Toward Real-world Activity Recognition: An SVM Based System Using Fuzzy Directional Features

Samy Sadek, Ayoub Al-Hamadi, Bernd Michaelis
Otto-von-Guericke University Magdeburg
Department of Electrical Engineering and Information Technology
Universitätsplatz 2, 39016 Magdeburg
GERMANY
{*Samy.Bakheet,Ayoub.Al-Hamadi*}*@ovgu.de*

Abstract: Despite their attractive properties of invariance, robustness and reliability, fuzzy directional features are not hitherto paid the attention they deserve in the activity recognition literature. In this paper, we propose to adopt an innovative approach for activity recognition in real-world scenes, where a new fuzzy motion descriptor is developed to model activities as time series of fuzzy directional features. A set of one-vs.-all SVM classifiers is trained on these features for activity classification. When evaluated on our dataset (i.e., IESK action dataset) incorporating a large and diverse collection of realistic video data, the proposed approach yields encouraging results that compare very favorably with those reported in the literature, while maintaining real-time performance.

Key–Words: Human activity recognition, fuzzy directional features, one-vs.-all SVM, video interpretation

1 Introduction

Recognizing human activities in unconstrained settings is a longstanding and extremely challenging problem in computer vision and many of its related applications, due to a variety of challenging real-world conditions, including partial occlusion, substantial background clutter, drastic illumination variation, large intra-class variability within each class, extreme pose variation, and changes in scale, viewpoint, and appearance [1]. Specifically, in this work, we propose to focus on the recognition of human activities in real-world scenarios which is an important but challenging problem with prosperous applicability into human-computer interactions and security industry. Real-world datasets for the evaluation of human action recognition systems generally consist of a large collection of real-world video streams (or video clips) about the actions of interest. Each video stream includes an individual (i.e. action subject) performing a single action or a series of successive actions. All videos belonging to the same action category can be annotated with a categorical label describing the type of action performed within them. It is clear that developing good algorithms for solving the problem of the recognition of human activities in real-world scenes would yield huge potential for a large number of potential real-life applications, e.g., human-computer interaction, video surveillance, gesture recognition, and robot learning and control, etc. 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. 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. The automatic recognition of human actions is still an underdeveloped area due to the lack of a general purpose model and most approaches proposed in the literature remain limited in their ability. For this, much research still needs to be undertaken to address the ongoing challenges. The remainder of this paper is organized as follows. Section 2 reviews related work in the literature. The proposed framework for activity recognition is presented and explained in details in Section 3, whereas Section 4 outlines the details of our evaluation procedure. Finally, in Section 5, we summarize our results and draw conclusions.

2 Related literature

Over the course of the last couple of decades or so, a great deal of work has been done (and still being done) on the recognition of human activities from both still images and video sequences. Despite these years of work, the problem is still open and provides a big challenge to the researchers and more rigorous research is needed to come around it. Human action can generally be recognized using various visual cues such as motion [2–4] and shape [5, 6]. Scanning the literature, one notices that a significant body of work in action recognition focuses on using spatial-temporal keypoints and local feature descriptors [7–9]. 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 [3], periodic motions are detected and classified to recognize actions. In [2] 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 [10], Bobick and Davis use temporal templates, including motion-energy images and motion-history images to recognize human movement. In [11] the authors detect the similarity between video segments using a space-time correlation model. While in [12], Rodriguez et al. present a template-based approach using a Maximum Average Correlation Height (MACH) filter to capture intra-class variabilities. Jhuang et al. [13] perform actions recognition by building a neurobiological model using spatio-temporal gradient. In [14], 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 [15–18]. Finally, there is also an attractive area of research that concentrates on using generative topic models for visual recognition based on the so-called Bag-of-Words (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. Topic models are built and then applied to the BoW representation. Three of the most popularly used topic models are Latent Dirichlet Allocation (LDA) [19], Correlated Topic Models (CTM) [20] and probabilistic Latent Semantic Analysis (pLSA) [21].

3 Proposed Methodology

In this section, we present a new approach for action recognition in real-world video sequences, based on a modified fuzzy version of HOF (Histogram of Optical Flow), so-called fuzzy histogram of optical flow as a new motion descriptor to model action in a real-world scene as a time-series of fuzzy directional features. A set of one-vs.-all SVM classifiers are trained on these features for the action classification. We evaluate the approach on IESK dataset which incorporates a collection of real-world video data.

3.1 Motion estimation

To detect moving objects (i.e., action subjects), we use an algorithm that works based on the same principles as the two-frame motion estimation algorithm in [22]. The key idea of the algorithm is to approximate a pixel neighborhood in a frame by a quadratic polynomial:

$$f(\mathbf{x}) \sim p(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c \tag{1}$$

where $\mathbf{A}, \mathbf{b}$, and c are the expansion coefficients that are determined using a Gaussian-weighted least-squares fitting of the signal f by the polynomial p. Hence, the new frame can be derived from the last one by a global translation $\mathbf{d}$:

$$\begin{aligned}\tilde{f}(\mathbf{x}) &\sim p(\mathbf{x}-\mathbf{d}) \\ &= (\mathbf{x}-\mathbf{d})^\top \mathbf{A}(\mathbf{x}-\mathbf{d}) + \mathbf{b}^\top(\mathbf{x}-\mathbf{d}) + c \\ &= \mathbf{x}^\top \tilde{\mathbf{A}}\mathbf{x} + \tilde{\mathbf{b}}^\top \mathbf{x} + \tilde{c} \end{aligned} \tag{2}$$

It is easy to see that these two sets of expansion coefficients are related by

$$\begin{aligned}\tilde{\mathbf{A}} &= \mathbf{A}, \\ \tilde{\mathbf{b}} &= \mathbf{b} - 2\mathbf{A}\mathbf{d}, \\ \tilde{c} &= c + \mathbf{d}^\top \mathbf{A}\mathbf{d} - \mathbf{b}^\top \mathbf{d}. \end{aligned} \tag{3}$$

Looking at Eq. (3), one realizes that a solution for the translation $\mathbf{d}$ exists only if

$$\mathbf{d} = \frac{1}{2}\mathbf{A}^{-1}(\tilde{\mathbf{b}} - \mathbf{b}) \tag{4}$$

