Automated Detection of Similar Human Actions using Motion Descriptors

Ammar Ladjailia*†, Imed Bouchrika†, Hayet Farida Merouani* and Nouzha Harrati†‡
†*Faculty of Science and Technology, University of Souk Ahras, Algeria*
**Department of Computer Science, University of Annaba, Algeria*
‡*Department of Computer Science, University of Bejaia, Algeria*
a.ladjailia@univ-soukahras.dz

***Abstract*—As computing becomes ubiquitous in our modern society, automated recognition of human activities emerges as a crucial topic where it can be applied to many real-life human-centric scenarios such as smart automated surveillance, human computer interaction and automated refereeing. In this research study, a motion descriptor is constructed based on the extraction of optical flow features across consecutive frames for the classification of human activities. A histogram of features is derived from images taking into account the solely local properties embedded within the motion map. Feature selection which is based on the proximity of instances belonging to the same class is performed to obtain the most distinctive features. Experimental results carried out on the Weizmann dataset confirmed the potency for the proposed method with a high recognition rate of 95.02 % to distinguish between different basic human action classes such as running, walking, waving and jumping. The dataset is made of 19 basic actions for 9 different subjects. Further experiments are conducted to assess the ability of the proposed approach to recognize similar actions based on the intra and inter class distribution analysis.**

***Keywords*-Activity Recognition, Motion Descriptor, Optical Flow, Feature Selection**

I. Introduction

Much research within the computer vision community is dedicated towards the analysis of and understanding of human motion. Such study is fueled by the wide range of applications where human motion analysis can be deployed such as smart automated surveillance, biometrics, human computer interaction and sport refereeing and analysis. As computing becomes ubiquitous in our modern society, the recognition of human activities emerges as a crucial topic where it can be applied to many real-life human-centric scenarios [1], [2]. Furthermore, given the immense expansion of video data being recorded in everyday life from security surveillance cameras, movies productions and internet video uploads, it becomes an essential need to automatically analyse and understand video content semantically. This is to ease the process of video indexing and fast retrieval of data when dealing with large multimedia content and big data. Hence, the importance of automated systems for human activity recognition is central to the success of such applications.

Human activity recognition aims to automatically infer the action or activity being performed by a person. For instance, recognizing whether someone is walking, raising hands, or performing other types of activities. In the vision literature, both terms "Activity" and "Action" are used interchangeably and contentiously but every term has its rough definition [1]. An $action$ is considered as a simple activity referring to simple pattern performed by a person during a short period of time lasting a few seconds. Examples of actions may include raising hands, bending, sitting and walking. Poppe [1] described additionally the term $action\ primitive$ which refers to an atomic movement at the limb level. On the other hand, an $activity$ is considered as a composite sequence of actions executed by either a single person or several people interacting with each other. Examples of activities are like leaving an unattended bag, shaking hands or assaulting a pedestrian. A visual automated system for human activity analysis consists of three main consecutive stages: detection, camera-intra tracking and perception of the action or activity being performed.

The automated marker-less extraction and recognition of human activities are proven to be a challenging task. Although, the problem can be stated in simple terms, given a sequence of frames with one or more people performing a given activity, can an automated system recognize the activity being performed. The solution is difficult to devise or implement. The difficulties stem from three substantial factors related either to : person, acquisition environment and activity understanding. Most of the existing methods proposed for human activity recognition rely on sensors or special markers mounted on the subject [1], [2]. For a marker-less approach, the articulated nature of human body which encompasses a wide range of possible motion transformations in addition to self-occlusion and appearance variability, exacerbate further complexity on the task of visual feature extraction [3]. Challenges related to the acquisition environment may include background clutter, illumination, camera movement and viewpoint as well as occlusion by other objects in the scene. Lastly, an activity can be performed at various ways by different people depending on the context [4] or even culture of the performer. Inversely, the same activity performed by different people can have different semantic meanings. Furthermore, activities can interleave within each other and performed in parallel rather than a sequential fashion. For instance a person can use their computer whilst eating at the same time or answering the phone.

