

# Recognition of Human actions Based on Temporal Motion Templates

## Abstract

Despite their attractive properties of invariance, robustness and reliability, statistical descriptions from the temporal templates are not hitherto paid the attention they deserve in the action recognition literature. In this paper, we propose an innovative approach for action recognition, where a novel fuzzy representation based on temporal motion templates is developed to model human actions as time series of low-dimensional descriptors. An NB (Naïve Bayes) classifier is trained on these features for action classification. When evaluated on our dataset incorporating a large collection of video data, the approach yields encouraging results that compare favorably with those reported in the literature, while maintaining real-time performance.

*Keywords:* Skin Segmentation; AdaBoost; Gesture Detection; Pattern Recognition.

2010 Mathematics Subject Classification: 53C25; 83C05; 57N16

## 1 Introduction

The automatic recognition of human actions in unconstrained settings is still an underdeveloped area due to the lack of a general purpose model. Furthermore, most approaches in the literature remain limited in their ability. For this, much research still needs to be undertaken to address the ongoing challenges. The non-rigid nature of human body and clothes in video sequences, that results from drastic illumination changes, changing in pose, and erratic motion patterns presents a grand challenge to action recognition (Wang et al., 2015)

It is clear that developing good algorithms for recognizing human actions in real-world scenes provides huge potential for a large number of real-life applications, such as human-computer

interaction, video surveillance, gesture recognition, and robot learning and control (Sadek et al., 2012b). While the real-time performance is a major concern in computer vision, especially for embedded computer vision systems, the majority of recognition systems often employ sophisticated feature extraction and learning techniques, creating a barrier to the real-time performance of these systems. In this work, we attempt to address the recognition of human actions in real-world scenarios which is an important but challenging problem with prosperous applicability into human-computer interactions and security industry.

The remainder of this paper is structured as follows. Section 2 briefly reviews previous work. The proposed framework for action recognition is described in Section 4. In Section 5, the results of the preliminary experiments conducted to evaluate the stability of the system and its effectiveness in recognizing actions are presented and discussed. Finally, Section 6, concludes the paper by results summary and possible extensions.

## 2 Related literature

In the past four decades or so, a great deal of research has been conducted into the recognition of human actions. Despite these years of work, the problem still provides a big challenge to the researchers. Human actions can generally be recognized using various visual cues such as motion (Efros et al., 2003) and shape (Lu et al., 2009). Scanning the literature, we notice that a great deal of research focuses on using spatial-temporal keypoints and local feature descriptors (Liu and Shah, 2008). Another thread of research focuses on analyzing patterns of motion to recognize actions. For instance, in (Little and Boyd, 1998) the authors analyze the periodic structure of flow patterns for gait recognition.

Alternatively, some researchers have opted to use both motion and shape cues. For example, in (Shechtman and Irani, 2005) the authors detect the similarity between video segments using a space-time correlation model. While in (Rodriguez et al., 2008), Rodriguez et al. present a template-based approach using a Maximum Average Correlation Height (MACH) filter to capture intra-class variabilities, Jhuang et al. (Jhuang et al., 2007) perform action recognition by building a neurobiological model using spatio-temporal gradient. In parallel, a significant amount of work is targeted at modelling and understanding human motions by constructing elaborated temporal dynamic models (Laxton et al., 2007; Ikizler and Forsyth, 2007). There is also a research area that concentrates on using generative topic models for visual recognition based on the so-called Bag-of-Words (BoW) model.

## 3 Temporal Motion Templates

Temporal templates are 2D images constructed from image sequences, which effectively reduce a 3D spatio-temporal space to a 2D representation (Bobick and Davis, 2001). They are conformed with the successive images differences to show where and when motion in the image sequence occur; while one dimension is eliminated, the temporal information is retained and depicted in the related 2D image. To construct temporal templates, either the camera and the background are assumed to be static, or the motion of the object of interest is assumed to be separable from the motion induced by camera and/or background movements. When a temporal template is constructed without maintaining the information about the time instance at which the motion occurred, in this case the resulting temporal template is referred to as a Motion Energy Image (MEI). Instead, when the temporal information (i.e., motion history) is preserved through assigning different intensities to different moments of the motion, then the temporal template is termed a Motion History Image (MHI), see Figure 1.

