Research Article

Affine-Invariant Feature Extraction for Activity Recognition

Samy Sadek,¹ Ayoub Al-Hamadi,² Gerald Krell,² and Bernd Michaelis²

¹ Department of Mathematics and Computer Science, Faculty of Science, Sohag University, 82524 Sohag, Egypt

² Institute for Information Technology and Communications (IIKT), Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany

Correspondence should be addressed to Samy Sadek; samy.technik@gmail.com

Received 28 April 2013; Accepted 4 June 2013

Academic Editors: A. Gasteratos, D. P. Mukherjee, and A. Torsello

Copyright © 2013 Samy Sadek et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We propose an innovative approach for human activity recognition based on affine-invariant shape representation and SVM-based feature classification. In this approach, a compact computationally efficient affine-invariant representation of action shapes is developed by using affine moment invariants. Dynamic affine invariants are derived from the 3D spatiotemporal action volume and the average image created from the 3D volume and classified by an SVM classifier. On two standard benchmark action datasets (KTH and Weizmann datasets), the approach yields promising results that compare favorably with those previously reported in the literature, while maintaining real-time performance.

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 [1]. Although a great deal of progress has been made in automatic recognition of human actions during the last two decades, the approaches proposed in the literature remain limited in their ability. This leads to a need for much research work to be conducted to address the ongoing challenges and develop more efficient approaches. 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, for example, the search and the structuring of large video archives, human-computer interaction, video surveillance, gesture recognition, and robot learning and control. In fact, the nonrigid nature of human body and clothes in video sequences, resulting from drastic illumination changes, changing in pose, and erratic motion patterns, presents the grand challenge to human detection and action recognition. 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 human 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 performance. The remainder of this paper commences by briefly reviewing the most relevant literature in this area of human action recognition in Section 2. Then, in Section 3, we describe the details of the proposed method for action recognition. The experimental results corroborating the proposed method effectiveness are presented and analyzed in Section 4. Finally, in Section 5, we conclude and mention possible future work.

2. The Literature Overview

Recent few years have witnessed a resurgence of interest in more research on the analysis and interpretation of human motion motivated by the rise of security concerns and increased ubiquity and affordability of digital media production equipment. Human action can generally be recognized using various visual cues such as motion [2, 3] and shape [4, 5]. Scanning the literature, one notices that a significant body of work in human action recognition focuses on using spatial-temporal key points and local feature descriptors [6]. The local features are extracted from the region around each key point detected by the key point detection process. These features are then quantized to provide a discrete set of visual words before they are fed

FIGURE 1: GMM background subtraction: the first and third rows display two sequences of walking and running actions from KTH and Weizmann action datasets, respectively, while the second and fourth rows show the results of background subtraction where foreground objects are shown in cyan color.

into the classification module. Another thread of research is concerned with analyzing patterns of motion to recognize human actions. For instance, in [7], periodic motions are detected and classified to recognize actions. Alternatively, some researchers have opted to use both motion and shape cues. In [8], the authors detect the similarity between video segments using a space-time correlation model. In [9], Rodriguez et al. present a template-based approach using a Maximum Average Correlation Height (MACH) filter to capture intraclass variabilities. Likewise, a significant amount of work is targeted at modelling and understanding human motions by constructing elaborated temporal dynamic models [10]. There is also an attractive area of research which focuses on using generative topic models for visual recognition based on the so-called Bag-of-Words (BoW) model [11]. The underlying concept of a BoW is that each video sequence is 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 examples of commonly used topic models include Correlated Topic Models (CTMs) [11], Latent Dirichlet Allocation (LDA) [12], and probabilistic Latent Semantic Analysis (pLSA) [13].

3. Proposed Methodology

In this section, the proposed method for action recognition is described. The main steps of the framework are explained in detail along the following subsections.

3.1. Background Subtraction. In this paper, we use Gaussian Mixture Model (GMM) as a basis to model background distribution. Formally speaking, let X_t be a pixel in the current frame I_t , where t is the frame index. Then, each pixel can be modeled separately by a mixture of K Gaussians:

$$P(X_t) = \sum_{i=1}^K \omega_{i,t} \eta(X_t; \mu_{i,t}, \Sigma_{i,t}), \quad (1)$$

Where η is a Gaussian probability density function. $\mu_{i,t}$, $\Sigma_{i,t}$, and $\omega_{i,t}$ are the mean, covariance, and an estimate of the weight of the i th Gaussian in the mixture at time t , respectively. K is the number of distributions, which is set to 5 in experiments. Before the foreground is detected, the background is updated (see [14] for details about the updating procedure). After the updates are done, the weights $\omega_{i,t}$ are normalized. By applying a threshold T (set to 0.6 in our experiments), the background distribution remains on top with the lowest variance, where

$$B = \arg \min_b \left(\frac{\sum_{i=1}^b \omega_{i,t}}{\sum_{i=1}^K \omega_{i,t}} > T \right). \quad (2)$$

Finally, all pixels X_t that match none of the components are good candidates to be marked as foreground. An example of GMM background subtraction can be seen in Figure 1.

