On the use of Local Motion Information for Human Action Recognition via Feature Selection

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*—Automated recognition of human activities has received considerable attention within the computer vision community. This is mainly due to the plethora of applications where human activity recognition can be deployed such as smart automated surveillance and human computer interaction. In this research study, a motion descriptor is 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 solely 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 such as running, walking, waving and jumping. The dataset is made of 19 basic actions for 9 different subjects.**

***Keywords*-Activity Recognition, Motion Descriptor, Human Activity, Feature Selection**

I. Introduction

Much research within the computer vision community is devoted towards the analysis of human motion. Such research is fueled by the wide range of applications where human motion analysis can be deployed such as smart automated surveillance [1], biometrics, human computer interaction and sport automated 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 [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.

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 main 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. [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 feature extraction. Challenges related to the acquisition environment may include background clutter, illumination, camera movement and viewpoint as well as occlusion. Lastly, an activity can be performed at various ways by different people depending on the context [3] 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. For instance a person can use their computer whilst eating at the same time or answer the phone.

Because of the vital role of automated human activity recognition in smart surveillance and security applications, we explore 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. A histogram of features is constructed from consecutive images taking into account purely different various numerous the local properties embedded within the generated motion map based on matching of adjacent patches. 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.

II. Related Work

The recognition of human activity is of prime importance for various applications as automated visual surveillance. 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 [4]. Poppe

[5] and Vishwakarma [6] 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. 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 [5]. However, these methods are susceptible to noise, occlusions and variations in camera viewpoint. Weinland *et al* [7] described a compact and efficient representation which is based on matching a set of discriminative static landmark pose models. In their work, silhouette models are matched against edge data using the Chamfer distance. 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 [5]. 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). However, they have used bag of features for the classification stage instead together with SVM. Oshin [4] utilized the relative distribution of spatio-temporal interest points for activity recognition in unconstrained scenarios.

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. 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. 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. 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:

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

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 (5). 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.

$$d = AngIndex(v_{prev}) + AngIndex(v_{next}) * 8 \quad (5)$$

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 (6). 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

$$H_i = \sum_{t=1}^{5} \sum_{x,y} b(d(x, y, t) == i) \quad (6)$$

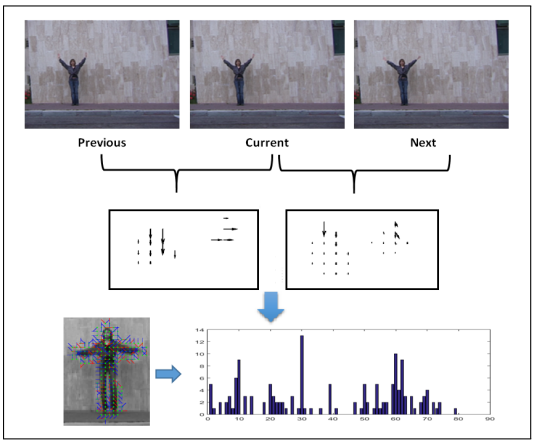

Figure 1. Histogram estimation using Optical Flow

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. The resulting action vector consists of 1,782 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.

B. Feature Selection

Feature selection is considered within this research to derive the discriminative features and remove the redundant and irrelevant components which may affect the classification performance. It is infeasible to apply an exhaustive search procedure for all possible combinations of feature subsets to derive the optimal feature subset because of the high dimensionality of the feature vector. Instead, the Adaptive Sequential Forward Floating Selection (ASFFS) search algorithm is employed to reduce the number of features.The feature subset selection method is purely based on an evaluation procedure that examines the discriminativeness of each feature or set of features in order to construct the best subset of features for the recognition process. We describe a validation-based evaluation criterion to choose the subset of features that would minimise the classification errors and ensure good inter-class separability between the different classes. As opposed to the voting paradigm used by the *KNN*, the evaluation criterion employs coefficients w that signify the importance of most nearest neighbours of the same class [11]. The probability score for a candidate s_c to belong to a cluster c is expressed in the following Equation (7):

$$f(s_c) = \frac{\sum_{i=1}^{N_c-1} z_i w_i}{\sum_{i=1}^{N_c-1} w_i} \quad (7)$$

where N_c is the number of instances within cluster c, and the coefficient w_i for the i^{th} nearest instance is inversely related to proximity as given:

$$w_i = (N_c - i)^2 \quad (8)$$

The value of z_i is defined as:

$$z_i = \begin{cases} 1 & \text{if } nearest(s_c, i) \in c \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