For practical considerations, the global polynomial in (4) are replaced with local polynomial approximations. Thus, giving two sets of expansion coefficients $\{\mathbf{A}_1(\mathbf{x}), \mathbf{b}_1(\mathbf{x}), c_1(\mathbf{x})\}$ and $\{\mathbf{A}_2(\mathbf{x}), \mathbf{b}_2(\mathbf{x}), c_2(\mathbf{x})\}$ for the first and second image frames respectively, it is possible to do a polynomial expansion of both frames. Ideally, this yields $\mathbf{A}_1 = \mathbf{A}_2$, however, in practice one is forced to settle for the approximation:

$$\mathbf{A}(\mathbf{x}) = \frac{\mathbf{A}_1(\mathbf{x}) + \mathbf{A}_2(\mathbf{x})}{2} \tag{5}$$

and further the following assumption

$$\Delta\mathbf{b}(\mathbf{x}) = -\frac{1}{2}(\mathbf{b}_2(\mathbf{x}) + \mathbf{b}_1(\mathbf{x})) \tag{6}$$

is made, which leads to the primary constraint

$$\mathbf{A}(\mathbf{x})\mathbf{d}(\mathbf{x}) = \Delta\mathbf{b}(\mathbf{x}) \tag{7}$$

where $\mathbf{d}(\mathbf{x})$ implies that the global displacement in (2) is replaced with a spatially varying displacement field. Under the assumption that the displacement field is only slowly varying, information over a neighborhood Ω of each pixel can be integrated. Consequently, $\mathbf{d}(\mathbf{x})$ satisfying (7) and minimizing

$$\sum_{\mathbf{\Delta x}\in\Omega} w(\mathbf{\Delta x})\|\mathbf{A}(\mathbf{x}+\mathbf{\Delta x})\mathbf{d}(\mathbf{x})-\Delta\mathbf{b}(\mathbf{x}+\mathbf{\Delta x})\|^2 \tag{8}$$

can be found, where $w(\mathbf{\Delta x})$ is a Gaussian weight function. Therefore, the minimum value is given by

$$e(\mathbf{x}) = \left(\sum w\mathbf{\Delta b}^\top\mathbf{\Delta b}\right) - \mathbf{d}(\mathbf{x})^\top \sum w\mathbf{\Delta A}^\top\mathbf{\Delta b},$$

which is obtained for

$$\mathbf{d}(\mathbf{x}) = \left(\sum w\mathbf{\Delta A}^\top\mathbf{\Delta A}\right)^{-1} \sum w\mathbf{A}^\top\mathbf{\Delta b} \tag{9}$$

It was shown, in [22] that in many cases it might be advantageous to introduce a certainty weight $\mathbf{c}(\mathbf{x}+\mathbf{\Delta x})$ to Eq. (8) that can be most conveniently achieved by scaling $\mathbf{A}$ and $\mathbf{\Delta b}$. To detect moving objects, particularly people (i.e., action subjects), the displacement field should be parameterized according to some motion model (e.g., affine motion model or eight-parameter model). For the eight-parameter model in 2D, the motion field can be expressed as,

$$\mathbf{d} = \mathbf{Sp} \tag{10}$$

where,

$$\begin{aligned}\mathbf{S} &= \begin{pmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 \end{pmatrix}, \\ \mathbf{p} &= \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \end{pmatrix}^\top \end{aligned} \tag{11}$$

Substituting from (10) into (8) yields the weighted least squares problem:

$$\sum_i w_i\|\mathbf{A}_i\mathbf{S}_i - \mathbf{\Delta b}_i\|^2 \tag{12}$$

which in turn has the solution

$$\mathbf{p} = \left(\sum_i w_i\mathbf{S}_i^\top\mathbf{A}_i^\top\mathbf{A}_i\mathbf{S}_i\right)^{-1} \sum_i w_i\mathbf{S}_i^\top\mathbf{A}_i^\top\mathbf{\Delta b}_i \tag{13}$$

The actual solution involves the accumulation of the coefficients of the 8×8 system of equations (13) over all points and then solving for the parameters. To improve the chances for a better displacement estimate in the algorithm, it is crucial to exploit some a priori knowledge about the displacement field that allow comparing the polynomial at $\mathbf{x}$ in the first signal to the polynomial at $\mathbf{x}+\tilde{\mathbf{d}}(\mathbf{x})$ in the second signal, where $\tilde{\mathbf{d}}(\mathbf{x})$ is the a priori displacement field. In this case, $\mathbf{A}(\mathbf{x})$ and $\Delta\mathbf{b}(\mathbf{x})$ introduced in Eq. (5) and Eq. (6) are substituted by

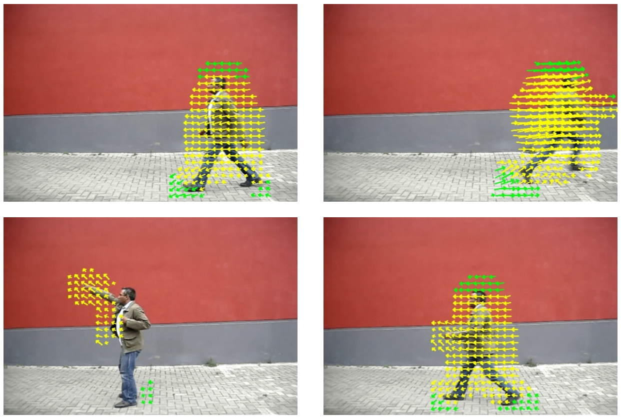


Figure 1: Sample pruning results for a setup with $\lambda = 0.25\ell$; the vectors labeled in yellow are accepted as valid flow components, while the vectors labeled in green are considered as noisy flow components and thus filtered out.

$$\begin{aligned}\mathbf{A}(\mathbf{x}) &= \frac{\mathbf{A}_1(\mathbf{x})+\mathbf{A}_2(\tilde{\mathbf{x}})}{2}, \\ \Delta\mathbf{b}(\mathbf{x}) &= -\frac{1}{2}(\mathbf{b}_2(\tilde{\mathbf{x}})+\mathbf{b}_1(\mathbf{x}))+\mathbf{A}(\mathbf{x})\tilde{\mathbf{d}}(\mathbf{x})\end{aligned}$$

where $\tilde{\mathbf{x}} = \mathbf{x}+\tilde{\mathbf{d}}(\mathbf{x})$.