3.2. Average Images from 3D Action Volumes. The 3D volume in the spatio-temporal (XYT) domain is formed by piling

FIGURE 2: 2D average image created from the 3D spatio-temporal volume of a walking sequence.

up the target region in the image sequences of one action cycle, which is used to partition the sequences for the spatiotemporal volume. An action cycle is a fundamental unit to describe the action. In this work, we assume that the spatio-temporal volume consists of a number of small voxels. The average image $I_{av}(x, y)$ is defined as

$$I_{av}(x, y) = \frac{1}{\tau} \sum_{t=0}^{\tau-1} I(x, y, t), \quad (3)$$

where τ is the number of frames in action cycle (we use $\tau = 25$ in our experiments). $I(x, y, t)$ represents the density of the voxels at time t . An example of average image created from the 3D spatio-temporal volume of the running sequence is shown in Figure 2. For characterizing these 2D average images, the 2D affine moment invariants are considered as features [26].

3.3. Feature Extraction. As is well known, the moments describe shape properties of an object as it appears. Affine moment invariants are moment-based descriptors, which are invariant under a general affine transform. Six affine moment invariants can be conventionally derived from the central moments [27] as follows:

$$\begin{aligned} I_1 &= \frac{1}{\eta_{00}^4} [\eta_{20}\eta_{02} - \eta_{11}^2], \\ I_2 &= \frac{1}{\eta_{00}^{10}} [\eta_{03}^2\eta_{30}^2 - 6\eta_{30}\eta_{21}\eta_{12}\eta_{03} + 4\eta_{30}^3\eta_{12}^2 \\ &\quad + 4\eta_{03}^3\eta_{21}^2 - 3\eta_{21}^2\eta_{12}^2], \\ I_3 &= \frac{1}{\eta_{00}^7} [\eta_{20}(\eta_{21}\eta_{03} - \eta_{12}^2) - \eta_{11}(\eta_{30}\eta_{03} - \eta_{21}\eta_{12}) \\ &\quad + \eta_{02}(\eta_{03}\eta_{12} - \eta_{21}^2)], \end{aligned}$$

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

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

For a spatio-temporal (XYT) space, the 3D moment of order $(p + q + r)$ of 3D object $\mathcal{O}$ is derived using the same procedure of the 2D centralized moment:

$$\eta_{pqr} = \sum_{(x,y,t) \in \mathcal{O}} (x - x_g)^p (y - y_g)^q (t - t_g)^r I(x, y, t), \quad (5)$$

Where (x_g, y_g, t_g) is the centroid of object in the spatio-temporal space. Based on the definition of the 3D moment in (5), six 3D affine moment invariants can be defined. The first two of these moment invariants are given by

$$\begin{aligned} J_1 &= \frac{1}{\eta_{000}^5} [\eta_{200}\eta_{020}\eta_{002} + 2\eta_{110}\eta_{101}\eta_{011} - \eta_{200}\eta_{011}^2 \\ &\quad - \eta_{020}\eta_{101}^2 - \eta_{002}\eta_{110}^2], \\ J_2 &= \frac{1}{\eta_{000}^7} [\eta_{400}(\eta_{040}\eta_{004} + 3\eta_{022}^2 - 4\eta_{013}\eta_{031}) \\ &\quad + 3\eta_{202}(\eta_{040}\eta_{202} - 4\eta_{112}\eta_{130} + 4\eta_{121}^2) \\ &\quad + 12\eta_{211}(\eta_{022}\eta_{211} + \eta_{103}\eta_{130} - \eta_{031}\eta_{202} \\ &\quad - \eta_{121}\eta_{112}) \\ &\quad + 4\eta_{310}(\eta_{031}\eta_{103} - \eta_{004}\eta_{220} \\ &\quad + 3\eta_{013}\eta_{121} - 3\eta_{022}\eta_{112}) \\ &\quad + 3\eta_{220}(\eta_{004}\eta_{220} + 2\eta_{022}\eta_{202} \\ &\quad + 4\eta_{112} - 4\eta_{013}\eta_{311} - 4\eta_{121}\eta_{103}) \\ &\quad + 4\eta_{301}(\eta_{013}\eta_{130} - \eta_{040}\eta_{103} + 3\eta_{031}\eta_{112} \\ &\quad - 3\eta_{022}\eta_{121})]. \end{aligned} \quad (6)$$