A serious drawback inherent to the originally proposed temporal template approach (Bobick and Davis, 2001) is the so-called "motion self-occlusion" problem due to overwriting. To show this problem,

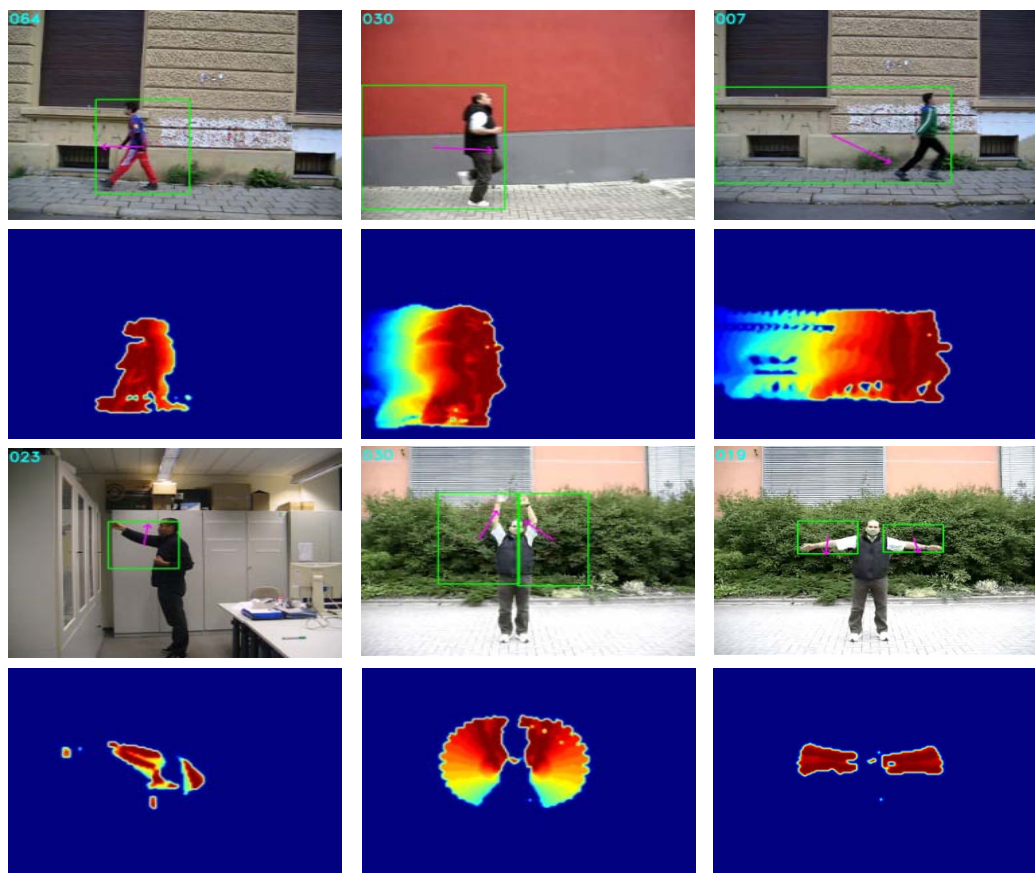


Figure 1: Action moves along with their MHIs; from left to right the top row actions are: walking, jogging, running, boxing, waving, and clapping.

let us consider, for instance, a motion occurs on a spatial location  $\chi$  at two time instances  $t_1$  and  $t_2$  where  $t_2 > t_1$ , then the recent motion that occurs at  $t_2$  will overwrite the prior motion (occurring at  $t_1$ ). One way to circumvent this problem is to record motion history at multiple time intervals (i.e., multilevel MHIs), instead of recording the motion history once for the entire image sequence (single MHI).

## 4 Proposed Recognition Approach

In this section, the proposed method for action recognition is described. The main steps in this method can be summarized as follow. First, temporal templates called Motion History Images (MHI's) are constructed from the image sequence. Then, few statistical low-level features characterizing human motion parametrically are extracted from the temporal templates. For dimensionality reduction, we present an adaptive fuzzy feature selection technique to reduce the size of the extracted features without or with little recognition performance degradation. Finally, feature vectors are applied to train an NB classifier for action recognition. The technical details of each of these steps are provided in the following subsections.