3.2 Flow Pattern pruning

It has to be admitted that despite over two decades of intensive research, most existing methods for the extraction of optical flow still lack robustness, and optical flow estimates are relatively inaccurate, particularly with respect to flow magnitude. This might be attributed to the large residual error in solving the equations for optical flow. Therefore, pruning of computed flow values appears to be a clue to accurate flow fields which in turn allows for better motion estimation. To tackle this problem, we introduce a particular kind of filter that straightens up noisy vectors in the flow field, while maintaining significant ones.

In our work, we perform this type of pruning stepwise. In other words, it involves two passes, each based on the magnitude (Euclidean length) of optical flow vectors to separate relevant from irrelevant flow vectors. In the first pass, we attempt to remove all flow vectors whose magnitudes are either relatively very small or very large. For this purpose, two predefined thresholds (i.e., minimum and maximum thresholds) are used that control the filtering of flow vectors in this step. Formally speaking, given two thresholds ρ_1 and ρ_2, a flow vector $\vec{v} = [x, y]^\top$ is only accepted as valid if it satisfies the validity constraint: $\rho_1 < \|\vec{v}\| < \rho_2$, where $\|\cdot\|$ denotes the magnitude of the flow vector with respect to the Euclidean metric; otherwise it is assumed to be a noisy flow component and thus removed. In our experiments, when ρ_1 and ρ_2 are given 5 and 20 respectively, satisfactory results can be achieved. We go then with a second pass of our pruning based on the Euclidean distance between the centroid of flow field and the flow points. Therefore, in this pass a vector $\vec{v}$ is treated as a valid flow component if the Euclidean distance between the center of flow and the vector being analyzed does not exceed a specific threshold λ. Formally, this is expressed as:

$$\|\vec{v} - \vec{c}\| < \lambda \tag{14}$$

where $\vec{c}$ is the motion region's centroid. In experiments, we found that setting the value of λ to one-fourth of the average of the image's width and height (i.e., $(w + h)/8$) yields a good pruning performance (see Figure 1 for visual examples).

3.3 Directional feature extraction

In the literature, several existing theoretical approaches to action recognition tend to put much more emphasis on providing practical methods which are consistently applicable only to various joint angles acquired from motion capture data. However, when applying these approaches to video data, we are regularly faced with the complex problem of segmenting and tracking of human joints. This problem is considerably more challenging and error-prone, particularly in dynamically complex environments where the tracking objects frequently undergo large changes in pose, scale, and lighting conditions.

Motivated by the potential benefits in performance of histogram of features (e.g. HOG [23]) for object recognition, in this work, we propose to compute a new motion-related descriptor based on optical flow analysis. However, most optical flow computations turn out to be most sensitive to background noise, and changes in scale and/or directionality of motion. Furthermore, the number of moving pixels is subject to change with time. Due to these restrictions, raw values of optical flow would likely be less suitable or unsuitable as features for motion analysis. In order to overcome these difficulties, we can here use the characteristics of distribution of optical flow as features to describe motion. As a matter of fact, one can see that the motion activity of an individual moving in a scene with a static background can be characterized fully by its own self-induced optical flow profile. In Figure 2, sample flow patterns for a video sequence showing a person performing several actions are shown.

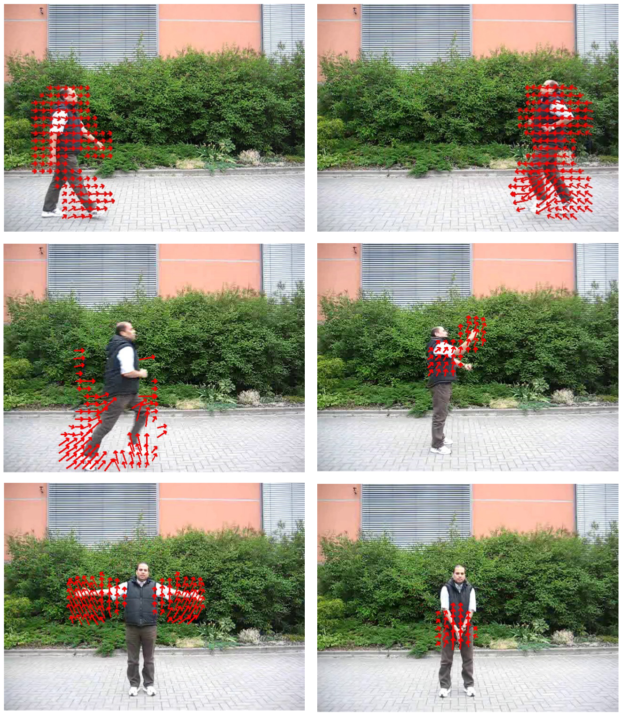


Figure 2: Flow estimation results for a video sequence showing a single person performing various actions, i.e. walking, jogging, boxing, waving, and clapping from left to right and top to bottom, respectively.

The main thrust of our work is to develop a new descriptor based on improved optical flow measurements over a spatiotemporal volume centered on a human figure to represent actions, and SVM classifiers are then used to classify these descriptors. To generate a robust and discriminative motion descriptor invariant to pose variation and directionality of motion, two aspects should be kept in mind, one referring to the dependency of the observed flow profile on the scale of motion activity, the other relating to the dependency of the orientation of optical flow on the directionality of motion. Moving from these considerations and requirements, we propose here the FHOF (Fuzzy Histogram of Optical Flow). A formal definition and implementation scheme of this new descriptor are as follows. Given an estimate for optical flow field at each

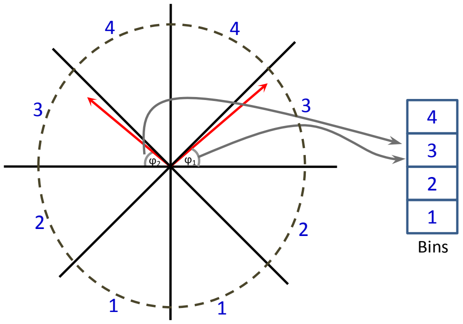


Figure 3: An example for orientation histogram with four bins ($K = 4$).

