Toward Robust Action Retrieval in Video

Samy Sadek[1]
Samy.Bakheet@ovgu.de
Ayoub Al-Hamadi[1]
Ayoub.Al-Hamadi@ovgu.de
Bernd Michaelis[1]
http://www.iesk.ovgu.de
Usama Sayed[2]
http://www.aun.edu.eg

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

Abstract

Retrieving human actions from video databases is a paramount but challenging task in computer vision. In this work, we develop such a framework for robustly recognizing human actions in video sequences. The contribution of the paper is twofold. First a reliable neural model, the Multi-level Sigmoidal Neural Network (MSNN) as a classifier for the task of action recognition is presented. Second we unfold how the temporal shape variations can be accurately captured based on both temporal self-similarities and fuzzy log-polar histograms. When the method is evaluated on the popular KTH dataset, an average recognition rate of 94.3% is obtained. Such results have the potential to compare very favorably to those of other investigators published in the literature. Further the approach is amenable for real-time applications due to its low computational requirements.

1 Introduction

Visual recognition and interpretation of human-induced actions and events are among the most active research areas in computer vision, pattern recognition, and image understanding communities [22]. Undoubtedly, the automatic recognition and understanding of actions in video sequences are still an underdeveloped area due to the lack of a robust general purpose model and the approaches proposed in literature remain limited in their ability. For this, much research still needs to be undertaken to address the ongoing challenges. It is clear that developing good algorithms for solving the problem of human action recognition would yield huge potential for a large number of potential applications (e.g., human-computer interaction, video surveillance, gesture recognition, and robot learning and control). 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 [15]. 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 trade-off between accuracy and real-time requirements.

BMVC 2010 doi:10.5244/C.24.44

The remainder of this paper commences by briefly reviewing the most relevant literature in Section 2, In Section 3, the MSNN model as a classifier is presented. Section 4 describes the details of the proposed action recognition method. The experimental results that assess the effectiveness of the method are presented and analyzed in Section 5. At last, in Section 6, we conclude the paper and outline future research directions.

2 Related work

Recent years have witnessed a flurry of interest in more research on the analysis and interpretation of human motion boosted by the rise of security concerns and increasing ubiquity and affordability of digital media production equipment [6]. Although a lot of papers have been published in this field, the problem is still open for investigation, posing great challenges to the researchers. Thus more research needs to be conducted to come around it. Human action can be generally recognized using various visual cues such as motion [7, 9, 20, 25] and shape [30]. Scanning the literature, one notices that a significant body of work in action recognition focuses on using spatial-temporal keypoints and local feature descriptors [8, 16, 18, 21, 23]. The local features are extracted from the region around each keypoint detected by the keypoint 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 [7, 25], periodic motions are detected and classified to recognize actions. In [20] 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 [5], Bobick and Davis use temporal templates, including motion-energy images and motion-history images to recognize human movement. In [29] the authors detect the similarity between video segments using a space-time correlation model. While in [26], Rodriguez *et al.* present a template-based approach using a Maximum Average Correlation Height (MACH) filter to capture intra-class variabilities. Jhuang *et al.* [14] perform actions recognition by building a neurobiological model using spatio-temporal gradients. In [27], 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 [10, 13, 19, 24]. In addition, there is more research focusing on using generative topic models for visual recognition based on the so-called Bag-of-Words (BoW) model. The key 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 Correlated Topic Models (CTM) [3], Latent Dirichlet Allocation (LDA) [4] and probabilistic Latent Semantic Analysis (pLSA) [12].

3 Multi-level neural networks

In this section, the neural model used by the proposed action recognition system is described. Neural network classifier has many advantages over other competitive machine learning classifiers. Some of these advantages include the high rapidity, easiness of training, realistic generalization capability, high selectivity and great capability to create arbitrary partitions of feature space. However, the neural model, in the standard form, might have low classi-

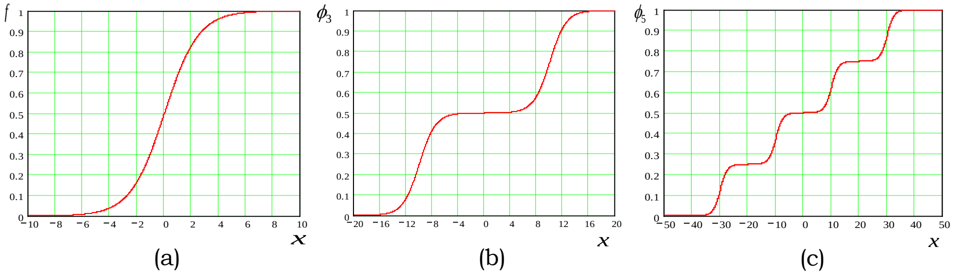

Figure 1: Standard sigmoidal function and its Multi-level versions:(a) Standard sigmoidal function, (b) Multi-level function for $r = 3$, (c) Multi-level function for $r = 5$.