## 4.1 Temporal Template Construction

Let  $I(x, y, t)$  be the image brightness that changes in time to provide an image sequence. Further, let  $D(x, y, t)$  be the binary image resulting from brightness variation detection ( e.g., obtained from frame subtraction:  $D(x, y, t) = |I(x, y, t) - I(x, y, t \pm \Delta)|$ , where  $(x, y)$  are the spatial image coordinates and  $\Delta$  is the difference distance). Specifically, within an MHI denoted as  $H_\tau$  with  $\tau$  decides the temporal duration of the MHI (e.g., in terms of frames), the intensity value at each point is a function of the motion properties at the corresponding spatial location in the image sequence. Therefore,  $H_\tau(x, y, t)$  can be computed from an update function  $\psi(x, y, t)$  with the following recursive formula:

$$H_\tau(x, y, t) = \begin{cases} \tau, & \text{if } \psi(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t - 1) - \varepsilon), & \text{o.w} \end{cases} \quad (4.1)$$

where  $\psi(x, y, t)$  signals object's presence (or motion) in the current frame and  $\varepsilon$  is the decay parameter. The MHI is generated from  $D(x, y, t)$  using a threshold  $\xi$ :

$$\psi(x, y, t) = \begin{cases} 1, & \text{if } D(x, y, t) > \xi \\ 0, & \text{o.w} \end{cases} \quad (4.2)$$

It is worthwhile mentioning here that if the information about when the motion of interest begins and ends is not available, it may be necessary to vary the observed period  $\tau$  and then attempt to classify all MHIs. While on the contrary, when the beginning and the end of action are known and they coincide with the duration of image sequence, there will be no need to vary the parameter  $\tau$ . The temporal behavior can be efficiently normalized by distributing the gray values inside the MHI over the available range (e.g. [0-255], assuming 8-bit gray scale level representation). Successively, variations in display duration of an action unit can be wiped out. This allows actions having a different period but otherwise identical to be compared against to each other.

Initially, each input image sequence may comprise of different number of frames. Hence, the history levels in MHIs would have different numbers from one sequence to another. In order to compare the sequences properly, the multilevel MHI (MMHI) approach tries to create all MHIs such that they have a fixed number of history levels  $n$ . Therefore each image sequence is sampled to  $n + 1$  frames. Using the known parameter  $n$ , the MHI operator is modified into:

$$H(x, y, t) = \begin{cases} \alpha t, & \text{if } \psi(x, y, t) = 1 \\ H(x, y, t - 1), & \text{o.w} \end{cases} \quad (4.3)$$

where  $\alpha = 255/t$  is the intensity step between two history levels. Note that (4.3) implies that for  $t = 0$ ,  $H(x, y, t) = 0$ . With a MMHI, the main objective is to encode motion occurring at different time instances on the same pixel location such that it is uniquely decodable afterwards. To achieve this objective, a simple bitwise coding scheme is used. If motion occurs at pixel location  $(x, y)$  and time  $t$ , the term  $2^{t-1}$  is added to the old value of the MMHI as follows,

$$M(x, y, t) = M(x, y, t - 1) + 2^{t-1}\psi(x, y, t) \quad (4.4)$$

where  $M(x, y, t) = 0$  for  $t = 0$ . With this bitwise coding scheme, it is possible to separate multiple motions occurring at the same location. Figure 1 shows six actions and their MHIs. In this figure, we observe that the regions of newer motion appear more reddish than those of old motion in MHIs.

## 4.2 Feature Extraction

The first set of extracted feature is statistical features, which includes: (1) The filling ratio of the bounding box around the largest ROI of the MHI region. This feature represents the percentage of the bounding box of the motion occurring at a specific time instance, which provides us with a significant information about motion shape. (2) The average direction in each moving region, which is the estimated angle between the motion orientation and the horizontal axis. This direction is

calculated from the weighted orientation histogram, where a recent motion has a larger weight and the motion occurred in the past has a smaller weight, as recorded in MHI. (3) The ratio between the width and height of the bounding box around the motion part of MHI. This feature serves as a most characteristic to the motion shape. (4) The shortest distance between the start and end-point of the motion trajectory.