frame of the action sequence, the magnitude and the orientation of each flow vector $\vec{v} = [x, y]^\top$ are specially defined as follows,

$$\begin{aligned}\rho &= \sqrt{x^2 + y^2} \\ \varphi &= \mathrm{atan2}(y, |x|) \end{aligned} \qquad (15)$$

where $| \cdot |$ denotes the ordinary absolute value, and $\frac{-\pi}{2} < \varphi \leq \frac{\pi}{2}$ that gives the smallest angle between the x-axis and $\vec{v}$ axis, as shown in Figure 3. It should be noted that the orientation angle φ in (15) has been defined so as to allow our histogram representation to be independent of the directionality of movement. A histogram can be derived at each frame by binning the flow vectors into a fixed number of bins based on their primary angles and their magnitudes. More formally, the directional histogram is created where each flow vector $\vec{v}$ with direction φ in the range:

$$-\frac{\pi}{2} + \pi\frac{k-1}{K} \leq \varphi < -\frac{\pi}{2} + \pi\frac{k}{K} \qquad (16)$$

gives a contribution proportional to ρ to its corresponding bin k, $1 \leq k \leq K$ where K is the number of bins. As seen in Figure 3, the resulting histogram representation is invariant to direction of motion. To achieve invariance to scale changes, the histogram is normalized by the overall magnitude of flow vectors, so that the bins integrate to unity. Moreover, as flow vectors contribute to the histogram proportionally to their magnitudes, the resulting descriptor would be more robust to noisy flow measurements. A visualization of the descriptor for the applied features is given in Figure 4. From a close inspective look at the plots in the figure, one can see that there is a remarkable similarity in feature structure (leading to similar color values in the Figure) among sequences of walk, jog, and run actions, and between sequences of wave and clap actions. Intuitively, this is due to the high closeness of similar types of actions.

3.4 Fuzzy feature selection

In this section, we describe our method for feature selection based on temporally adaptive decomposition of action sequences into a finite number of time slices in a fuzzy way, which is targeted at the removal of irrelevance and redundancy in the features set, so that not only does the reduced set of features speed up the action classification process by removing class irrelevant features, but it also provides at least the same quality of action classification as the original one. Eventually, this enables the proposed approach to achieve better feature reduction ratios without losses in recognition accuracy. As discussed in the previous section, a normalized histogram based on the HOG features can be derived at time instant t:

$$\mathbf{h}_t = (h_{t;1}, h_{t;2}, \ldots, h_{t;K})^\top \qquad (17)$$

where K (the number of histogram bins) is a parameter of choice, which has a direct influence on the performance of the recognition system. Since the features in (17) can be computed at a time instant of a given sequence (i.e., action snippet), the action snippet can be represented as a time series of these features: $A = \{\mathbf{h_t}\}_{t=0}^{\tau-1}$ which provides us a rigorous approach to classify and recognize actions.

To obtain the final feature vector for each action snippet, we temporally partition each action snippet into several time-slices defined by linguistic intervals [24]. A Gaussian fuzzy membership function is used to describe each of these intervals. The general forms of these functions is given as follows

$$\mathcal{G}_j(t; \alpha, \beta, \gamma) = e^{-\left|\frac{t-\alpha}{\beta}\right|^\gamma} \qquad (18)$$

where $\alpha, \beta,$ andγ are three scalar parameters of the fuzzy function; i.e., the center, width, and the fuzzification factor which is a weighting exponent on each fuzzy membership, respectively. Therefore, a feature vector for a time-slice can be generated by calculating the weighted average feature vector of all frames within the time-slice. More formally, the directional feature vector for time-slice j is given by,

$$\mathcal{H}_j = \frac{1}{\Delta t} \sum_{t \in slice_j} \mathcal{G}_j(t)\mathbf{h}_t, \; j = 1, 2, \ldots, m \qquad (19)$$

where $\mathcal{G}_j(t)$ is the Gaussian membership function representing the j-th time slice, Δt is the duration of the time slice in frames, and m is the total number of time slices into which the action snippet is divided. Accordingly, the full feature vector for an action snippet can be straightforwardly derived by concatenating all m feature vectors of its time slices as follows,

$$\mathcal{A} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \cdots \oplus \mathcal{H}_m \qquad (20)$$

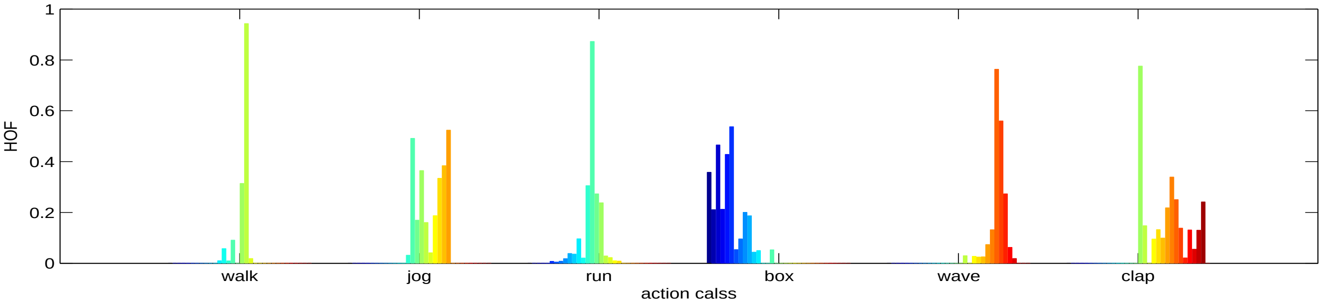


Figure 4: Visualization of the proposed descriptor (with $K = 32$) for HOF features extracted from sample sequences of walk, jog, run, box, wave, and clap actions.

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 would 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 of actions. For this purpose, each action sequence is treated as a time series composed of low-dimensional feature vectors corresponding to decomposition of the sequence into several time slices. More specifically, we keep only m multidimensional feature vectors corresponding to the m time slices, instead of taking all the feature vectors of all the frames in the video sequence. These m vectors form the feature space for action representation and classification.