fication accuracy and poor generalization properties because its neurons employ a standard bi-level function that gives only two values (i.e., binary responses) [2]. To relax this restriction and allow the neurons to generate multiple responses, a new functional extension for the standard sigmoidal functions should be developed. This extension is termed the Multi-level Activation Function, and the model that uses this extension is termed the Multi-level Sigmoidal Neural Network (MSNN). There are several multi-level versions corresponding to several standard activation functions. It is straightforward to derive a multi-level version from a given bi-level standard sigmoidal activation function. Let us suppose the general form of the standard sigmoidal function $f(x)$ is given by (see Figure 1(a))

$$f(x) = \frac{1}{1+e^{-\beta x}} \tag{1}$$

where β is a constant (sometimes called a steepness factor). Therefore the multi-level form can easily be derived from the previous standard form as follows

$$\varphi_r(x) \leftarrow f(x) + (\lambda - 1) f(c) \tag{2}$$

where λ is an integer index running from 1 to $r-1$; r is the number of levels, and c is an arbitrary constant. Multi-level sigmoidal functions for $r = 3$ and 5 are depicted in Figure 1(b) and Figure 1(c) respectively.

4 Proposed recognition framework

In this section, the proposed action recognition system is presented, which is schematically illustrated in Figure 2. As seen from the block diagram, the process of action recognition systematically runs as follows. For each action snippet, the keypoints are first detected using the Harris corner detector. To make the method more robust against time warping effects, action snippets are temporally divided into a number of overlapping time-slices 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 relevant to the current task, the final features (so-called hybrid features) fed into the MSNN classifier are constructed by combining both local and global features. The next subsections explain each of these components of the system in further detail.

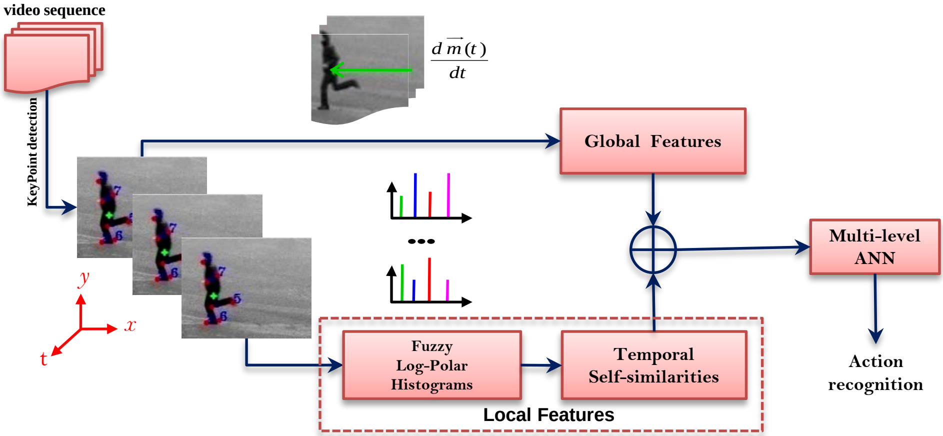

Figure 2: Main components of the human action recognition system.

4.1 Pre-processing and keypoint detection

Preprocessing generally aims to prepare the representative features desirable for knowledge generation. The frames of each video clip containing a person performing a certain action are smoothed by Gaussian convolution with a kernel of size 3×3 and variance $\sigma = 0.5$ to wipe off noise and weaken image distortion. Many previous evaluations of keypoint detectors [28], have revealed that the Harris detector still demonstrates its superior performance to that of many competitors. However Harris detector originally is not invariant to scale changes , so that it is highly needed to make Harris detector more robust to scale changes. The idea is quite simple. This can be done by joining the original Harris detector with automatic scale selection [11]. In this case, the local second moment matrix quantifying scale adaptive Harris detector are adapted to scale changes to make it independent of image resolution. The scale-invariant keypoints are then detected using the scale-adaptive Harris detector. The obtained keypoints are then filtered, so that under a certain amount of additive noise, only stable and more localized keypoints are retained. This is done in two steps: first low contrast keypoints are discarded, and second isolated keypoints not satisfying the spatial constraints of feature points are excluded (because they are out of the spatial scope of a target object).