Due to the incontestable role of automated human activity recognition in smart surveillance and security applications, we investigate in this research a marker-less motion-based descriptor for the classification of human actions. The method is not dependent on background

segmentation due to the nature of surveillance imageries subjected to various conditions. Instead, motion features are estimated through computing optical flow from a triplet of consecutive frames to obtain a descriptive number describing the temporal flow orientation at every pixel. A histogram of features is constructed from consecutive images taking into account purely different various numerous the local properties. Feature selection based on the proximity of instances belonging to the same class is applied to derive the most discriminative features. Experimental results carried out on the Weizmann dataset confirmed the potency for the proposed method to better distinguish between different human action classes such as running, walking, waving and jumping with the potency to extend the training procedure to recognize further complex activities. Based on analysing the intra and inter class distributions for various human action classes, the system is further trained to infer the degree of similar actions based on the proposed motion descriptor.

II. Related Work

The recognition of human activity is of prime importance for various applications as automated visual surveillance. The research area of human activity recognition is closely related to other fields of research that analyze human motion such as human computer interaction and biomechanical engineering. Although, there is a considerable body of work devoted to human action recognition, most of the methods are evaluated on datasets recorded in simplified settings. More recent research has shifted focus to natural activity recognition in unconstrained scenes [5]. Poppe [1] and Vishwakarma [2] surveyed the recent methods, research studies and datasets devoted to this area of research. Existing methods can be broadly classified into two major categories in terms of image representation which are either global or local representation. From another perspective, the temporal dimension is taken into account explicitly for image representations in addition to the spatial information meanwhile other methods extract image features on a frame by frame basis.

For the global representation, the region of interest (ROI) of a person is encoded as a whole. The subject is usually derived from an image through applying background subtraction. The processing of global representations is based on low-level information taken from silhouettes, edges or optical flow [1]. However, these methods are susceptible to noise, occlusions and variations in camera viewpoint. Wang *et al* [6] applied the R transform on the extracted silhouettes reporting that the obtained representation is translation and scale invariant. The main benefit of the R transform is its low computational cost as well as it geometric invariance. A set of HMMs are employed for training the extracted features in order to recognize activities. Weinland *et al* [7] described a compact and efficient representation which is based on matching a set of discriminative static landmark pose models. The method does not depend on or take into account the temporal ordering of sequences. In their work, silhouette models are matched against edge data using the Chamfer distance and therefore eliminating the need for background segmentation. Ali and Shah [8] derived a set of kinematic-based features from the optical flow such as divergence, velocity, symmetric and anti-symmetric flow fields. Multiple instance learning method is used together with Principal Component Analysis to determine the kinematic modes.

For activity recognition using local representations, a collection of independent patches within an image are analyzed to generate a discriminative feature vector for the observed activity. Local representations do not require accurate localization or background subtraction and enjoy the benefits of being to some extent invariant to appearance transformation, background clutter and partial occlusion [1]. Local patches are described by local grid-based descriptors that would summarize locally the observation within grid cells for the case of still frames. Yeffet *et al* [9] proposed a local trinary pattern descriptor for encoding human motion from a sequence of frames. The trinary number is generated from a matching process of patches of a given frame against adjacent patches residing on both the previous and next frames respectively. A histogram-based feature vector is constructed from the concatenation resulting from the image divided into a grid. As an extension of their work, Kliper-Gross [10] employed the same approach of the local trinary motion pattern renamed as Motion Interchange Pattern (MIP) for the automated recognition of human activities. However, they have used bag of features for the classification stage instead together with SVM. Oshin [5] utilized the relative distribution of spatio-temporal interest points for activity recognition in unconstrained scenarios.