Due to their long formulae, the remaining four moment invariants are not displayed here (refer to [28]). Figure 3

FIGURE 3: Plots of 2D affine moment invariants (I_i , $i = 1, \dots, 6$) computed on the average images of walking, jogging, running, boxing, waving, and clapping sequences.

shows a series of plots of 2D dynamic affine invariants with different action classes computed on the average images of action sequences.

3.4. Action Classification Using SVM. In this section, we formulate the action recognition task as a multiclass learning problem, where there is one class for each action, and the goal is to assign an action to an individual in each video sequence [1, 29]. There are various supervised learning algorithms by which action recognizer can be trained. Support Vector Machines (SVMs) are used in this work due to their outstanding generalization capability and reputation of a highly accurate paradigm [30]. SVMs that provide a best solution to data overfitting in neural networks are based on the structural risk minimization principle from computational theory. Originally, SVMs were designed to handle dichotomic classes in a higher dimensional space where a maximal separating hyperplane is created. On each side of this hyperplane, two parallel hyperplanes are conducted. Then, SVM attempts to find the separating hyperplane that maximizes the distance between the two parallel hyperplanes (see Figure 4). Intuitively, a good separation is achieved by the hyperplane having the largest distance. Hence, the larger the margin, the lower the generalization error of the classifier. Formally, let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}\}$ be a training dataset; Vapnik [30] shows that the problem is best

FIGURE 4: Generalized optimal separating hyperplane.

addressed by allowing some examples to violate the margin constraints. These potential violations are formulated with some positive slack variables ξ_i and a penalty parameter $C \geq 0$ that penalize the margin violations. Thus, the generalized optimal separating hyperplane is determined by solving the following quadratic programming problem:

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_i \xi_i \quad (7)$$

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

Geometrically, $\beta \in \mathbb{R}^d$ is a vector going through the center and perpendicular to the separating hyperplane. The offset parameter β_0 is added to allow the margin to increase and not to force the hyperplane to pass through the origin that restricts the solution. For computational purposes, it is more convenient to solve SVM in its dual formulation. This can be accomplished by forming the Lagrangian and then optimizing over the Lagrange multiplier α . The resulting decision function has weight vector $\beta = \sum_i \alpha_i \mathbf{x}_i y_i$, $0 \leq \alpha_i \leq C$. The instances $\mathbf{x}_i$ with $\alpha_i > 0$ are called *support vectors*, as they uniquely define the maximum margin hyperplane.

In the current approach, several classes of actions are created. Several one-versus-all SVM classifiers are trained using affine moment features extracted from action sequences in the training dataset. For each action sequence, a set of six 2D affine moment invariants is extracted from the average image. Also, another set of six 3D affine moment invariants is extracted from the spatio-temporal silhouette sequence. Then, SVM classifiers are trained on these features to learn various categories of actions.

4. Experiments and Results

To evaluate the proposed approach, two main experiments were carried out, and the results we achieved were compared with those reported by other state-of-the-art methods.

4.1. Experiment 1. We conducted this experiment using KTH action dataset [31]. To illustrate the effectiveness of the method, the obtained results are compared with those of other similar state-of-the-art methods. The KTH dataset contains action sequences, comprised of six types of human actions (i.e., walking, jogging, running, boxing, hand waving, and hand clapping). These 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 25 fps and a spatial resolution of 160×120 pixels over homogeneous backgrounds. To the best of our knowledge, there is no other similar dataset already available in the literature of sequences acquired on different environments. In order to prepare the experiments and to provide an unbiased estimation of the generalization abilities of the classification process, a set of sequences (75% of all sequences) performed by 18 subjects was used for training, and other sequences (the remaining 25%) performed by the other 7 subjects were set aside as a test set. SVMs with Gaussian radial basis function (RBF) kernel are trained on the training set, while the evaluation of the recognition performance is performed on the test set.