4.2 Extracted local features

Feature extraction is the most important part of the action recognition procedure because the accuracy of the recognizer often relies on how well the features are extracted. In this section, the features that describe the local spatio-temporal shape characteristics are described.

4.2.1 Fuzzy log-polar histograms

Initially, we temporally divide a video sequence into several time-slices to compensate the time warping effects. These slices are defined by linguistic intervals. Gaussian membership functions are used to describe such intervals, which are given by

$$\mu_j(t;\varepsilon_j,\sigma,m) = e^{-\frac{1}{2}\left|\frac{t-\varepsilon_j}{\sigma}\right|^m}, \quad j = 1,2,\ldots,s \tag{3}$$

where ε_j, σ, and m are the center, width, and fuzzification factor, respectively, and s is the total number of time-slices. 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 3. By using these fuzzy functions, not only are temporal information extracted successfully, the performance decline due to time warping effects can also be dramatically weakened or obliterated. To extract the local features of the shape representing action at an instance of time, the basic idea of the shape context proposed first by Belongie *et al.* in [▯] should be improved. Their shape contexts were applied to images after aligning transforms. The shape descriptor presented in this work calculates the log-polar histograms on condition that they are invariant to simple transforms like scaling, rotation and translation. These histograms are normalized for all affine transforms as well. The idea behind a modified shape context is based on computing rich descriptors for fewer keypoints. Furthermore the shape context is reasonably extended by combining local descriptors with fuzzy memberships functions and temporal self-similarities paradigms. In general, human action is composed of a sequence of poses over time. Reasonable estimate of a human 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 a set of sampled keypoints $\{(x_i, y_i)\}_{i=1}^{n}$ representing an action pose at an instance time t. Then, log-polar coordinates (ρ_i, η_i) for each keypoint $p_i \in B$, are given by

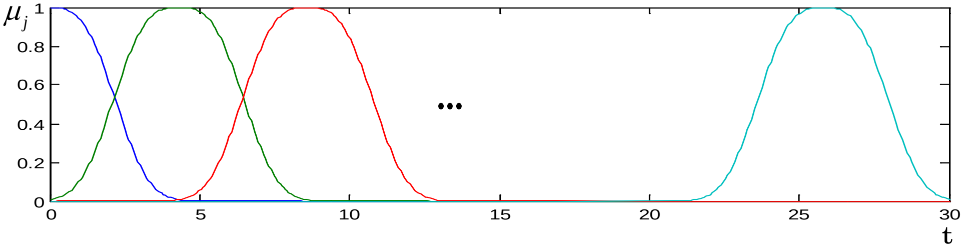

Figure 3: Gaussian membership functions used to represent the temporal intervals, with $\varepsilon_j = \{0, 4, 8, \ldots\}$, $\sigma = 2$, and $m = 3$.

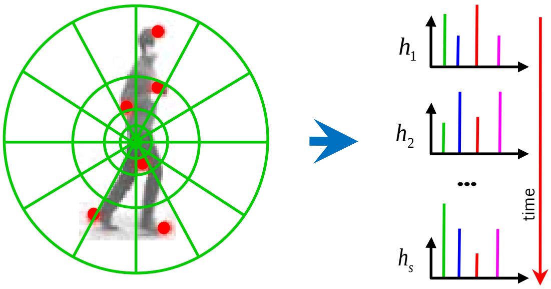

Figure 4: Fuzzy Log-polar histograms representing the spatio-temporal shape contextual information of "jogging" action.

$$\rho_i = \log\left(\sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}\right), \; \eta_i = \arctan\left(\frac{y_i - y_c}{x_i - x_c}\right), \; i = 1, 2, \ldots, n \tag{4}$$

where (x_c, y_c) is the center of mass, which is invariant to scaling, rotation or translation. Hence the angle is computed with respect to a horizontal line passing through the center of gravity. To calculate the modified shape context of action pose, a log-polar histogram is overlaid on the shape of pose as illustrated in Figure 4. Thus the fuzzy log-polar histograms representing action at a time-slice j is constructed by using the membership functions:

$$h_j(k_1,k_2) = \sum_{\substack{\rho_i \in bin(k_1),\\ \eta_i \in bin(k_2)}} \mu_j(t_i),\ j = 1,2,\ldots,s \tag{5}$$