As we are interested to represent actions in more local level, we propose to define an additional set of geometric features. To achieve this goal, the bounding box around the detected motion is partitioned into a number of sub-regions. Then, a set of affine geometric invariants can be derived from each of the sub-regions. This set is invariant under affine transformations and moment-based descriptors (Flusser and Suk, 1993), and is given as,

$$\begin{aligned}
 I_1 &= \frac{1}{\eta_{00}^4} [\eta_{20}\eta_{02} - \eta_{11}^2] \\
 I_2 &= \frac{1}{\eta_{10}^2 \eta_{00}^2} [\eta_{03}^2 \eta_{30}^2 - 6\eta_{30}\eta_{21}\eta_{12}\eta_{03} \\
 &\quad + 4\eta_{30}\eta_{12}^3 + 4\eta_{03}\eta_{21}^3 - 3\eta_{21}^2 \eta_{12}^2] \\
 I_3 &= \frac{1}{\eta_{00}^7} [\eta_{20}(\eta_{21}\eta_{03} - \eta_{12}^2) - \eta_{11}(\eta_{30}\eta_{03} \\
 &\quad - \eta_{21}\eta_{12}) + \eta_{02}(\eta_{03}\eta_{12} - \eta_{21}^2)] \\
 I_4 &= \frac{1}{\eta_{00}^{11}} [\eta_{20}^3 \eta_{03}^2 - 6\eta_{20}^2 \eta_{11} \eta_{12} \eta_{03} \\
 &\quad - 6\eta_{20}^2 \eta_{02} \eta_{21} \eta_{03} + 9\eta_{20}^2 \eta_{02} \eta_{12}^2 \\
 &\quad + 12\eta_{20} \eta_{11}^2 \eta_{21} \eta_{03} + 6\eta_{20} \eta_{11} \eta_{02} \eta_{30} \eta_{03} \\
 &\quad + 18\eta_{20} \eta_{11} \eta_{02} \eta_{30} \eta_{12} - 8\eta_{11}^3 \eta_{30} \eta_{03} \\
 &\quad - 6\eta_{20} \eta_{02}^2 \eta_{30} \eta_{12} + 9\eta_{20} \eta_{02}^2 \eta_{21}^2 \\
 &\quad + 12\eta_{11}^2 \eta_{02} \eta_{30} \eta_{12} - 6\eta_{11} \eta_{02}^2 \eta_{30} \eta_{12} \\
 &\quad + \eta_{02}^3 \eta_{30}^3] \\
 I_5 &= \frac{1}{\eta_{00}^6} [\eta_{40}\eta_{04} - 4\eta_{31}\eta_{13} + 3\eta_{22}^2] \\
 I_6 &= \frac{1}{\eta_{00}^9} [\eta_{40}\eta_{04}\eta_{22} + 2\eta_{31}\eta_{13}\eta_{22} - \eta_{40}\eta_{13}^2 \\
 &\quad - \eta_{04}\eta_{13}^2 - \eta_{22}^3]
 \end{aligned} \tag{4.5}$$

where  $\eta_{pq}$  is the central moment of order  $p + q$ .

### 4.3 Fuzzy Feature Selection

The key idea of our method for feature selection is to temporally decompose action sequences (i.e. snippets) into a finite number of time slices in a fuzzy way. This would enable the approach to achieve better feature reduction ratios without loss in recognition accuracy. Formally, at each time instant  $t$ , feature vector can be created from extracted features as follows,

$$\mathbf{f}_t = (f_{t;1}, f_{t;2}, \dots, f_{t;k})^\top \tag{4.6}$$

where  $k$  is the total number of features at time  $t$ . Since the features in (4.6) are computed at each time instant of a given snippet, the snippet can then be represented as a time series of these features:  $A = \{\mathbf{f}_t\}_{t=0}^{\tau-1}$ , which provide a rigorous approach to classify and recognize actions. To obtain the final feature vector for a snippet, it is partitioned into several time-slices defined by linguistic intervals (Sadek et al., 2010). A Gaussian fuzzy membership function is used to describe each of