Kliper-Gross *et al* [11] proposed the ASLAN Action Similarity LAbeliNg database with an evaluation benchmark protocol. The main contribution of their work is constructing a wide database of videos with hundreds of complex actions dedicated mainly to inferring the similarity of actions, rather than the classification or recognition of human actions. The ASLAN dataset includes over 400 complex action classes. The benchmark protocols focus on action similarity as a binary classification problem with $same$ or $not - same$ output values. For their proposed benchmark protocol, they have reported a success rate of 65.3% using a fused set of descriptors. Burghouts [12] described a saliency measure for each individual feature point that would enhance the distinctiveness for a given action through the use of bag of features.

III. Proposed Approach

A. Motion Flow Descriptor

The propose approach encodes a sequence of frames into a feature vector describing the performed basic action by a person [13]. The method does not depend on background subtraction for the derivation of motion features. This is because it is computationally expensive and complex to deploy background subtraction for real-time surveillance applications due the process of updating

the background model which is influenced by a number of factors such as background clutter, weather conditions and other outdoor environmental effects. Inspired by the work of Kliper-Gross [10] for proposing the Motion Interchange Pattern for action recognition together with the fact that local descriptors are known for their effectiveness and robustness for encoding texture for recognition purposes including biometrics, we have proposed a local descriptor which captures the motion of the local structure based on estimating optical flow. Provided that there is a motion of a small patch at frame t to the next frame $t+1$, there is a high probability that a similar patch would be induced within the neighboring region of the original patch position at the previous frame. The proposed descriptor is based on constructing a feature that reflects the patch displacement from frame to frame based.

Because of the common increase of image self-similarity regions, the block matching using simple similarity operators can fail in distinguishing to between similarity caused by motion and similar static textures. In addition, the matching can be difficult as moving patches may have their appearances changed due to the non-rigid nature of the human motion. In this research, the optical flow is instead harnessed to better estimate the motion information from video sequences. Optical flow is one of the most active research area in computer vision due to their central role in various fields of applications such as autonomous vehicle or robot navigation, visual surveillance and fluid flow analysis. The main basis of optical flow is to observe the displacement of intensity patterns [14]. This pattern is a result of the apparent motion of objects, surfaces, and edges in a visual scene caused by the relative movement between an observer and the scene [15]. In other words, optical flow can arise either from the relative motion of the object or camera. For a given image I, the constraint for optical flow states that the gray intensity value of a moving pixel $I(x, y, t)$ at time t stays constant over time as expressed as:

$$I(x, y, t) - I(x + V_x, y + V_y, t + 1) = 0 \tag{1}$$

such that V_x, V_y is the optical flow velocity vector for a pixel $p(x, y)$ from time t to $t+1$. The intensity constancy hypothesis can also be written in the differential form shown in the following Equation:

$$\frac{dI}{dt} = 0 \tag{2}$$

Equation (2) can be rewritten using the chain rule of differentiation as given below :

$$\frac{\partial I}{\partial x}\ \frac{dx}{dt} + \frac{\partial I}{\partial y}\ \frac{dy}{dt} + \frac{\partial I}{\partial t} = 0 \tag{3}$$

Such that $\partial I/x$ which is abbreviated as E_x in this paper, is the partial derivative for the image with respect to x. The two unknowns which are the optical flow parameters are given in the following equation:

$$v = \begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{bmatrix} \tag{4}$$

To solve Equation (3) which has two unknowns, constraints are required to be added to ease finding a solution. There are several solutions proposed in the literature. Differential methods are the most used method. In this article, the method of Lucas-Kanade [16] is considered for estimating the optical flow vector. The method is based on the principle that relative motion of brightness content between two successive images is small and approximatively constant within a local neighborhood of a given point p. Therefore, the optical flow equation is assumed to hold for all points within the smaller neighborhood region centered at p. The lucas-kanade method solves the inherent ambiguity of the optical flow equation via combining information for several close pixels. The local image flow vector ($V_x,\ V_y$) must satisfy the following :