Each of these histograms is a 2-d matrix of $d\rho \times d\eta$ dimensional, where $d\rho$ and $d\eta$ are the radial and angular dimensions, respectively. By applying a simple linear transformation on the indices k_1 and k_2, the 2-d histograms is converted into one dimensional (1-d) as follows

$$h_j(k) = h_j(k_1 d\eta + k_2), \quad k = 0,1,\ldots,d\rho\, d\eta - 1 \tag{6}$$

Next, the resulting histograms are normalized to the integral value of unity to achieve robustness to scale variations and to reduce the influence of illumination. The normalized histograms obtained can now be used as shape contextual information for classification and matching. Many conventional approaches in various computer vision applications directly combine such normalized histograms to obtain one feature vector per video clip, which, in turn, can be classified by any classification algorithm such as Bayes decision rule, ANN, HMM, SVM, etc. In contrast, in this research, we aim to enrich these histograms with the self-similarity analysis after using a suitable distance function to measure similarity (more precisely dissimilarity) between each pairs of these histograms. This is of most importance to accurately discriminate between temporal variations of different human actions.

4.2.2 Similarity measure of action snippets

Video analysis is seldom carried out directly on row video data. Instead feature vectors extracted from small portions of video (i.e., so-called 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, Mahalanobis metric, etc. Whilst such metrics may 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 relative changes are considered. Such metric is based on Pearson Linear Correlation (PLC), and given by:

$$\rho(\vec{u},\vec{v}) = \frac{\sum_{i=1}^{m}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{m}(u_i - \bar{u})^2 \sum_{i=1}^{m}(v_i - \bar{v})^2}} \tag{7}$$

where $\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i$ and $\bar{v} = \frac{1}{m}\sum_{i=1}^{m} v_i$ are the means of $\vec{u}$ and $\vec{v}$ respectively. The expression profiles are shifted down (by subtracting the means) and scaled by the standard deviations (i.e., the data have $\mu = 0$ and $\sigma = 1$). Note that Pearson linear correlation (PLC) is a measure that is invariant to scaling and shifting of the expression values. The value of PLC is constrained between -1 and $+1$ (perfectly anti-correlated and perfectly correlated). This is a similarity measure, but it can be easily enforced to be a dissimilarity measure by:

$$s(\vec{u},\vec{v}) = \frac{1-\rho(\vec{u},\vec{v})}{2} \tag{8}$$

4.2.3 Temporal self-similarities construction

To reveal the inner structure of a human action in a video clip, second statistical moments (i.e., mean and variance) are not enough. Instead, self-similarity analysis seems to be a much more appropriate paradigm, which can formally formulated as follows. Given a histogram series $H = \{h_1, h_2, \ldots, h_m\}$ that represents m time-slice of an action, then the temporal self-similarity matrix is given by

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

where $s_{ij} = s(h_i, h_j), i, j = 1, 2, \ldots, m$. The diagonal entries are zero, as $s(h_i, h_i) = 0$. Moreover since $s_{ij} = s_{ji}$, S is a symmetric matrix.

4.3 Fusion of global and local features

It follows from the previous subsections that the features extracted based on the fuzzy log-polar histograms and temporal self-similarities are emphasized. Such features extracted on each temporal slice are considered as temporally local features while those features extracted during the entire motion of the action actor in a video sequence are regarded as temporally global features. The use of global features has proven to be very beneficial in many applications of object recognition. This motivates us to fuse global and local features to form the final MSNN classifier. All the global features extracted here are based on calculating the center of mass $\vec{m}(t)$ that delivers the center of motion. Therefore the global features $\vec{F}(t)$ describing the general distribution of motion are given by

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

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 global and 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

In this section action recognition is modeled as a multi-class classification task, where there is one class for each human 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. The MSNN classifier described in Section 3 is used for the current classification task due to its outstanding generalization capability and reputation of a highly accurate paradigm. The basic model of MSNN is a multi-layer feedforward network with two hidden layers of 20 neurons each, which is most similar to the classical network structures but with improving in the hidden-unit adaptive activation functions (i.e., Multi-level

Activation Functions). Before the training phase, the classifier starts with random weights at the connections between the neurons. The learning procedure followed by the MSNN classifier is similar to the well-known backpropagation procedure. In our approach, several classes of actions are created. During the learning stage, the MSNN classifier is trained using the features extracted from the action snippets in the training dataset. The up diagonal elements of the similarity matrix representing the local features are first transformed into plain vectors, and then fused with the global features. All feature vectors are finally fed into the MSNN classifier to distinguish all action classes. After the learning stage is finished, the system is able to recognize and identify unseen actions. In fact, the classifier produces a real value between 0 and 1, which can easily be binarized by using a predetermined threshold.