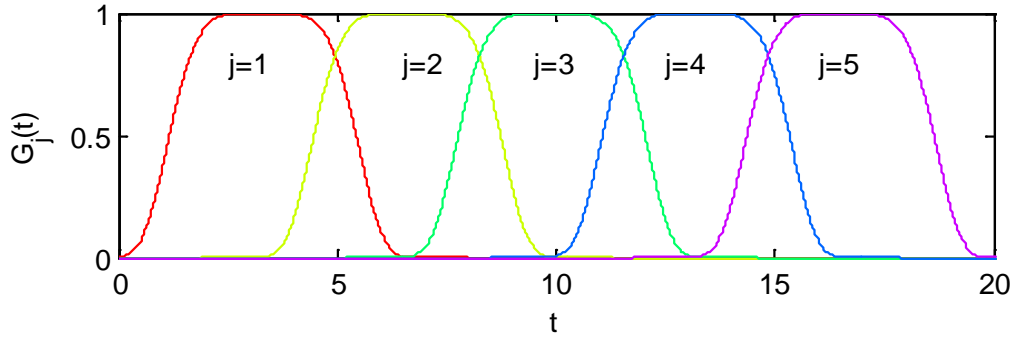


Figure 2: Fuzzy membership functions used to represent five temporal steps.

these intervals,

$$\mathcal{G}_j(t; \alpha, \beta, \gamma) = e^{-\left|\frac{t-\alpha}{\beta}\right|^\gamma} \quad (4.7)$$

where  $\alpha, \beta$  and  $\gamma$  are scalar parameters, i.e. the center, width, and fuzzification factor, respectively. Therefore, a feature vector for a time-slice can be created by calculating the weighted average feature vector of all frames within the time-slice. Formally, the feature vector for time-slice  $j$  is given by,

$$\mathcal{F}_j = \frac{1}{\Delta t} \sum_{t \in \text{slice}_j} \mathcal{G}_j(t) \mathbf{f}_t, \quad j = 1, 2, \dots, m \quad (4.8)$$

where  $\mathcal{G}_j(t)$  is the membership function representing the  $j$ -th time slice,  $\Delta t$  is the duration of the time slice, and  $m$  is the total number of time slices (see Figure 2). Thus, the full feature vector for an action snippet can straightforwardly be derived by concatenating all  $m$  feature vectors of time slices:

$$\mathcal{A} = \mathcal{F}_1 \oplus \mathcal{F}_2 \oplus \dots \oplus \mathcal{F}_m \quad (4.9)$$

where  $\oplus$  is the concatenation operator. From the above mentioned, it follows that the process of slicing action snippets into a finite number of temporal steps achieve the primary goal of effective feature dimensionality reduction and de-correlation by removing probable redundancy in the features set, while retaining the information essential for effective recognition.

#### 4.4 Action classification

In this work, we formulate the action recognition task as a multi-class 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. Naïve Bayesian (NB) classifier is used for action classification. The main advantage of the NB is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. In spite of their naive design and apparently over-simplified assumptions, NB classifiers have shown to work quite well in many complex real-world situations (P. and Pazzani, 1997). Formally speaking, given a final feature vector  $X$  extracted from a test action snippet, posteriori probabilities are calculated using training action snippets. This is accomplished using Bayes rule, as follows,

$$p(\omega_i|X) = \frac{p(X|\omega_i)p(\omega_i)}{p(X)} \quad (4.10)$$

where  $p(X) = \sum_{i=1}^K p(X|\omega_i)p(\omega_i)$ .  $p(\omega_i|X)$  is the posteriori probability of observing the class  $\omega_i$  given the feature vector  $X$ .  $p(\omega_i)$  is the priori probability of observing the class  $\omega_i$ ,  $p(X|\omega_i)$  is the conditional density, and  $K$  is the total number of classes. For this recognition task, it is assumed



Figure 3: Example frames from video sequences in our IKT action dataset.

that each action is uniquely described by the value of its a posteriori probability. Moreover, all priori probabilities are assumed to be equal, and thus find the density functions for each class. Hence, such  $K$  densities are found, corresponding to  $K$  different actions. Having obtained  $K$  values for all action classes, the most likely action is given by,

$$P = \max[p_1, p_2, \dots, p_K] \quad (4.11)$$

where  $P$  is the probability of the most likely class and  $p_1, p_2, \dots, p_K$  are the probabilities of  $K$  actions.