The confusion matrix that shows the recognition results achieved on the KTH action dataset is given in Table 1, while the comparison of the obtained results with those obtained by other methods available in the literature is shown in Table 3. As follows from the figures tabulated in Table 1, most actions are correctly classified. Furthermore, there is a high distinction between arm actions and leg actions. Most of the mistakes where confusions occur are between “jogging” and “running” actions and between “boxing” and

TABLE 1: Confusion matrix for the KTH dataset.

Action	Walking	Running	Jogging	Boxing	Waving	Clapping
Walking	0.94	0.01	0.05	0.00	0.00	0.00
Running	0.00	0.96	0.04	0.00	0.00	0.00
Jogging	0.04	0.08	0.88	0.00	0.00	0.00
Boxing	0.00	0.00	0.00	0.94	0.02	0.04
Waving	0.00	0.00	0.00	0.02	0.93	0.05
Clapping	0.00	0.00	0.00	0.01	0.03	0.96

“clapping” actions. This is intuitively plausible due to the fact of high similarity between each pair of these actions. From the comparison given by Table 3, it turns out that our method performs competitively with other state-of-the-art methods. It is pertinent to mention here that the state-of-the-art methods with which we compare our method have used the same dataset and the same experimental conditions; therefore, the comparison seems to be quite fair.

4.2. Experiment 2. This second experiment was conducted using the Weizmann action dataset provided by Blank et al. [32] in 2005, which contains a total of 90 video clips (i.e., 5098 frames) performed by 9 individuals. Each video clip contains one person performing an action. There are 10 categories of action involved in the dataset, namely, *walking, running, jumping, jumping in place, bending, jacking, skipping, galloping sideways, one-hand waving, and two-hand waving*. Typically, all the clips in the dataset are sampled at 25 Hz and last about 2 seconds with image frame size of 180×144 . In order to provide an unbiased estimate of the generalization abilities of the proposed method, we have used the leave-one-out cross-validation (LOOCV) technique in the validation process. As the name suggests, this involves using a group of sequences from a single subject in the original dataset as the testing data and the remaining sequences as the training data. This is repeated such that each group of sequences in the dataset is used once as the validation. Again, as with the first experiment, SVMs with Gaussian RBF kernel are trained on the training set, while the evaluation of the recognition performance is performed on the test set.

The confusion matrix in Table 2 provides the recognition results obtained by the proposed method, where correct responses define the main diagonal. From the figures in the matrix, a number of points can be drawn. The majority of actions are correctly classified. An average recognition rate of 97.8% is achieved with our proposed method. What is more, there is a clear distinction between arm actions and leg actions. The mistakes where confusions occur are only between *skip* and *jump* actions and between *jump* and *run* actions. This intuitively seems to be reasonable due to the fact of high closeness or similarity among the actions in each pair of these actions. In order to quantify the effectiveness of the method, the obtained results are compared qualitatively with those obtained previously by other investigators. The outcome of this comparison is presented in Table 3. In the light of this comparison, one can see that the proposed method is competitive with the state-of-the-art methods.

TABLE 2: Confusion matrix for the Weizmann dataset.

Action	Bend	Jump	Pjump	Walk	Run	Side	Jack	Skip	Wave 1	Wave 2
Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pjump	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Walk	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Run	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.10	0.00	0.00
Side	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Jack	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Skip	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.90	0.00	0.00
Wave 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

TABLE 3: Comparison with the state of the art on the KTH and Weizmann datasets.

Method	KTH	Weizmann
Our method	93.5%	98.0%
Liu and Shah [15]	92.8%	—
Wang and Mori [16]	92.5%	—
Jhuang et al. [17]	91.7%	—
Rodriguez et al. [9]	88.6%	—
Rapantzikos et al. [18]	88.3%	—
Dollár et al. [19]	81.2%	—
Ke et al. [20]	63.0%	—
Fathi and Mori [21]	—	100%
Bregonzio et al. [22]	—	96.6%
Zhang et al. [23]	—	92.8%
Niebles et al. [24]	—	90.0%
Dollár et al. [19]	—	85.2%
Kläser et al. [25]	—	84.3%

It is worthwhile to mention that all the methods that we compared our method with, except the method proposed in [21], have used similar experimental setups; thus, the comparison seems to be meaningful and fair. A final remark concerns the real-time performance of our approach. The proposed action recognizer runs at 18fps on average (using a 2.8 GHz Intel dual core machine with 4 GB of RAM, running 32-bit Windows 7 Professional).