$$\begin{cases} E_x(q_1)\,V_x + E_y(q_1)\,V_y = -E_t(q_1) \\ E_x(q_2)\,V_x + E_y(q_2)\,V_y = -E_t(q_2) \\ \vdots \\ E_x(q_n)\,V_x + E_y(q_n)\,V_y = -E_t(q_n) \end{cases} \tag{5}$$

such that q_a is a^{th} pixel within the small region centered at the point $p(x, y)$. n is the number of points. These equations can be written in matrix form $Av = b$, where:

$$A = \begin{bmatrix} E_x(q_1) & E_y(q_1) \\ \vdots & \vdots \\ E_x(q_n) & E_y(q_n) \end{bmatrix} \tag{6}$$

and

$$b = \begin{bmatrix} -E_t(q_1) \\ \vdots \\ -E_t(q_n) \end{bmatrix} \tag{7}$$

The above system has more equations than unknowns and therefore it is considered over-determined. Through the use of the least squares principle, the LucasKanade method finds a compromise solution for the 2×2 system as given in the following Equation :

$$v = \left(A^T A\right)^{-1} A^T b \tag{8}$$

Consequently, the optical flow vector v is estimated as:

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = C^{-1} K \tag{9}$$

such that:

$$C = \begin{bmatrix} \sum_i^n E_x(q_i)^2 & \sum_i^n E_x(q_i) E_y(q_i) \\ \sum_i^n E_y(q_i) E_x(q_i) & \sum_i^n E_y(q_i)^2 \end{bmatrix} \tag{10}$$

$$k = \begin{bmatrix} -\sum_i^n E_x(q_i) E_t(q_i) \\ -\sum_i^n E_y(q_i) E_t(q_i) \end{bmatrix} \tag{11}$$

Based on a triplet of frames denoted as $previous$, $current$ and $next$, a descriptor number d is constructed for every pixel for the current image through computing two optical flow images for $v_{prev} : \{previous, current\}$ and $v_{next} : \{current, next\}$. We apply thresholding based on the magnitude of the velocity flow considering only values greater than $\tau = 0.5$. Based on the location of the

angular values within the polar coordinate system which is equally divided into 8 numbered sections of 40 degrees from 1 to 8, the optical flow vector is converted into a number reflecting the order within the eighth circular portions. This is denoted using the function $AngIndex$ as expressed in Equation (12). The zero indexes refer that there is no motion where the magnitude of the optical flow is less than the threshold τ . Both of the two digits resulting at every pixel from the next and previous frames are concatenated together to generate a number of base 8 which is converted to a decimal number.

$$d = AngIndex(v_{prev}) + AngIndex(v_{next}) * 8 \quad (12)$$

The number d serves as a descriptor for the motion at a pixel level. Experimentally, we have observed that a simple action can be fully contained within only 15 frames based on video recorded at a frame rate of 25. Therefore, the encoding process is performed for every pixel for the seven different triplets of consecutive frames taken from a video. The motion orientation histogram for a triplet is computed as shown in Equation (13). Figure (1) outlines the procedure to estimate the histogram of motion-based features using optical flow. b is a Boolean function returning 1 for true cases and 0 for false conditions