It bears mentioning that m is a parameter of choice, where $m \ll n, n$ is the number of frames in the action sequence. To investigate whether and how the overall recognition results are affected by different values for m, in our experiments, different values of the parameter m were tried, each lies in the range of 1 to 5. The value that generates the highest average recognition accuracy over all runs would be selected. As a final note here it should also to be mentioned that the directional features are efficiently computed using fuzzy histograms that enables real-time implementation of the proposed action recognition method.

3.5 SVM based action classification

In this section, our goal is to classify actions according to the fuzzy descriptors mentioned previously. Human action recognition can be modeled as a multidimensional classification problem having one class for each action, and the goal is to assign a class label to a given action. For this purpose, we use one-vs.-rest SVMs (Support Vector Machines) with RBF (Radial Basis Function) kernels. For SVMs, the one-vs.-rest approach is widely adopted for handling the multi-class problem by constructing the decision rule based on multiple binary classification tasks.

Generally speaking, there are various supervised learning algorithms by which an action recognizer can be trained to recognize patterns of motion over time. In this work, we propose to employ SVMs in our framework due to their outstanding generalization capability and reputation of a highly accurate paradigm. SVMs [25] 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 5). 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, let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|\, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}\}$ be a training set, Coretes and Vapnik [25] have argued that this problem is best approached by allowing some examples to violate the margin constraints . These potential violations can be formulated using some positive slack variables ξ_i and a penalty parameter $C \geq 0$ that penalize the margin violations. Thus the optimal separating hyperplane is determined by solving the following QP problem:

$$\min_{\boldsymbol{\beta},\beta_0} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_i \xi_i \tag{21}$$

$$\text{subject to}$$
$$(y_i(\langle \mathbf{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \geq 1 - \xi_i \quad \forall i) \wedge (\xi_i \geq 0 \quad \forall i).$$

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 introduced to allow the margin to increase and to not force the hyperplane

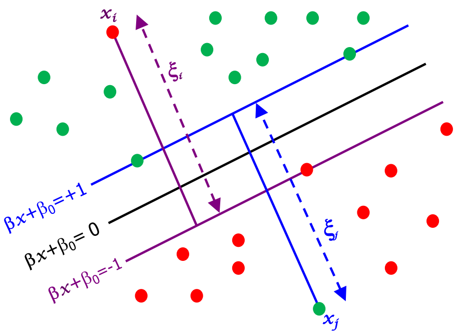


Figure 5: Generalized optimal separating hyperplane.

to pass through the origin that restricts the solution. For computational purposes, it is more convenient to solve SVM in dual space. To do this, we form the Lagrangian and then optimize over the Lagrange multiplier $\boldsymbol{\alpha}$. The resulting decision function has a 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 hyperplane.

For this approach, several classes of actions are defined and hence several one-vs.-all SVM classifiers are trained on the fuzzy directional features extracted from the action sequences in the training dataset. The feature vectors of the training set are fed into SVM classifiers in order to learn the differences among the features of each action class. In this work, we used one of the most popular and successful kernels, the RBF (or exponential) kernel, defined as

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp(\|\mathbf{x} - \mathbf{y}\|^2/(2\sigma^2)) \tag{22}$$

where σ is the kernel width which can be regarded as a tuning parameter. It is noteworthy to mention here that the SVMs with RBF have evolved as a flexible and powerful tool which is potentially able to create models that handle non-linearly separable data by mapping original features of the training data to a higher dimensional feature space to enable linear separation for classification. In this higher dimensional space, linear functions (or separators) can be constructed, which is potentially able to produce non-linear boundaries (see Figure 6) when mapped back to the original input space. Another important point to underscore here is that, for RBF kernel, there is a set of parameters (e.g, c and γ) for which several tests were carried out in order to establish their optimum values.

4 Experiments and discussion

In this section, we commence our discussion with a description of the action dataset on which the experiments are conducted. Thereafter, in the forthcoming sections, we present detailed descriptions of how the experiments were carried out and what their results show. To evaluate the performance of the proposed approach for action recognition in real world scenarios, we decided to create our own realistic action recognition dataset (hereinafter called as IESK 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 the KTH [26] action dataset, a total of six action categories are contained in the IESK action dataset; three "leg actions" (i.e., walking, jogging, and running) and three "arm actions" (i.e., boxing, hand-waving, and hand-clapping). The video sequences were typically acquired by a Canon IXUS 65 digital camera and stored in a resolution of 640 × 480 pixels represented in 256 grayscale levels. We believe that this resolution will likely be sufficient to reduce the high impact of the camera artifacts on the recognition results, since the data are internally stored in a lossy MPEG-format by the camera. Contrary to the KTH dataset, the sequences in IESK 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. Each action sequence was then segmented into shorter video clips of about 53sec duration which we term 'action snippets'. Figure 7 shows example frames from action sequences of different categories in the IESK dataset.

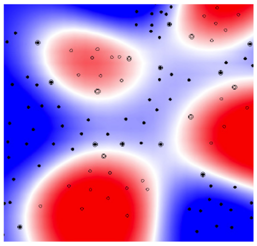


Figure 6: An example for nonlinear RBF kernel.

The reported results here are based on our feature extraction technique described in detail in Section 3.3 and 3.4 (i.e., fuzzy HOF-based features) and obtained with the IESK action recognition dataset that we created for the purpose of recognizing human actions in realistic scenarios. In this study, first of all, the experiments have been conducted to gauge the potential recognition capabilities of the proposed recognition system. This section shows also the results of a series of experiments performed to quantify the effect on recognition performance of altering the feature de-

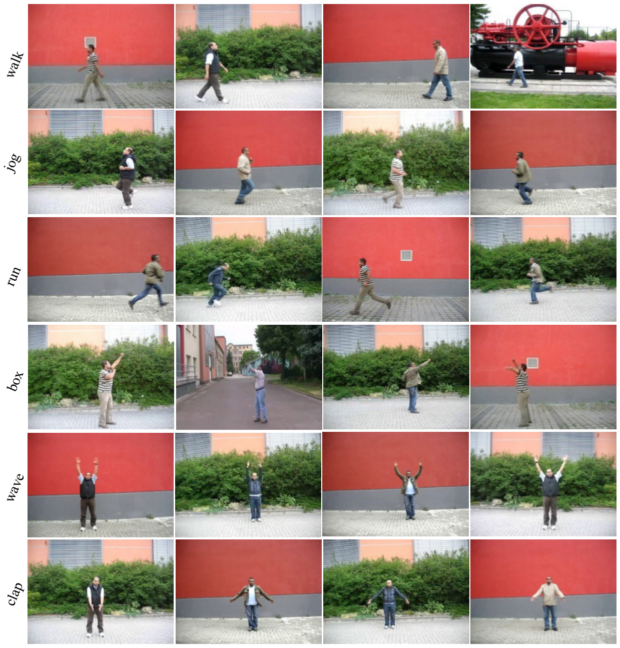


Figure 7: Sample frames from the action sequences in IESK action dataset.

scription parameters (i.e., K and m) in order to establish the optimum recognition rate.