5. Conclusion and Future Work

In this paper, we have introduced an approach for activity recognition based on affine moment invariants for activity representation and SVMs for feature classification. On two

benchmark action datasets, the results obtained by the proposed approach were compared favorably with those published in the literature. The primary focus of our future work will be to investigate the empirical validation of the approach on more realistic datasets presenting many technical challenges in data handling, such as object articulation, occlusion, and significant background clutter.

References

- [1] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "Recognizing human actions: a fuzzy approach via chord-length shape features," *ISRN Machine Vision*, vol. 1, pp. 1–9, 2012.
- [2] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, vol. 2, pp. 726–733, October 2003.
- [3] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "Towards robust human action retrieval in video," in *Proceedings of the British Machine Vision Conference (BMVC '10)*, Aberystwyth, UK, September 2010.
- [4] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "Human activity recognition: a scheme using multiple cues," in *Proceedings of the International Symposium on Visual Computing (ISVC '10)*, vol. 1, pp. 574–583, Las Vegas, Nev, USA, November 2010.
- [5] S. Sadek, A. Al-Hamadi, M. Elmezain, B. Michaelis, and U. Sayed, "Human activity recognition via temporal moment invariants," in *Proceedings of the 10th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT '10)*, pp. 79–84, Luxor, Egypt, December 2010.
- [6] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, Article ID 540375, 2011.
- [7] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on*

- Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, 2000.
- [8] E. Shechtman and M. Irani, “Space-time behavior based correlation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’05)*, vol. 1, pp. 405–412, June 2005.
 - [9] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action MACH: a spatio-temporal maximum average correlation height filter for action recognition,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’08)*, June 2008.
 - [10] N. Ikizler and D. Forsyth, “Searching video for complex activities with finite state models,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’07)*, June 2007.
 - [11] D. M. Blei and J. D. Lafferty, “Correlated topic models,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 18, pp. 147–154, 2006.
 - [12] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.
 - [13] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’99)*, pp. 50–57, 1999.
 - [14] S. J. McKenna, Y. Raja, and S. Gong, “Tracking colour objects using adaptive mixture models,” *Image and Vision Computing*, vol. 17, no. 3-4, pp. 225–231, 1999.
 - [15] J. Liu and M. Shah, “Learning human actions via information maximization,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’08)*, June 2008.
 - [16] Y. Wang and G. Mori, “Max-Margin hidden conditional random fields for human action recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR ’09)*, pp. 872–879, June 2009.
 - [17] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, “A biologically inspired system for action recognition,” in *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV ’07)*, pp. 257–267, October 2007.
 - [18] K. Rapantzikos, Y. Avrithis, and S. Kollias, “Dense saliency-based spatiotemporal feature points for action recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR ’09)*, pp. 1454–1461, June 2009.
 - [19] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS ’05)*, pp. 65–72, October 2005.
 - [20] Y. Ke, R. Sukthankar, and M. Hebert, “Efficient visual event detection using volumetric features,” in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV ’05)*, pp. 166–173, October 2005.
 - [21] A. Fathi and G. Mori, “Action recognition by learning mid-level motion features,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’08)*, June 2008.
 - [22] M. Bregonzio, S. Gong, and T. Xiang, “Recognising action as clouds of space-time interest points,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR ’09)*, pp. 1948–1955, June 2009.
 - [23] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, “Motion context: a new representation for human action recognition,” in *Proceeding of the European Conference on Computer Vision (ECCV ’08)*, vol. 4, pp. 817–829, 2008.
 - [24] J. C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.
 - [25] A. Kläser, M. Marszaek, and C. Schmid, “A spatiotemporal descriptor based on 3D-gradients,” in *Proceedings of the British Machine Vision Conference (BMVC ’08)*, 2008.
 - [26] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, “Human action recognition via affine moment invariants,” in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR ’12)*, pp. 218–221, Tsukuba Science City, Japan, November 2012.
 - [27] J. Flusser and T. Suk, “Pattern recognition by affine moment invariants,” *Pattern Recognition*, vol. 26, no. 1, pp. 167–174, 1993.
 - [28] D. Xu and H. Li, “3-D affine moment invariants generated by geometric primitives,” in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR ’06)*, pp. 544–547, August 2006.
 - [29] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, “An SVM approach for activity recognition based on chord-length-function shape features,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP ’12)*, pp. 767–770, Orlando, Fla, USA, October 2012.
 - [30] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.
 - [31] C. Schödl, I. Laptev, and B. Caputo, “Recognizing human actions: a local SVM approach,” in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR ’04)*, pp. 32–36, 2004.
 - [32] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV ’05)*, vol. 2, pp. 1395–1402, October 2005.