzzy log-polar histograms and temporal self-similarities have been highlighted. Such features obtained at each temporal stage are considered as temporally local features, while the features that are extracted along the entire motion are regarded as temporally global features. Though we should note that each of the two types of features is spatially local. Global features have previously proven to be successful in many applications of object recognition. This encourages us to extend the idea to the temporally global features and to fuse global features and local features to form the final SVM classifier. All global features extracted herein are based on calculating the center of gravity $\vec{m}(t)$ that

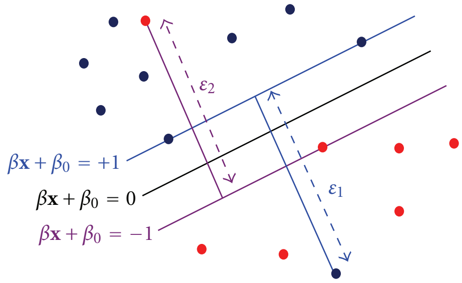
ε_2
$\beta\mathbf{x} + \beta_0 = +1$
$\beta\mathbf{x} + \beta_0 = 0$
ε_1
$\beta\mathbf{x} + \beta_0 = -1$

Figure 4: Generalized optimal separating hyperplane.

delivers the center of motion. Thus the global features $\vec{F}(t)$ describing the distribution of motion are given by

$$\vec{F}(t) = \frac{\Delta\vec{m}(t)}{\Delta t}, \quad \vec{m}(t) = \frac{1}{n}\sum_{i=1}^{n} p_i(t). \tag{10}$$

Such features are very informative 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 in which motion occurs over a relatively large area (e.g., running) and an action localized in a smaller region, where only small parts are in motion (e.g., boxing). Hence significant improvements in recognition performance are expected to be achieved by fusing global and local features.

4.4. SVM Classification. 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 [33] 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 (see Figure 4). Intuitively, a good separation is achieved by the hyperplane having the largest distance. Hence the larger the margin the lower the generalization error of the classifier. More formally, leting $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}\}$ be a training dataset, Vapnik [33] show 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 ξ_i and a penalty parameter $C \geq 0$ that penalize the margin violations.

Table 1: Confusion matrix obtained on KTH dataset.

Action	Walking	Running	Jogging	Waving	Clapping	Boxing
walking	0.98	0.00	0.02	0.00	0.00	0.00
running	0.00	0.97	0.03	0.00	0.00	0.00
jogging	0.05	0.11	0.83	0.00	0.01	0.00
waving	0.00	0.00	0.00	0.94	0.00	0.06
clapping	0.00	0.00	0.00	0.00	0.92	0.08
boxing	0.00	0.00	0.00	0.00	0.01	0.99

Table 2: Comparison with other methods done using KTH dataset.

Method	Accuracy
Our method	**93.6%**
Liu and shah [15]	**92.8%**
Wang and Mori [35]	92.5%
Jhuang et al. [22]	91.7%
Rodriguez et al. [21]	88.6%
Rapantzikos et al. [36]	88.3%
Dollár et al. [37]	81.2%
Ke et al. [12]	63.0%

Thus the optimal separating hyperplane is determined by solving the following QP problem:

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_i \xi_i \tag{11}$$

subject to $(y_i(\langle \mathbf{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \geq 1 - \xi_i \ \forall i) \wedge (\xi_i \geq 0 \ \forall i)$. Geometrically, $\boldsymbol{\beta} \in \mathbb{R}^d$ is a vector going through the origin point and perpendicular to the separating hyperplane. The offset parameter β_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 termed *support vectors*, as they uniquely define the maximum margin hyperplane. In our approach, several classes of actions are created. Several one-versus-all SVM classifiers are trained using the features extracted from the action snippets in the training dataset. The up diagonal elements of the temporal similarity matrix representing the features are first transformed into plain vectors based on the element scan order. All feature vectors are then fed into the SVM classifiers for the final decision.

5. Experiments

We present our experimental results in this section. The experiments presented here are divided into two parts. For each part, we summarize the experimental setup and the dataset we used. In this work, two popular and publicly available action datasets, namely, KTH dataset [16] and Weizmann [34], were used to demonstrate and validate our proposed approach. To assess the feasibility/reliability of the approach, the results obtained from both experiments were