As there was no control over the video capturing process, the action sequences in the dataset that we used in these experiments 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. As mentioned before, this dataset contains a total of six categories, namely walking, jogging, running, boxing, hand waving and hand clapping, performed several times by nine subjects. 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 × 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 six-action SVM model.

A series of experiments with different feature description parameters (K and m) was run to assess the effectiveness of the proposed technique for action recognition in realistic settings. We extracted about 360 directional features (for the case $K = 18$) from each action video, and then applied our fuzzy approach for feature selection described in Section 3.4 to reduce the dimension of the fuzzy feature descriptor to 90. Figure 8 shows an example of visualization of the proposed fuzzy descriptor for the directional features extracted from different action categories. By inspecting the figure, one can observe that the descriptor reflects the actual similarity/dissimilarity between different categories of actions at each temporal step. Thus, to quantify the degree of similarity or dissimilarity between two actions, a measure of similarity can be reliably computed based on a distance (e.g. Euclidean distance) between these descriptors. 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 action recognition system, we investigate the recognition performance of the proposed recognition framework under the values of the feature description pa-

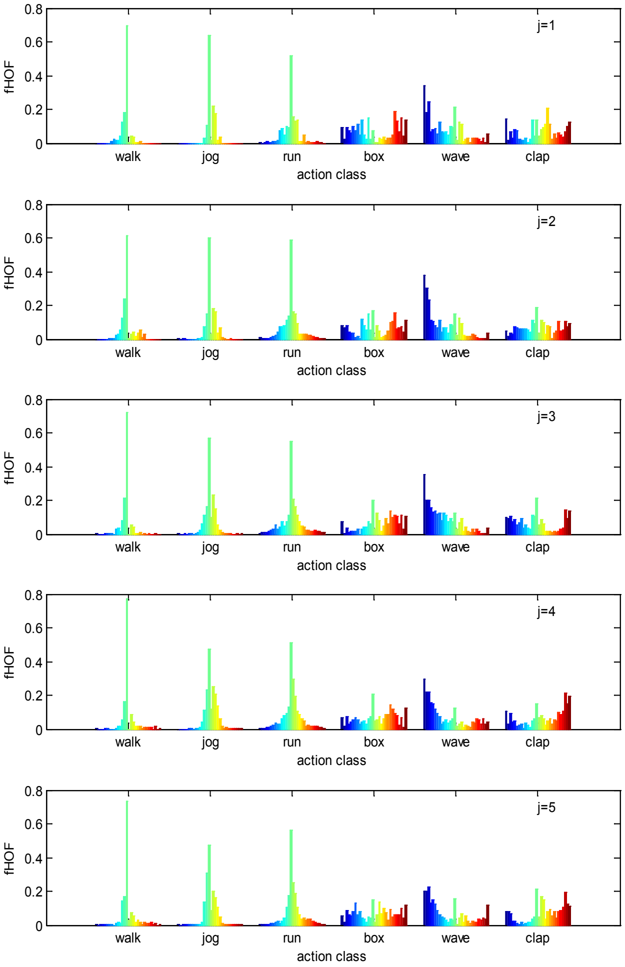


Figure 8: An example of visualization of the proposed descriptor for directional features extracted from different action categories at five temporal steps $m = 5$.

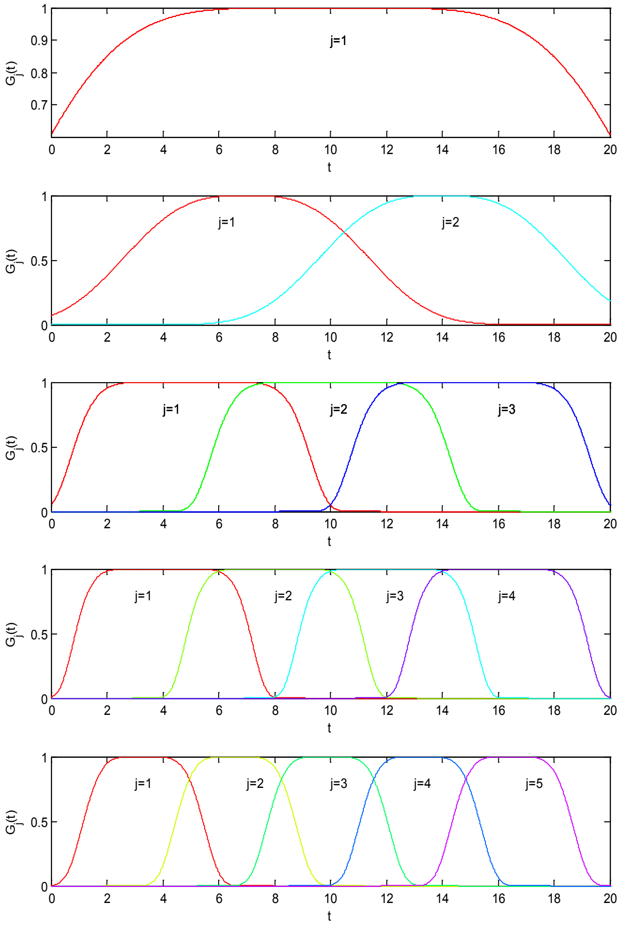


Figure 9: Fuzzy Gaussian membership functions used to represent temporal steps.

rameters (K and m) varying. Towards this goal, we compute such descriptors a total of 20 times for all samples in training set (i.e., the number of all possible combinations of values of the parameters m and K, where $m \in \{1, 2, 3, 4, 5\}$ and $K \in \{4, 8, 12, 18\}$). Therefore, m fuzzy membership functions should be defined to represent different time slices of a given action sequence, as shown in Figure 9. Note that for the sake of visualization, each fuzzy membership function in the Figure is plotted in a unique color.