5 Experimental results

In this section, the proposed method is evaluated on the popular KTH dataset [18]. To illustrate the effectiveness of our approach, the results obtained are compared with those of other similar state-of-the-art methods in the literature. The KTH dataset contains a total of 2391 sequences, which, in turn, are comprised of six types of human actions (i.e., walking, jogging, running, boxing, hand waving and hand clapping). The actions are 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 sequences were acquired by a static camera at 25fps and a spatial resolution of 160 × 120 pixels over homogeneous backgrounds. There is, to the best of our knowledge, no other similar action dataset already available in the literature with sequences acquired on different environments. Typical example frames from the KTH dataset for each action in each scenario can be seen in Figure 5. In order to prepare the simulation and to provide an unbiased estimation of the generalization abilities of the classification process, the sequences, for each action, were divided into a training set (two

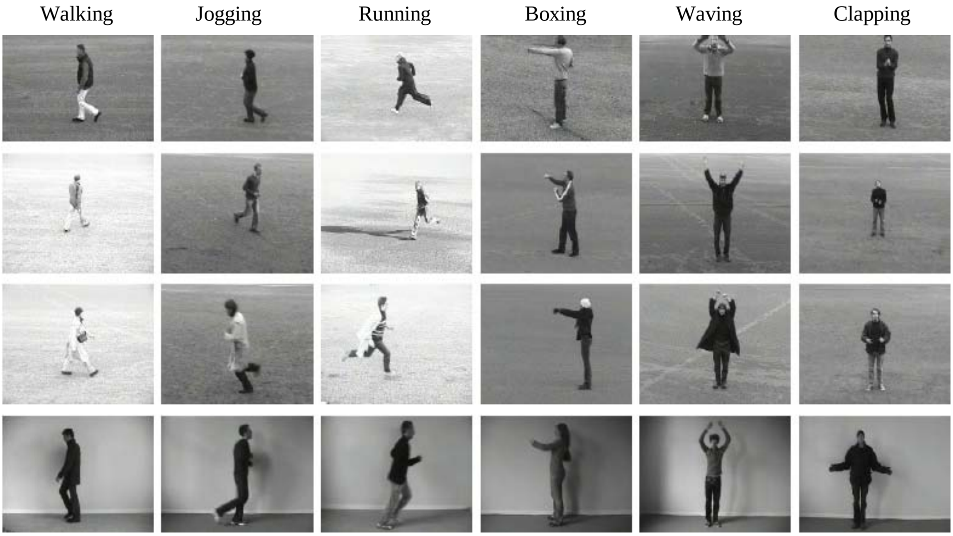

Figure 5: Sample frames of the KTH dataset actions with different scenarios.

ACTION	Walking	Running	Jogging	Waving	Clapping	Boxing
Walking	0.99	0.00	0.01	0.00	0.00	0.00
Running	0.00	0.96	0.04	0.00	0.00	0.00
Jogging	0.07	0.06	0.87	0.00	0.00	0.00
Waving	0.00	0.00	0.00	0.95	0.00	0.05
Clapping	0.00	0.00	0.00	0.00	0.94	0.06
Boxing	0.00	0.00	0.00	0.00	0.02	0.98

(a)

Related studies	Accuracy
Our method	**94.3%**
Ke *et al.* [16]	63.0%
Dollàr *et al.* [8]	81.2%
Rodriguez *et al.* [26]	88.6%
Jhuang *et al.* [14]	91.7%
Liu *et al.* [21]	92.8%
Kim *et al.* [17]	95.3%

(b)

Figure 6: Confusion matrix obtained with the developed method (a) and the comparison of our results with those of six other widely quoted studies from the literature (b).

thirds) and a test set (one third). This was done such that both sets contained data from all 540 sequences. The MSNN was trained on the training set while the evaluation of the recognition performance was performed on the test set. The confusion matrix depicting the results of action recognition obtained with the proposed method and the comparison of our results with those of other previously published studies in the literature are shown in Figure 6 (a) and Figure 6 (b) respectively. As follows from the figures tabulated above, most of 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 in Figure 6, it turns out that the proposed method performs competitively with other state-of-the-art methods and the results obtained compare very favorably to those reported in the literature. Notably all the methods we compare to have used the same experimental setup, except the method of Kim *et al.* [17] that has achieved an impressive accuracy (i.e., 95.3%), but spatio-temporal alignment of video sequences was manually carried out. Thus the comparison is reasonably fair. Finally, using the self-similarity analysis provides an efficient way of doing feature reduction that allows the method to be processed in real-time and thus amenable to real-time applications.