#### 4.5 Action localization

The analysis of the spatial location distribution of the features can efficiently contribute to a fast and accurate estimation of the 2D position of the centroid of these features based on the average of the coordinates of all feature points in motion ROI (Sadek et al., 2012a). More formally, the centroid of an action, at each time instant, is calculated according to the following expression:

$$\mu_x = \frac{1}{n} \sum_{i=1}^n x_i, \quad \mu_y = \frac{1}{n} \sum_{i=1}^n y_i \quad (4.12)$$

where  $(\mu_x, \mu_y)$  denote 2D coordinates of the centroid of the features. This centroid coincides with the estimated center of mass of the moving ROI (i.e., action actor here). In a similar vein, the dimensions of the moving object are estimated by

$$\sigma_x = 2\sqrt{3\eta_{xx}}, \quad \sigma_y = 2\sqrt{3\eta_{yy}} \quad (4.13)$$

where  $\eta_{xx}$  and  $\eta_{yy}$  are the central moments of the corresponding centroid.

### 5 Experimental Results

To evaluate the performance of our recognition approach, we decided to create our own realistic action dataset (i.e. so called IKT action dataset) which is going to be publicly available free of restrictions on use for action recognition research on the Web very soon. Analogous to benchmark KTH dataset (Schüldt et al., 2004), six action categories are contained in our IKT dataset; three “leg actions” (i.e., walking, jogging, and running) and three “arm actions” (i.e., boxing, hand-waving, and hand-clapping). Contrary to KTH dataset, the sequences in IKT dataset were taken over various non-homogeneous backgrounds at 30 fps frame rate. Within the sequences, actions are performed by nine subjects, each subject was asked to wear a different clothing item. This is expected to make recognizing actions slightly more challenging. Figure 3 shows example actions in IKT dataset. A series of experiments have been carried out to quantify the effect on recognition performance of altering the feature description parameter (i.e.,  $m$ ) in order to establish the optimum recognition rate.



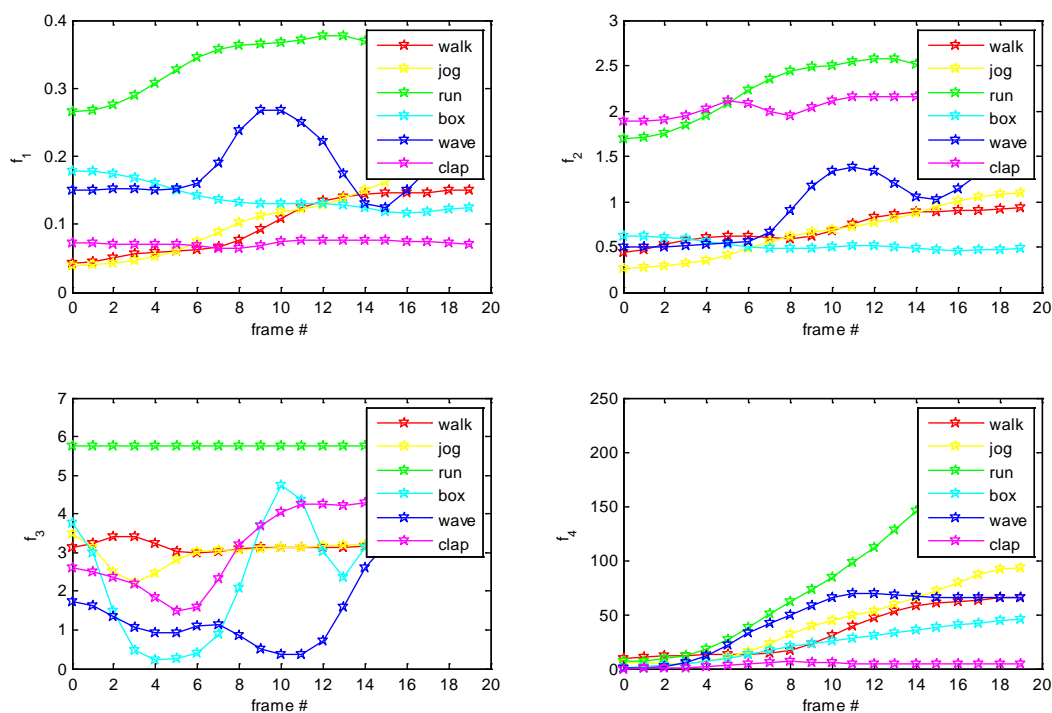


Figure 4: Plots of simple statistical features computed on the temporal motion templates of walking, jogging, running, boxing, waving, and clapping actions.