In order to evaluate qualitatively and quantitatively the system's performance, we performed the previous experiments for all possible combinations of values of the feature parameters. 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 10. In this figure, we see a series of 3D plots visualizing the confusion in recognition results for each action category, each corresponding to a combination of feature representation parameters. By inspecting all plots shown in the figure, it is explicitly observed, as expected, that the feature representation parameters K and m are both significant and directly affect the results of the recognition.

Furthermore, the overall accuracy (or correct recognition rate) metric is employed to gauge the holistic performance of the proposed recognition scheme. The dependency of the overall recognition rate on the feature parameters has a shape similar to shown in Figure 11. Having a closer look at the figure, one can see that in terms of recognition rate, the larger values of both parameters provide the greatest improvement in performance, and generally are the most important. In other words, the larger the values of feature parameters are, the better the holistic performance is. For the sake of brevity, as a final remark in this section, we only mention that in our computational experiments, all the routines considered in this study were 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.

Action Localization:

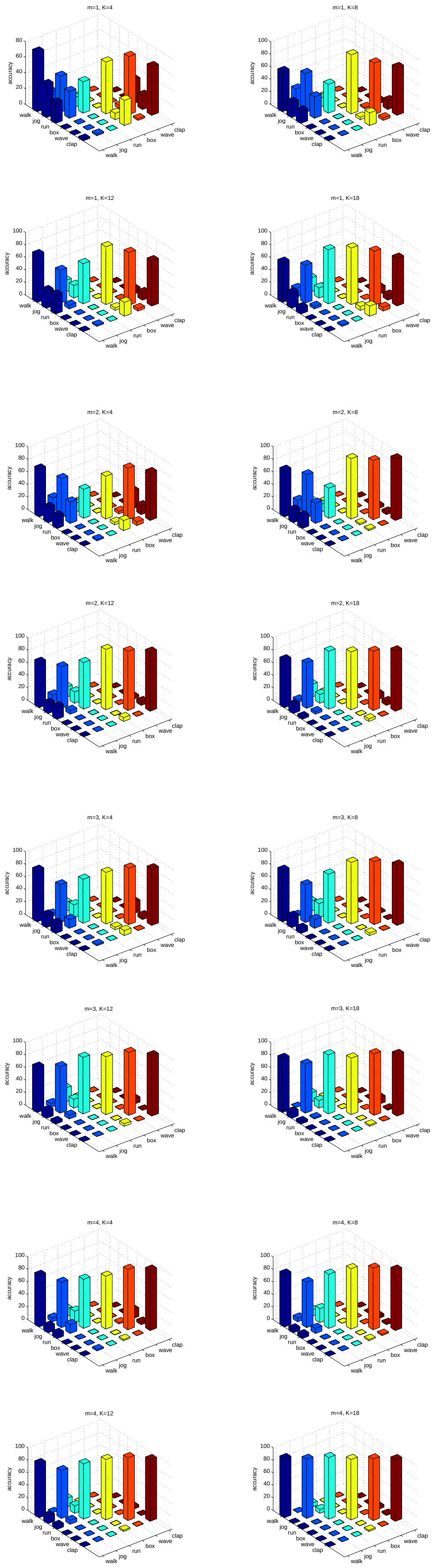
In this subsection, we describe the results of a final simple experiment conducted with the purpose of localizing the moving objects as motion regions of interest (ROI) identified by motion information. The analysis of the spatial location distribution of the flow features generated by our proposed fuzzy framework 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. More formally, the centroid of an action, at each time instant, is calculated according to the following expression:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{23}$$

where (μ_x, μ_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

$$\sigma_x = 2\sqrt{3\eta_{xx}}, \quad \sigma_y = 2\sqrt{3\eta_{yy}} \tag{24}$$

where η_{xx} and η_{yy} are the central moments of the corresponding centroid. In practice, this approach has



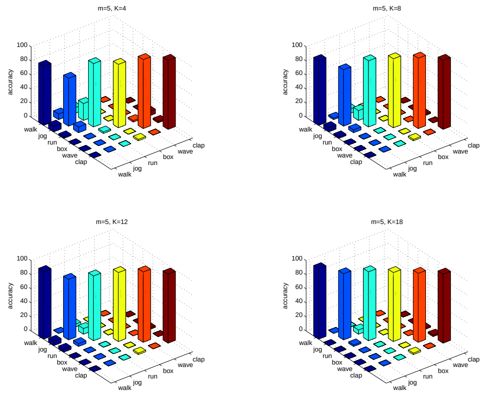


Figure 10: 3D bar plots visualizing the confusion in the action recognition results, each corresponding to different values of the feature parameters K and m.

proved to be significantly more efficient for scenes with a relatively stable background, even with very high levels of noise. In Figure 12, some results of action localization are depicted following this approach.

5 Summary and Conclusion

In this paper, we have presented an innovative approach for action recognition. On the basis of the proposed approach towards action recognition in realistic scenarios, a new fuzzy framework for representing and recognizing human actions in real-world video sequences has been presented. In this work, a compact and computationally-efficient descriptor; the fuzzy motion descriptor is constructed based on directional features of optical flow and fuzzy temporal slicing. The one-vs.-rest SVM classifiers have been trained automatically in the feature space for activity classification. The simplicity and computational efficiency of the employed features allow this approach to be more amenable for real-time implementation. It is noteworthy to point out here that the presented experiments conducted so far have demonstrated two points of considerable interest. First, the feature representation parameters K and m are both significant and directly affect the recognition results. Secondly, in terms of holistic performance, the larger values of both parameters provide the greatest improvement in overall recognition rate, and generally are the most important. In other words, the larger the values of the feature parameters are, the better the overall recognition performance is. Finally, for the sake of brevity here, we only affirm that the best overall recognition accuracy (corresponding to $K = 18$ and $m = 5$)

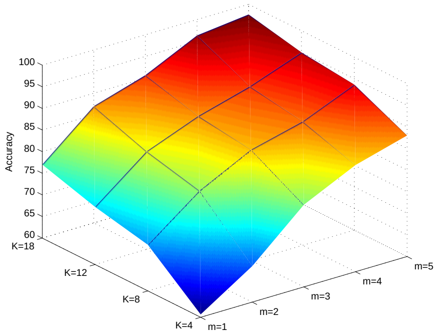


Figure 11: Overall action recognition performance of the proposed framework as a two-dimensional function of the feature parameters K and m.

achieved by the proposed approach is 96.3 % which can be regarded as ”encouraging“, and confirm the basic correctness of the approach, considering the realistic working environments. However, some further investigations on larger realistic datasets may be necessary to discuss the substantive correctness, robustness, and large-scale feasibility of the approach.

