Vision-Based Human Activity Recognition Using LDCRFs

Mahmoud Elmezain[1,2] and Ayoub Al-Hamadi[3]
[1]Faculty of Science and Computer Engineering, Taibah University, KSA
[2]Computer Science Division, Faculty of Science, Tanta University, Egypt
[3]Institute of Information Technology and Communications, Otto-Von-Guericke-University, Germany

Abstract: *In this paper, an innovative approach for human activity relies on affine-invariant shape descriptors and motion flow is proposed. The first phase of this approach is to employ the modelling background that uses an adaptive Gaussian mixture to distinguish moving foregrounds from their moving cast shadows. Accordingly, the extracted features are derived from 3D spatio-temporal action volume like elliptic Fourier, Zernike moments, mass center and optical flow. Finally, the discriminative model of Latent-dynamic Conditional Random Fields (LCDRFs) performs the training and testing action processes using the combined features that conforms vigorous view-invariant task. Our experiment on an action Weizmann dataset demonstrates that the proposed approach is robust and more efficient to problematic phenomena than previously reported. It also can take place with no sacrificing real-time performance for many practical action applications.*

Keywords: *Action recognition, Invariant elliptic fourier, Invariant zernike moments, latent-dynamic conditional random fields.*

Received August 15, 2015; accepted January 11, 2016

1. Introduction

Human activities that are automatically recognizing from video sequences are still need to be studied consequent to their major prospects for many implementations in numerous domain and situation [13]. For human action recognition, there is a probably one common field so-called Human Computer Interaction (HCI), in which there is no available actions as mouse and keystrokes to get the user input. Throughout the state-of-the-art, the human activity can be recognized and classified using diverse graphic twines like motion [4] and shape [14]. Several frameworks characterize the action as a sequence of emissions (i.e., observations) matrices, which represent the extracted feature vectors from video data. Thus, the activity of human can be recognized via searching for such sequence [18]. In [3], the authors explore an approach for representing and recognizing the actions of human, where the main idea was based on capturing the silhouettes of both shape and motion as temporal templates. Here, 2D images of motion history and motion energy instead of upholding the 3D volume (i.e., spatio temporal) are used as an action templates for classification. Using space-time shapes, Gorelick *et. al.* [7] introduces an action model to represent shape, which obtained from the detection of silhouette information with respect to background subtraction. Numerous features such as local saliency, angle and shape are extracted by the properties of Poisson's equation [11]. Figure 1 demonstrations three examples of these shapes that appeared in [7]. In [17], the authors propose a framework, which extracts the spatio temporal features in various scales to cluster and isolate the human actions. So, they analysed and scaled the video volumes temporally to interact with the speed disparities of human activities. In addition, the local intensity gradients are calculated and then normalized for every point through the 3D volume.

Figure 1. Space-time volume depended on silhouette information.

Shectman and Irani [16] introduced a framework to compute the human motion flows to recognize the 3D volume correlation, which spot correspondences amongst image segments. Ahmad and Lee proposed an approach to recognize human action from multivites videos sequence, which use the integration between shape and the information of motion flow with variability investigation [1]. In this method, the features vector of shape flow in addition to local global optic are united as a multi-dimensional Hidden Markov Model (HMM) set for modelling the human activity. The key contribution of our framework is to motivate and recognize human action relied on the descriptor of affine

invariant shape as an elliptic Fourier, Zernike moments, optic flow and mass center for wave features. Using 3D spatio temporal volumes, the extracted features are positioned to the discriminative model of LDCRFs for classifying the human activity in video sequences.

Our experiments on standard benchmark action Weizmann dataset are carried out and show that the proposed approach yields promising results than previously reported anywhere the literature with no sacrificing real-time execution.

2. Proposed Methodology

In this section, the main steps of human activity throughout the proposed framework are explored in the next subsections (Figure 2).