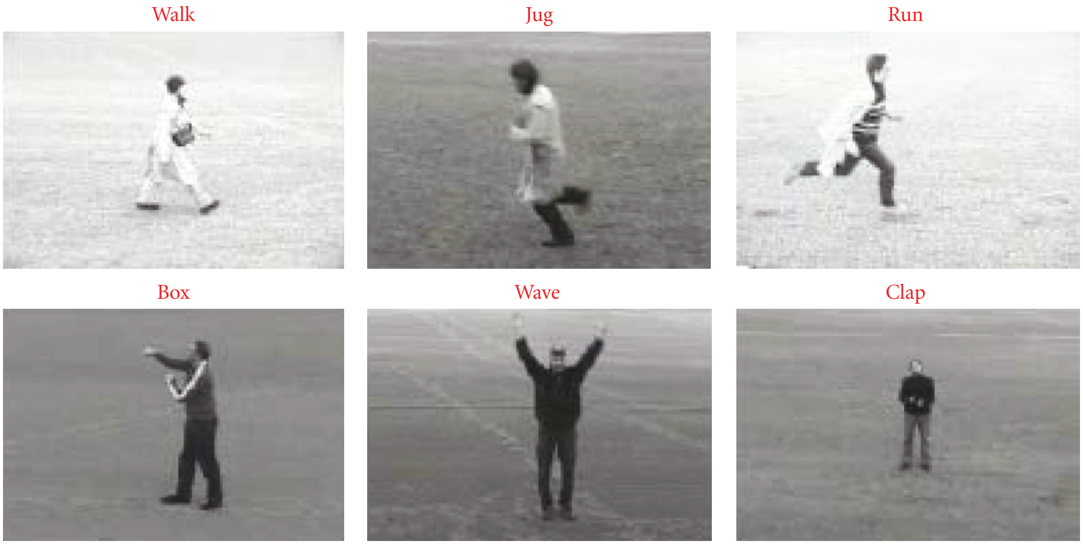

Figure 5: Example sequences from the KTH action dataset.

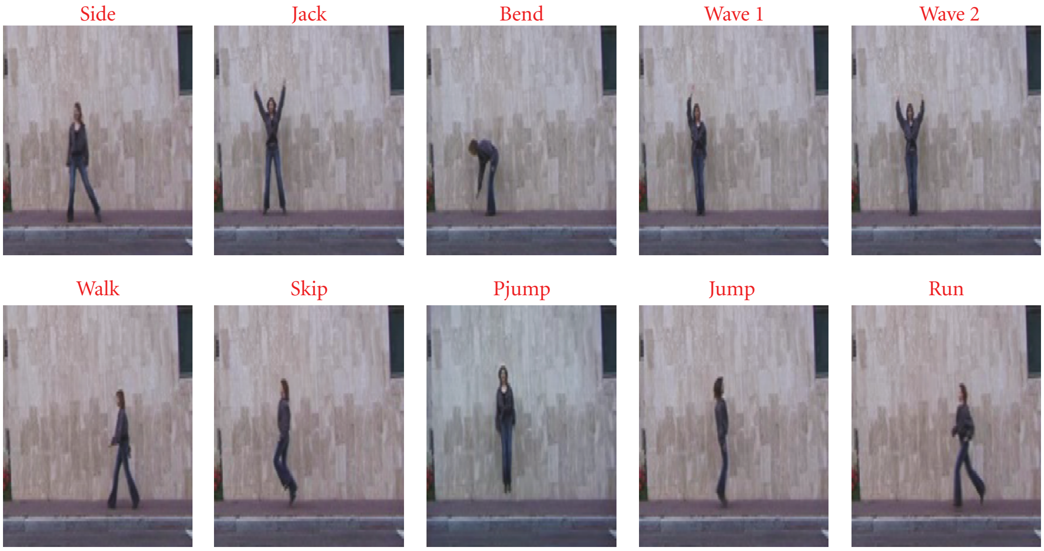

Figure 6: Sample sequences from the Weizmann action dataset.

Table 3: Confusion matrix obtained on 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	0.00	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 recent methods on Weizmann dataset.

Method	Accuracy
Our method	**97.8**%
Fathi and Mori [42]	100%
Bregonzio et al. [38]	96.6%
Zhang et al. [39]	92.8%
Niebles et al. [40]	90.0%
Dollár et al. [37]	85.2%
Kläser et al. [41]	84.3%

then compared with those reported by other investigators in similar studies.

5.1. Experiment-1. We conducted the first experiment using the KTH dataset in which a total of 2391 sequences are involved. The sequences include six types of human actions (i.e., walking, jogging, running, boxing, hand waving and hand clapping). Each of these actions is performed by a total of 25 individuals in four different settings (i.e., outdoors, outdoors with scale variation, outdoors with different clothes, and indoors). All action sequences were taken with a static camera at 25 fps frame rate and a spatial resolution of 160 × 120 pixels over homogeneous backgrounds. Although the KTH dataset is actually not a real-world dataset and thus not so much challenging, there are, to the best of our knowledge, only very few similar datasets already available in the literature with sequences acquired on different environments. An example sequence for each action from the KTH dataset is shown in Figure 5. In order to prepare the simulation and to provide an unbiased estimation of the generalization abilities of the classification process, we partition the sequences for each action into a training set (two thirds) and a test set (one third). This was done such that both sets contained data from all the sequences in the dataset. The SVMs were trained on the training set while the evaluation of the recognition performance was performed on the test set. Table 1 shows the confusion matrix that depicts the recognition results obtained on KTH dataset. As follows from the figures tabulated in Table 1, most actions are correctly classified. Additionally, 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 intuitively seems to be reasonable due to the fact of high similarity between each pair of these actions. To assess the reliability of the proposed approach, our results obtained for this experiment are compared with those obtained by other authors in similar studies (see Table 2). From this comparison, it turns out that our method performs competitively with other state-of-the-art methods and its results are compared favorably with previously published results. Here we would like to contend that all the methods that we compared our method with have used similar experimental setups, thus the comparison is most unbiased.