6 Conclusion and future work

This paper has introduced such a fuzzy framework to recognize human actions from video sequences. In the proposed method, the temporal shape contextual information of action is obtained based on the local temporal self-similarities defined over fuzzy log-polar histograms. The incorporation of the fuzzy membership functions makes the method quite robust to shape deformations and time wrapping effects. When validated on the KTH action dataset, the results obtained compare very favorably with those achieved by much more sophisticated and computationally complex methods. Furthermore, the method runs in real-time and thus can offer latency guarantees to real-time applications. However, it would be beneficial to explore the empirical validation of the method on more complex realistic datasets presenting many technical challenges in data handling. These issues are of crucial interest and worthwhile to be investigated in our future work.

Acknowledgment

This work is supported by Förschungspraemie (BMBF, FKZ: 03FPB00213), BMBF Bernstein-Group (FKZ: 01GQ0702) and Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by DFG.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on PAMI*, 24(4):509–524, 2002.

[2] S. Bhattacharyya, P. Dutta, and U. Maulik. Object extraction using self organizing neural network with a multi-level sigmoidal transfer function. In *Proc. 5th International Conference on Advances in Pattern Recognition*, pages 435–438, 2003.

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

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on PAMI*, 23(3):257–267, 2001.

[6] B. Chakraborty, A. D. Bagdanov, and J. Gonzàlez. Towards real-time human action recognition. In *Proc. of 4th Iberian Conference on PRIA, LNCS*, pages 425–432, 2009.

[7] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on PAMI*, 22(8):781–796, 2000.

[8] P. Dòllar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. of IEEE Workshop on VS-PETS*, pages 65–72, 2005.

[9] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE Int. Conference on Computer Vision*, volume 2, pages 726–733, 2003.

[10] X. Feng and P. Perona. Human action recognition by sequence of movelet codewords. In *Proc. of 3D Data Processing Visualization and. Transmission*, pages 717–721, 2002.

[11] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the Fourth Alvey Vision Conference*, pages 147–151, 1988.

[12] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of SIGIR*, pages 50–57, 1999.

[13] N. Ikizler and D. Forsyth. Searching video for complex activities with finite state models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[14] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *IEEE Int. Conference on Computer Vision*, pages 257–267, 2007.

[15] I. N. Junejo, E. Dexter, I. Laptev, and P. Prez. Cross-view action recognition from temporal self-similarities. *Proc. European Conference Computer Vision (ECCVŠ08)*, 2:293–306, 2008.

[16] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. of ICCV*, volume 1, pages 166–173, 2005.

[17] T. K. Kim, S. F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *Proc. CVPR*, 2007.

[18] I. Laptev and P. Perez. Retrieving actions in movies. In *Proc. of ICCV*, pages 1–8, 2007.

[19] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *Proc. of CVPR*, pages 1–8, 2007.

[20] L. Little and J. E. Boyd. Recognizing people by their gait: The shape of motion. *International Journal of Computer Vision*, 1(2):1–32, 1998.

[21] J. Liu and M. Shah. Learning human actions via information maximization. In *CVPR*, 2008.

[22] 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*, 104(2):90–126, 2006.

[23] S. Nowozin, G. Bakir, and K. Tsuda. Discriminative subsequence mining for action classification. In *Proc. of ICCV*, pages 1919–1923, 2007.

[24] N. Olivera, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. and Image Underst.*, 96 (2):163–180, 2004.

[25] R. Polana and R. C. Nelson. Detection and recognition of periodic, non-rigid motion. *International Journal of Computer Vision*, 23(3):261–282, 1997.

[26] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[27] K. Schindler and L. Van Gool. Action snippets: How many frames does action recognition require? In *Proc. of CVPR*, pages 1–8, 2008.

[28] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[29] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proc. of CVPR*, volume 1, pages 405–412, 2005.

[30] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *Proc. of ECCV LNCS 2352*, volume 1, pages 629–644, 2002.