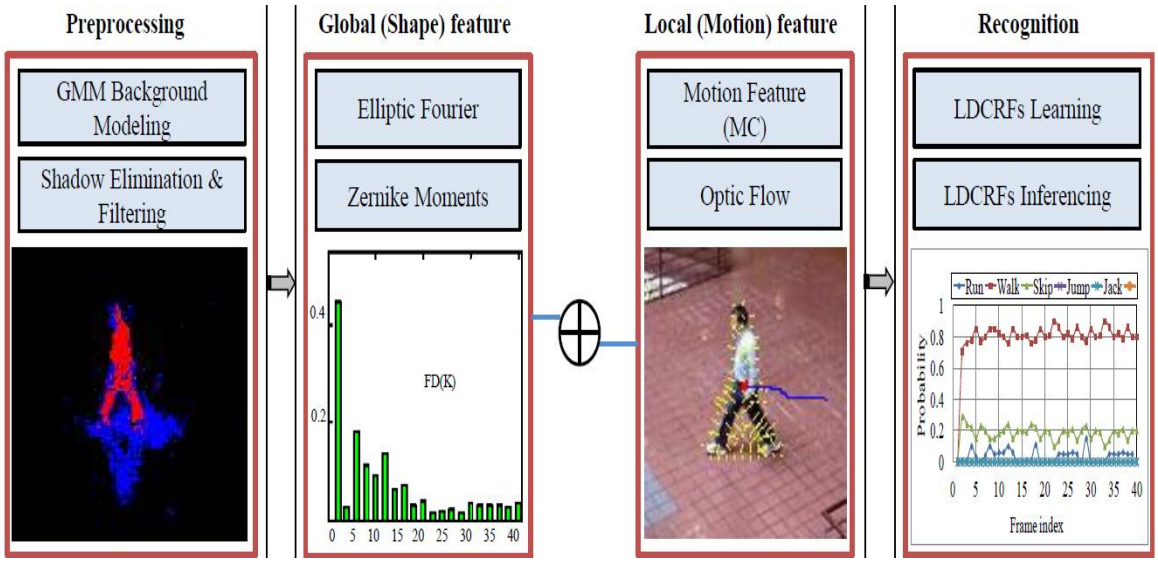


Figure 2. Concept of the action recognition approach.

2.1. Preprocessing

In this stage, an approach for modelling the background is proposed to discriminate the moving objects (i.e., foreground) from affecting cast shadows in image sequences. Briefly, these are detailed below.

2.1.1. Background Subtraction

The procedure of background subtraction is broadly used to detect the infrequent motion along scene. It is value stating that the Gaussian Mixture Model (GMM) is one instance of greater label for density model, which has numerous functions like preservative elements [2]. Properly talking, suppose that x refers to a pixel in current image frame j and M refers to Gaussian distributions number. Every pixel can be measured using a mixture of Gaussian M separately as next;

$$p(x) = \sum_{j=1}^{M} p(x \mid j).p(j) \quad (1)$$

To decide the number of Gaussian components, a method is used to observe the histogram of the dataset in which the selection of M is based on the number of peaks for this histogram. Based on the principles of maximizing likelihood function, a constructive algorithm is used in order to obtain the number of Gaussian elements [10]. In Equation (1), $p(j)$ represents the prior probability of j^{th} component. It is also called weighting function, which is generated from the component j of the mixture. $p(x|j)$ refers to Gaussian density of j^{th} element (Equation 2).

$$p(x \mid i) = \frac{1}{(2\pi)^{f/2}\sqrt{|\Sigma_i|}} . e^{-\frac{1}{2}.(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)} \quad (2)$$

where μ_j and Σ_j represent the mean and the covariance of j^{th} component, respectively. Additionally, f refers to the feature space. After deciding the number of component's M, the parameters of the mean, covariance and the prior probability for each component are calculated from given dataset. EM algorithm [10] is considered to compute these parameters and then they optimized relied on the minimization error of function E.

$$E = -\sum_{n=1}^{N} \ln\left(\sum_{j=1}^{M} p(x_n \mid j).p(j)\right) \quad (3)$$