5.2. Experiment-2. This experiment was conducted using the Weizmann action dataset provided by Blank et al. [34] in 2005. This dataset 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 action 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 25 Hz and last about 2 seconds with image frame size of 180 × 144. Figure 6 shows a sample image for each actions in the Weizmann dataset. Again, in order to provide an unbiased estimate of the generalization abilities of our method, the leave-one-out cross-validation technique was used 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. Then the SVM classifiers with Gaussian radial basis function kernel are trained on the training set, while the evaluation of the recognition performance is performed on the test set. In Table 3, the recognition results obtained on the Weizmann dataset are summarized in a confusion matrix, where correct responses define the main diagonal.

From the figures in the matrix, a number of points can be drawn. 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 is also due to the high closeness or similarity among the actions in each pair of these actions. Once more, in order to quantify the effectiveness of the proposed method, the obtained results 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, one can see that the proposed method is competitive with other state-of-the-art methods. It is worthwhile to mention here that all the methods [37–41] that we compared our method with, except the method proposed in [42], have used similar experimental setups, thus the comparison seems to be meaningful and fair. A final remark concerns the computational time performance of the approach. In both experiments, the proposed action recognizer runs at 28 fps on average (using a 2.8 GHz Intel dual core machine with 4 GB of RAM, running Microsoft Windows 7 Professional). This suggests that the approach is very amenable to working with real-time applications and embedded systems.

6. Conclusion and Future Work

In this paper, such a fuzzy approach to human activity recognition based on keypoint detection has been proposed. Although our model might seem to be similar to previous models of visual recognition, it differs substantially in some important aspects resulting in a considerably improved performance. Most importantly, in contrast to the motion features employed previously, local shape contextual information in this model is obtained through fuzzy log-polar histograms and local self-similarities. Additionally, the incorporation of fuzzy concepts allows the model to be most robust to shape deformations and time wrapping effects. The obtained results are either comparable to or surpass previous results obtained through much more sophisticated and computationally complex methods. Finally the method can offer timing guarantees to real-time applications. However it would be advantageous to explore the empirical validation of the method on more complex realistic datasets presenting many technical challenges in data handling such as object articulation, occlusion, and significant background clutter. Certainly, this issue is very important and will be at the forefront of our future work.

Acknowledgment

This work is supported by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by DFG and Bernstein-Group (BMBF/FKZ: 01GQ0702).

References

[1] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[2] B. Chakraborty, A. D. Bagdanov, and J. Gonzàlez, "Towards real-time human action recognition," in *Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA '09)*, vol. 5524 of *Lecture Notes in Computer Science*, pp. 425–432, June 2009.

[3] 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 (ICCV '03)*, pp. 726–733, October 2003.

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

[5] YU. G. Jiang, C. W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR '07)*, pp. 494–501, July 2007.

[6] 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, 2010.

[7] J. Sullivan and S. Carlsson, "Recognizing and tracking human action," in *Proceedings of the 7th European Conference on Computer Vision (ECCV '02)*, vol. 1, pp. 629–664, Copenhagen, Denmark, May-June 2002.

[8] 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.

[9] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "Human activity recognition: a scheme using multiple cues," in *Proceedings of the 6th International, Symposium on Visual Computing (ISVC '10)*, vol. 6454 of *Lecture Notes in Computer Science*, pp. 574–583, Las Vegas, Nev, USA, November-December 2010.

[10] C. Thurau and V. Hlaváč, "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)*, June 2008.

[11] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "Human activity recognition via temporal moment invariants," in *Proceedings of IEEE Symposium on Signal Processing and Information Technology (ISSPIT '10)*, 2010.

[12] 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.

[13] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2046–2053, San Francisco, Calif, USA, June 2010.

[14] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 925–931, October 2009.

[15] 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.

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

[17] 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.

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

[19] 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.

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

[21] 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.

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

[23] 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.

[24] 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, pp. 717–721, 2002.

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

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

[27] 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.

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

[29] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, 1988.

[30] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151–172, 2000.

[31] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[32] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

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

[34] 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)*, pp. 1395–1402, October 2005.

[35] Y. Wang and G. Mori, "Max-Margin hidden conditional random fields for human action recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 872–879, June 2009.

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

[37] 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.

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

[39] 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. 5305 of *Lecture Notes in Computer Science*, no. 4, pp. 817–829, October 2008.

[40] J. C. Niebles, H. Wang, and LI. 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.

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

[42] 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.