$$H_i = \sum_{t=1}^{5} \sum_{x,y} b(d(x,y,t) == i) \quad (13)$$

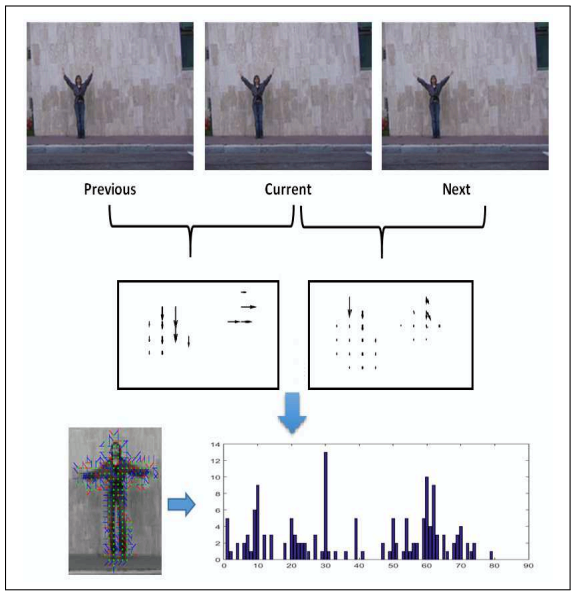

Figure 1. Histogram estimation using Optical Flow

In this research, various features that could potentially describe better the motion are generated based on simple fusion operations including summation and statistical operators being applied on the set of motion orientation histograms for the triplets of frames. Equation (14) shows the obtained feature vector by concatenation of different histograms. The resulting action vector consists of features describing purely local motion features of the human body without any information describing the global structure of the activity nor the anthropometric measurements of the human body.

$$F = [\sum his \quad mean\,(hist) \quad std(hist)] \quad (14)$$

B. Feature Selection

The feature selection process is considered in this research to derive the most discriminative features and suppress the redundant and irrelevant components which may degrade the classification rate. Because it is infeasible to conduct a brute force search procedure for all possible combinations of subsets to derive the optimal feature subset due to the high dimensionality of the raw feature vector. Alternatively, the Adaptive Sequential Forward Floating Selection (ASFFS) search algorithm [17] is harnessed to reduce the number of features.

The feature subset selection procedure is purely based on an evaluation function that assesses the discriminativeness of each component or set of features in order to derive the optimal subset of features for the classification process [18]. We present a validation-based evaluation criterion to pick up the subset of features that would minimise the classification errors and ensure larger inter-class separability between the different classes. As opposed to the voting paradigm employed by the *KNN* classifier, the evaluation function utilises coefficients w that signify the significance of the most nearest neighbours of the same class. The probability score for a candidate s_c to belong to a cluster c is expressed in the following Equation (15) as:

$$f(s_c) = \frac{\sum_{i=1}^{N_c-1} z_i w_i}{\sum_{i=1}^{N_c-1} w_i} \quad (15)$$

such that N_c is the number of candidates within the cluster c, and the coefficient w_i for the i^{th} nearest instance is inversely related to closeness as given:

$$w_i = (N_c - i)^2 \quad (16)$$

The value of z_i is given as:

$$z_i = \begin{cases} 1 & \text{if } nearest(s_c, i) \in c \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

Where the $nearest(s_c, i)$ function returns the i^{th} closest instance to the instance s_c. The Euclidean distance measure is employed to infer the nearest neighbours from the same class. The significance for a subset of features is entirely based on the validation-based metric which is estimated using the leave-one-out cross-validation rule. The human action signature is made as the subset of features S among the feature space F attaining the maximum value which is the average sum of f computed across the N samples x as shown the following equation:

$$Signature = \arg\max_{S \in F} \left(\frac{\sum_{x=1}^{N} f_S(x)}{N} \right) \quad (18)$$

In order to infer whether two actions are similar, intra and inter-class distribution analysis is performed using all possible pairs from the dataset via computing the

simple Euclidean Distance as a similarity measure. This is carried out using the derived subset of normalized features ensuring better discriminability between different action classes using the Weizmann dataset described in the following section. Based on the intra and inter class distributions being plotted for this case study, the threshold $\tau = 0.267$ is estimated as the intersection point between the two Gaussian distributions which is harnessed to deduce whether a given two instances of human actions belong to the same class or not.

IV. Experimental Results

For the evaluation of motion-based local features derived using the marker-less method for human action recognition, the proposed method is tested on the Weizmann dataset [19] which contains 90 video sequences with low-resolution of 180×144 recorded at frame rate of 25 frames per second. There are nine different people, each performing 10 actions such as running, walking, skipping, jumping-jack, jumping forward on two legs and waving one hand. Figure (2) In this study, we manually collected a dataset containing 218 video sequences for 19 different simple actions by decomposing an activity into primitive actions. Each video consists of 15 frames which are all checked to better describe the complete action. For instance, the activity of waving one hand can be split into two basic actions including rising the hand upwards and than lowering it down.