Here, N refers to the data points number x_n. Using threshold γ with 0.5, it is being noted that the background distribution be on the topmost with the lowermost change. So, each pixel with no of component is marked as foreground. Figure 3 illustrates the estimation results of background with M=5

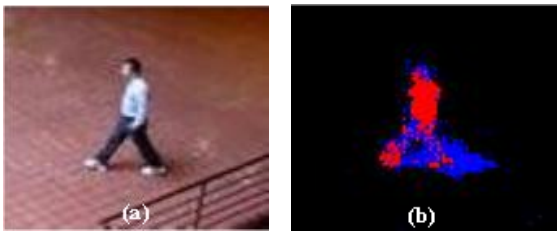
2.1.2. Shadow Elimination and Filtering

Designing a color model that separates the brightness from the chromaticity component plays a significant role in shadow elimination from images [1]. Generally, the background value $B_t = [\mu_r, \mu_g, \mu_b]$ in RGB color space is

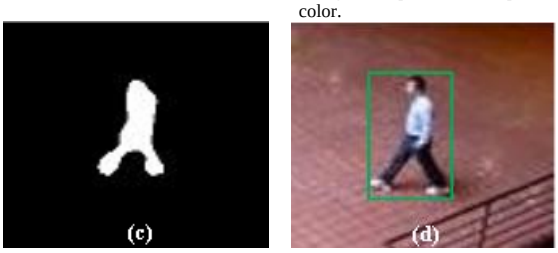
constituted by the brightness α and chromaticity distortion *C*. Furthermore, for a given pixel in each subsequent frame F_t, such that $C_r = F_r(t) - \alpha_t\mu_r$, $C_g = F_g(t)-\alpha_t\mu_g$ and $C_b = F_b(t)-\alpha_t\mu_b$; the brightness α_t and chromaticity C_t distortions are estimated from background model as follows;

$$\alpha_t = \frac{F_r(t)\frac{\mu_r}{\sigma_r^2} + F_g(t)\frac{\mu_g}{\sigma_g^2} + F_b(t)\frac{\mu_b}{\sigma_b^2}}{(\mu_r/\sigma_r)^2 + (\mu_g/\sigma_g)^2 + (\mu_b/\sigma_b)^2} \quad (4)$$

$$C_t = \sqrt{\left(\frac{C_r}{\sigma_r}\right)^2 + \left(\frac{C_g}{\sigma_g}\right)^2 + \left(\frac{C_b}{\sigma_b}\right)^2} \quad (5)$$



a) Source input foreground. b) Output from our background subtraction where the foreground pixels are overlaid by red, the shadows are overlaid by blue and the ordinary background pixels are kept as black color.



c) Silhouette detection after eliminating shadow. d) Foreground detection.

Figure 3. A sequence of person moving in indoor scene.

Where the chromaticity distortion C_t represents the perpendicular vector length between a pixel value F_t and a line that joins the background value μ and zero intensity point. Within the training phase, the variation β of the chromaticity distortion is estimated by Equation (6), in turn lead to a normalized chromaticity distortion.

$$\beta = \sqrt{\frac{\sum_{t=1}^{N} C_t^2}{N}} \;; \quad \hat{C} = \frac{C_t}{\beta} \quad (6)$$

For a given snapshot, the threshold τ_C is selected based on the shadow detection rate. As a result, pixels are either classified as a labelled background, foreground or cast shadow. Further precisely, a pixel is classified as a labelled cast shadow when the next two conditions are satisfied;

$$\hat{C} < \tau_C \;\; ; \;\; \alpha_{\min} < \alpha_t < 1 \quad (7)$$

After the shadow is eliminated, there still exists some small regions and noise. To remove these outlier's erosion and dilation are employed in mixture to produce a desired effect of frame processing. Then, the foreground image is obtained using the median filter with size 5×5. After a labelled foreground frame is obtained, they are localized using a blob analysis function. An example with the result of foreground, shadow and background detection is illustrated in Figure 3.