As there was no control over the video capturing process, the action sequences in the dataset exhibit some degree of variation in the actors, scale, pose, camera views, appearance inside the same action category, coupled with cluttered background and different illumination conditions. Considering that most previous research experiments were conducted in controlled or partially controlled environments (e.g., KTH and Weizmann datasets), we intuitively expect that the experimental results using this dataset will be more realistic. The test data consists of a total of 300 action snippets derived from the video sequences recorded in the dataset. These streams were saved in AVI format with a resolution of  $640 \times 480$ -pixel frame dimensions with 24-bit color depth at 30 fps frame rate. An additional total of 480 streams are used to train the NB classifier.

Figure 4 shows an example of visualization of the applied features extracted from different action categories. By inspecting the plots in this figure, one can observe that the features reflect the actual similarity/dissimilarity between different categories of actions. One more interesting observation is that the descriptor remains constant or slightly changes with time; this suggests that a relatively few number of time slices will suffice to construct such a descriptor.

With the eventual goal of developing a high performance recognition system, we investigate the recognition performance of the framework under the values of the feature description parameter  $m$  varying. To achieve this, we compute the feature descriptors a total of five times ( $m \in \{1, 2, 3, 4, 5\}$ ) for all samples in training set. To facilitate the visualization of the system's performance, the confusion matrices that tabulate the correct and incorrect classifications are calculated through majority voting. The performance of the system can be presented directly in the form of confusion tables. Instead, for the sake of clarity, we graphically represent these confusion tables through a series of 3D bar plots, presented in Figure 5. By inspecting all plots shown in the figure, it is explicitly observed that the feature representation parameter  $m$  is significant and directly affects the results of the recognition.



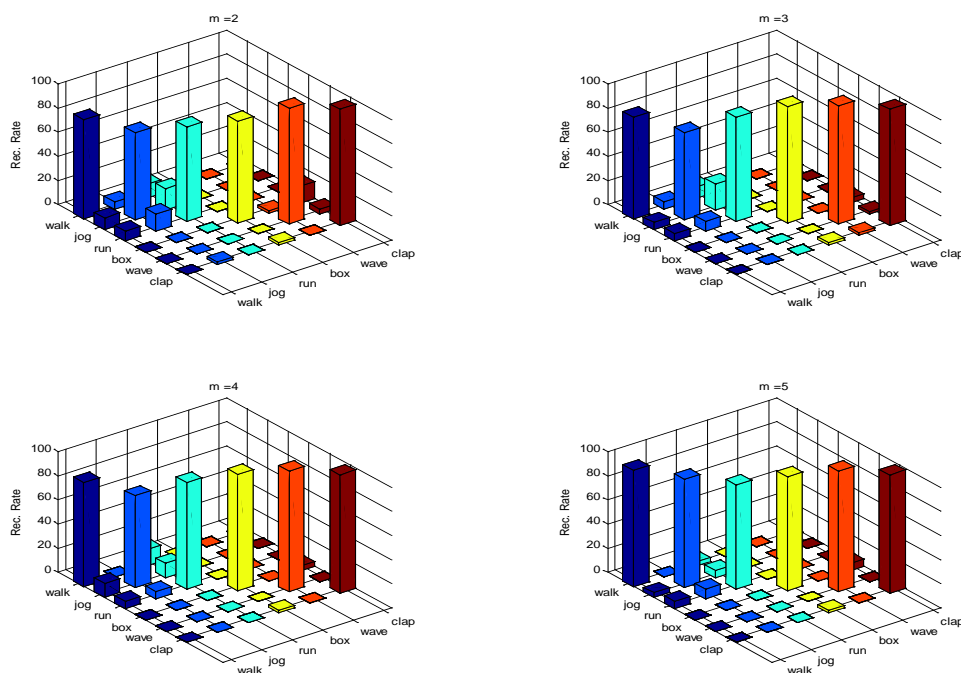


Figure 5: Sample 3D bar plots visualizing the confusion in action recognition results, each corresponding to different value of the feature parameter  $m$ .