References:

[1] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, “An SVM approach for activity recognition based on chord-length-function shape features,” in *IEEE International Conference on Image Processing (ICIP’12)*, Florida, U.S.A., October 2012, pp. 767–770.

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

[3] R. Cutler and L. S. Davis, “Robust real-time periodic motion detection, analysis, and applications,” *IEEE Transactions on PAMI*, vol. 22, no. 8, pp. 781–796, August 2000.

[4] A. Efros, A. Berg, G. Mori, , and J. Malik, “Recognizing action at a distance,” in *IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2003, pp. 726–733.

[5] C. Thuran and V. Hlaváč, “Pose primitive based human action recognition in videos or still images,” in *CVPR*, 2008.

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

[7] I. Laptev and P. Pérez, “Retrieving actions in movies,” in *ICCV*, 2007.

[8] P. Dóllar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, 2005, pp. 65–72.

[9] J. Liu and M. Shah, “Learning human actions via information maximization,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[10] A. Bobick and J. 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.

[11] E. Shechtman and M. Irani, “Space-time behavior based correlation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 405–412.

[12] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action MACH: A spatio-temporal maximum average correlation height filter for action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[13] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, “A biologically inspired system for action recognition,” in *ICCV*, 2007, pp. 1–8.

[14] K. Schindler and L. V. Gool, “Action snippets: How many frames does action recognition require?” in *CVPR*, 2008.

[15] B. Laxton, J. Lim, and D. Kriegman, “Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video,” in *CVPR*, 2007, pp. 1–8.

[16] N. Olivera, 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.

[17] X. Feng and P. Perona, “Human action recognition by sequence of movelet codewords,” in *1st Int. Symp. on 3D Data Processing Visualization and Transmission (3DPVT2002)*, 2002, pp. 717–721.

[18] N. Ikizler and D. Forsyth, “Searching video for complex activities with finite state models,” in *CVPR*, 2007.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

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

[21] T. Hofmann, “Probabilistic latent semantic indexing,” in *SIGIR*, 1999, pp. 50–57.

[22] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” *Lecture Notes in Computer Science*, vol. 2749, pp. 363–370, 2003.

[23] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *International Conference on Computer Vision & Pattern Recognition*, vol. 2, June 2005, pp. 886–893.

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

[25] V. Vapnik, *The nature of statistical learning theory*. New York: Springer, 1995.

[26] C. Schüldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach,” in *Proceedings of International Conference on Pattern Recognition (ICPR’04)*, vol. 3, Cambridge, UK, 2004, pp. 32–36.

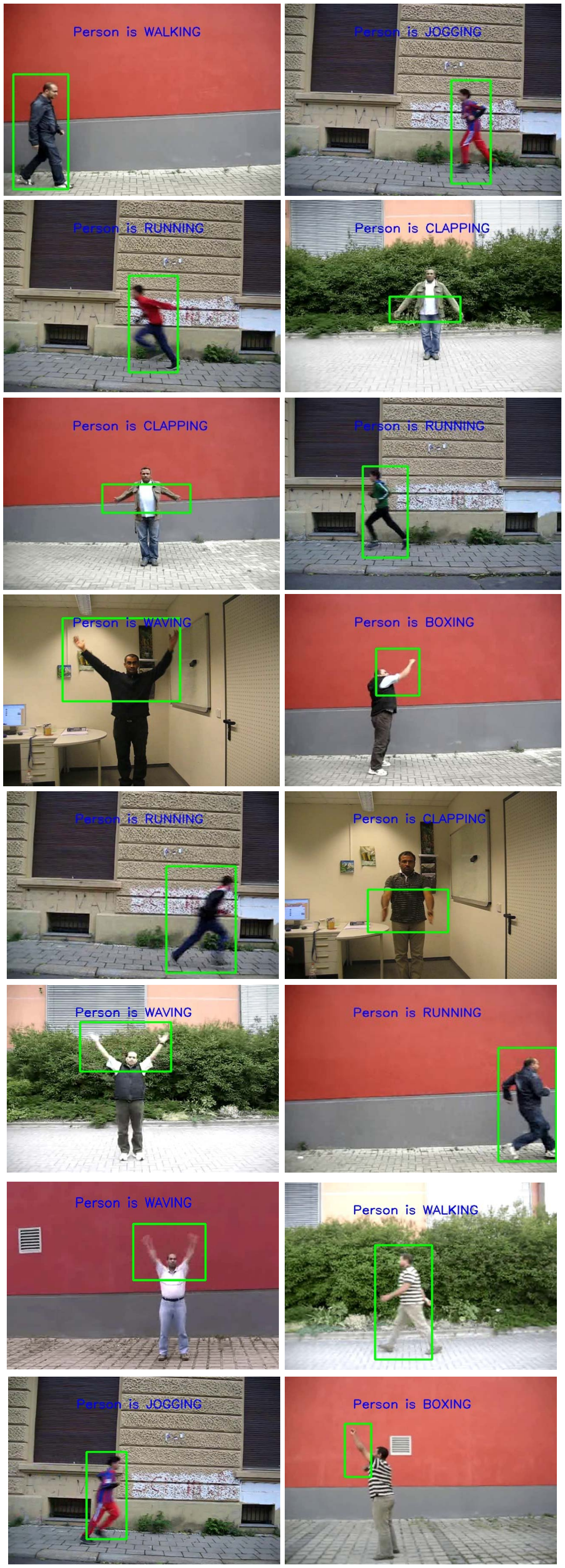


Figure 12: Some results of action localization and recognition in our dataset.