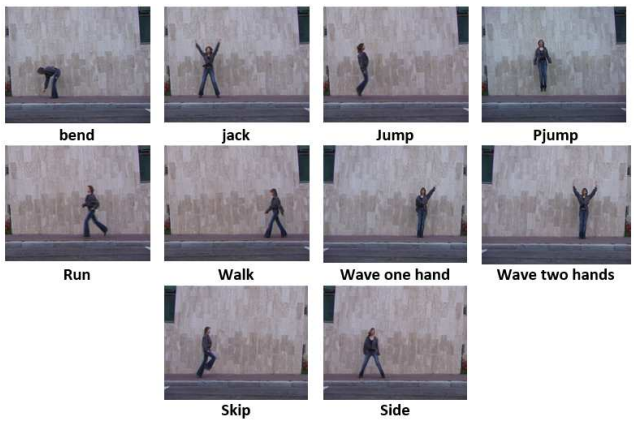

Figure 2. Weizmann dataset

The feature selection process with the proposed evaluation criterion is applied on the extracted raw visual features. An optimal action signature is derived containing only 30 features. The Correct Classification Rate is computed using the K-nearest neighbour (KNN) rule for the value of $k = 3$ using the leave-one-out cross-validation. The KNN classifier is chosen at the classification phase because of its simplicity and therefore fast computation besides the convenience of comparison to other existing studies. Using the Cumulative Match Score (CMS) evaluation method which was described by Phillips in the FERET protocol for face recognition, we have correctly classified 95.02% of the 19 basic actions at rank $R = 1$ and 100% at rank $R = 9$. The achieved results are promising because the recognition is based purely on local motion information and this can be boosted through adding global features. The achieved results promising because the recognition is based purely on local motion information and this can be boosted through adding global features. The confusion matrix is shown in Figure (3) which visualizes the separation results across the different clusters. The lighter squares reflect higher separation scores and thus better discriminability. The dark blue diagonal line reflects the zero distance between the same classes. The separation distance between the different clusters is computed using the Euclidean distance metric. Table (1) shows comparative results for different methods for human activity recognition on the same Weizmann dataset. Although, we have chosen to consider different number of action classes, the obtained results inspire to certain potentials in addressing the intricate issue of human activity recognition via decomposition into simple actions.

Method	Num. Actions	CCR
Our method	19	95.00 %
Yao & al [20]	10	92.20 %
Almotairi & Ribeiro[21]	10	92.22%
Liu & al [22]	10	90.40%

Table I
Comparative Results for the Weizmann Dataset

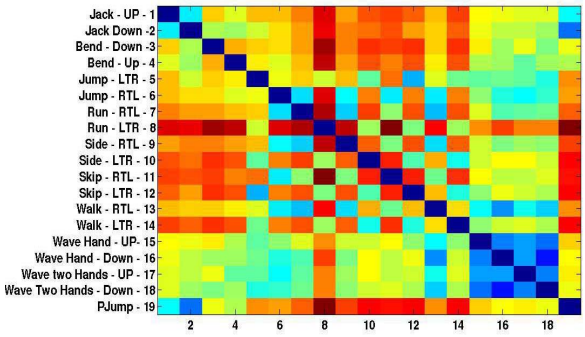

Figure 3. Confusion matrix for cross-matching of action recognition

The Receiver Operating Characteristics (ROC) curve is plotted in Figure 4 to show the verification results for estimating the similarity between two instances across all pairs. In the verification process, the instances from database are verified to check if they belong to the claimed class labels based on computing the Euclidean distance. The thresholding function described in Feature Selection section is used to express whether the two pairs belong to the claimed class. In order to plot the False Acceptance Rate (FAR) versus the False Rejection Rate (FRR), different score thresholds are used. Using the human action signature derived from dynamics, the system achieved equal error rate of 1.89% is obtained.