Such that the $nearest(s_c, i)$ function gives the i^{th} nearest instance to the instance s_c. The Euclidean distance metric is used to deduce the nearest neighbours from the same class. The significance for a subset of features is based on the validation-based metric which is computed 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 instances x as expressed the following equation:

$$Action = \arg\max_{S \in F} \left(\frac{\sum_{x=1}^{N} f_S(x)}{N} \right) \quad (10)$$

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 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 activities. Figure (2) In this study, we manually collected a dataset containing 241 video

sequences for 19 different basic actions by decomposing an activity into primitive actions. Each video consists of 15 frames which are all checked to better describe the complete action. The list actions include : jumping, walking, running, siding and skipping from right to left (RTL) and vice versa. Further, the activities of waving one hand, both hands, bending and jacking are split into two basic actions relating to the upward and downward portions of the motion.

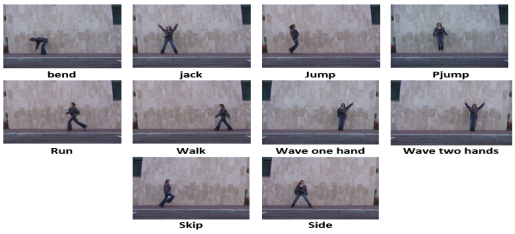

Figure 2. Weizmann dataset

After running the feature selection procedure on the obtained raw features, an optimal action signature is derived containing 648 features. The Correct Classification Rate is estimated using the K-nearest neighbour (KNN) classifier with $k = 3$ using the leave-one-out cross-validation rule. The KNN rule is applied at the classification phase due to its simplicity and therefore fast computation besides the ease of comparison to other existing methods. Using the Cumulative Match Score (CMS) evaluation method which was introduced by Phillips in the FERET protocol, we have correctly classified 95.02% of the 20 basic actions at rank $R = 1$ and 100% at rank $R = 9$. Figure (3) shows the CMS curve for the classification process. The achieved results promising because the recognition is based purely on local motion information and this can be boosted through adding global features.

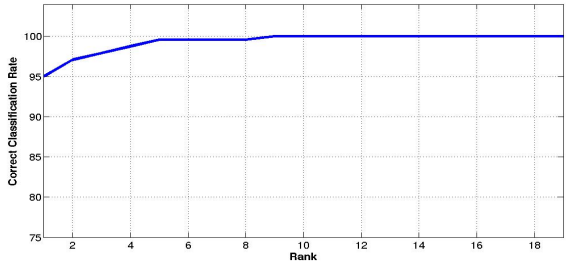

Figure 3. Cumulative Match Score for action recognition

The confusion matrix is shown in Figure (4) which visualizes the separation results across the different classes. The lighter squares reflect higher separation score and therefore higher discriminability. The dark blue diagonal line reflects the zero distance between the same class. The separation distance between the different clusters is computed using the Euclidean distance metric

V. Conclusions

The recognition of human activity is of prime importance for various applications as automated visual surveillance, sports analysis and automated refereeing. In this research study, a motion interchange descriptor is 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.

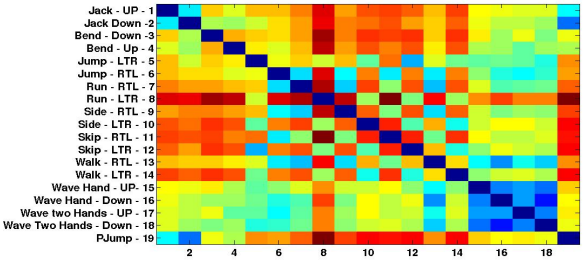

Figure 4. Confusion matrix for cross-matching of action recognition

References

[1] I. Bouchrika, J. N. Carter, M. S. Nixon, R. Mörzinger, and G. Thallinger, "Using gait features for improving walking people detection," in *20th International Conference on Pattern Recognition*, 2010, pp. 3097–3100.

[2] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *Communications Surveys & Tutorials*, vol. 15, no. 3, 2013.

[3] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," vol. 7, no. 1, 2013.

[4] O. Oshin, A. Gilbert, and R. Bowden, "Capturing relative motion and finding modes for action recognition in the wild," *CVIU*, vol. 125, 2014.

[5] R. Poppe, "A survey on vision-based human action recognition," *IVC*, vol. 28, no. 6, 2010.

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

[7] D. Weinland and E. Boyer, "Action recognition using exemplar-based embedding," in *CVPR 2008.*

[8] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *PAMI, IEEE Transactions on*, vol. 32, no. 2, 2010.

[9] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *ICCV*, 2009.

[10] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *ECCV*, 2012.

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