Furthermore, the accuracy metric is used to gauge the holistic performance of the recognition scheme. The experiments demonstrate two points of particular interest. First, the feature parameter  $m$  is significant and directly affects recognition results. Secondly, in terms of recognition performance, the larger values of  $m$  provide a greater improvement in recognition rate. The best recognition accuracy achieved by the approach is 94.7% that can be regarded as "encouraging" and confirms the basic correctness of the approach, regarding the realistic environments. All the routines have been coded in Visual Studio 2008 and executed on a PC equipped with an Intel Core 2 processor operating at 2.8 GHz with 8 MB of cache and 4 GB of SDRAM. The final experiment was conducted with the purpose of localizing action objects as moving regions of interests (ROIs) identified by motion information. To achieve this goal, eqs. (4.12) and (4.13) are used to determine of the location of the action subject. In practice, the approach has proved to be more efficient for scenes with a relatively stable background, even with very high levels of noise. Some results of action localization are depicted in Figure 6.

## 6 Conclusion and Future work

This paper has presented an innovative approach for action recognition, in which a compact fuzzy descriptor is constructed using temporal templates and fuzzy temporal slicing. The simplicity and computational efficiency of the employed features allow the approach to be amenable for integration into real-time applications. Future work will involve further investigations on larger realistic datasets to discuss the substantive correctness, robustness, and large-scale feasibility of the approach.

## Acknowledgment

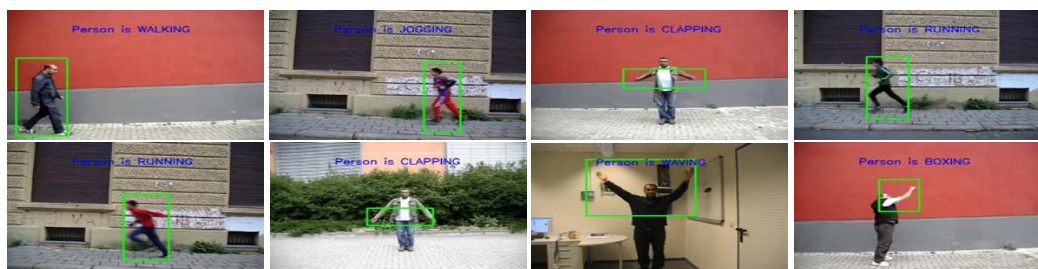


Figure 6: Some results of action localization and recognition in IKT dataset.

This research was supported by the Transregional Collaborative Research Center SFB/TRR 62 Companion Technology for Cognitive Technical Systems funded by the German Research Foundation (DFG). The financial support from the DFG is gratefully acknowledged.

## References

- Bobick, A. and Davis, J. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(3):257–267.
- Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *ICCV*, volume 2, pages 726–733.
- Flusser, J. and Suk, T. (1993). Pattern recognition by affine moment invariants. In *Pattern Recognition*, volume 26, pages 167–174.
- Ikizler, N. and Forsyth, D. (2007). Searching video for complex activities with finite state models. In *CVPR*.
- Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. In *ICCV*, pages 257–267.
- Laxton, B., Lim, J., and Kriegman, D. (2007). Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.
- Little, L. and Boyd, J. E. (1998). Recognizing people by their gait: The shape of motion. *Int. Journal of Computer Vision*, 1(2):1–32.
- Liu, J. and Shah, M. (2008). Learning human actions via information maximization. In *CVPR*.
- Lu, W.-L., Okuma, K., and Little, J. J. (2009). Tracking and recognizing actions of multiple hockey players using the boosted particle filter. *Image and Vision Computing*, 27(1):189–205.
- P., D. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–137.
- Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*.

- Sadek, S., Al-Hamadi, A., Michaelis, B., and Sayed, U. (2010). Human activity recognition: A scheme using multiple cues. In *Proceedings of the International Symposium on Visual Computing (ISVC'10)*, volume 1, pages 574–583.
- Sadek, S., Al-Hamadi, A., Michaelis, B., and Sayed, U. (2012a). Human action recognition via affine moment invariants. In *21st International Conference on Pattern Recognition (ICPR'12)*, pages 218–221.
- Sadek, S., Al-Hamadi, A., Michaelis, B., and Sayed, U. (2012b). An SVM approach for activity recognition based on chord-length-function shape features. In *IEEE ICIP'12*, pages 767–770.
- Schüldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *ICPR*, pages 32–36.
- Shechtman, E. and Irani, M. (2005). Space-time behavior based correlation. In *CVPR*, volume 1, pages 405–412.
- Wang, L., Qiao, Y., and Tang, X. (2015). Action recognition with trajectory-pooled deep-convolutional descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.