V. Conclusions

Automated recognition of human activity is central for success of various applications as smart visual surveillance. In this research, a motion interchange descriptor is

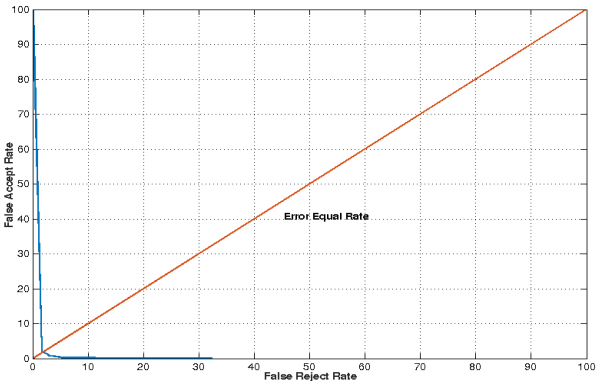

Figure 4. Receiver Operating Characteristic (ROC) Plot

employed for the extraction of features across consecutive frames for the classification of human activities. A histogram of features is constructed from the image taking into account the global and local properties embedded within the motion map. Feature selection based on the proximity of instances belonging to the same class is applied to derive the most discriminative features. Experimental results carried out on the Weizmann dataset confirmed the potency for the proposed method to better distinguish between different activity classes. Based on analysing the intra and inter class distributions for various human action classes, the system is further trained to infer the degree of similar actions based on the proposed motion descriptor. The obtained results reveal certain potentials in addressing the intricate issue of human activity recognition via decomposition into simple actions.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[2] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

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

[4] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 1, pp. 91–101, 2013.

[5] O. Oshin, A. Gilbert, and R. Bowden, "Capturing relative motion and finding modes for action recognition in the wild," *Computer Vision and Image Understanding*, vol. 125, pp. 155–171, 2014.

[6] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on r transform," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[7] D. Weinland and E. Boyer, "Action recognition using exemplar-based embedding," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

[8] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 288–303, 2010.

[9] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 492–497.

[10] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *European Conference on Computer Vision*. Springer, 2012, pp. 256–269.

[11] O. Kliper-Gross, T. Hassner, and L. Wolf, "The action similarity labeling challenge," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 3, pp. 615–621, 2012.

[12] G. J. Burghouts, S. van den Broek, and R. ten Hove, "Spatio-temporal saliency for action similarity," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*. IEEE, 2013, pp. 257–262.

[13] A. Ladjailia, I. Bouchrika, H. F. Merouani, and N. Harrati, "On the use of local motion information for human action recognition via feature selection," in *4th IEEE International Conference on Electrical Engineering (ICEE)*, 2015.

[14] D. Fortun, P. Bouthemy, and C. Kervrann, "Optical flow modeling and computation: a survey," *Computer Vision and Image Understanding*, vol. 134, pp. 1–21, 2015.

[15] A. Burton and J. Radford, *Thinking in perspective: critical essays in the study of thought processes*. Methuen, 1978.

[16] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision." in *IJCAI*, vol. 81, 1981, pp. 674–679.

[17] P. Somol, P. Pudil, J. Novovičová, and P. Paclık, "Adaptive floating search methods in feature selection," *Pattern recognition letters*, vol. 20, no. 11, pp. 1157–1163, 1999.

[18] I. Bouchrika, "Gait analysis and recognition for automated visual surveillance," Ph.D. dissertation, University of Southampton, 2008.

[19] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1395–1402.

[20] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2061–2068.

[21] S. Almotairi and E. Ribeiro, "Action classification using sequence alignment and shape context," in *The Twenty-Seventh International Flairs Conference*, 2014.

[22] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.