2.2. Feature Extraction

In this framework, we used an assortment of global and local features to designate the moving parts of human body (i.e., segmented silhouettes of *f(x, y, t)*). According to the global feature (i.e., shape feature), the silhouette image is segmented using a diversity of invariant descriptors like Zernike moments and elliptic Fourier's descriptors. Furthermore, the local feature (i.e., local feature) of foreground frame is obtained by the action motion trajectory in conjunction optic flow motion. Thus, the result of feature matrix is given below;

$$Action_{features} = \begin{pmatrix} G_i^{t_s} & G_i^{t_{s+1}} & \dots & G_i^{t_e} \\ L_i^{t_s} & L_i^{t_{s+1}} & \dots & L_i^{t_e} \end{pmatrix} \quad (8)$$

Where G_i and L_i are to Global and Local features, respectively. The length of action feature represents a difference between the starting frame (t_s) and the ending frame (t_e), in which $G_i^{t_s} = g_1^{t_s}, g_2^{t_s}, \dots, g_{F_1}^{t_s}$ and $L_i^{t_s} = l_1^{t_s}, l_2^{t_s}, \dots, l_{F_2}^{t_s}$. As a result, the feature matrix of global feature and local feature at every frame is equal to F_1+F_2.

2.2.1. Global Feature

The shape flow that considered as global flow of silhouette is itemised using an elliptic Fourier descriptor in addition to Zernike moments $G_i=[C_{xk}, C_{yk}, z_{00}, z_{11}, z_{22}]^T$. This process is described according to the following points.

- *Elliptic Fourier descriptor*: In this work, the feature of action silhouettes is determined by the trigonometric form of curve's shape C_k. The representation of trigonometric is supplementary instinctual to implement. Referring to Equation (9), the elliptic coefficients are calculated using Equation (10).

$$C_k = \frac{1}{T}\int_0^T c(t)e^{-jk\omega t} \quad (9)$$

$$C_k = C_{xk} + jC_{yk} \quad (10)$$

In which

$$C_{xk} = \frac{1}{T}\int_0^T x(t)e^{-jk\omega t}, \; C_{yk} = \frac{1}{T}\int_0^T y(t)e^{-jk\omega t} \quad (11)$$

ω represents the fundament frequency with value $T/2\pi$. Here, T represents function period and the harmonic number is assigned with k. It being noted that, the

selection of elements guarantee that the curve description is invariant to shape translation, scaling and rotation. In addition, they are self-regulating based on the selection of contour start point. For additional specifics, the reader can reference to [12].

- *Zernike Moments*: the silhouette image invariance can be realized via Zernike moments that provide rotation invariant moments orthogonal set. Furthermore, translation invariance and scale can proceed by the moment normalization [1]. Mathematically, Equation (12) gives the complex Zernike moment Z_{pq} of image intensity $f(\rho, \theta)$.

$$Z_{pq} = \frac{p+1}{\lambda_N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(\frac{x}{a} + \bar{x}, \frac{y}{a} + \bar{y}, t) R_{pq}(\rho) e^{-jq\theta} \quad (12)$$

where p (i.e., order) represents positive integer value, λ_N refers to a normalization factor, whereas q (i.e., repetition) either positive or negative integer with respect to conditions $|q| \leq p$ and p-$|q|$ = even. Additionally, the function f is then normalized according to translation and scaling via the scale factored a as well as the centre of silhouette image $(\bar{x}, \bar{y})$. $R_{pq}(\rho)$ is to a redial polynomial [12]. As a result, the invariant of Zernike moment features within the features of geometric to shape scaling, rotation and translation which salient likeness to invariant moments. Hu is specified via $G_z=[z_{00}, z_{11}, z_{22}]$. Experimentally, it is being noted that the percentage errors for invariants are lower than 0.5%.

2.2.2. Local Feature

Here, the local feature is expressed by the motion flow of foregrounds and is specified using the gravity center and optic flow $L_i = [\vec{z}(t), \vec{v}_{op}]$;

- *Center of Silhouettes Motion (CM):* the using of motion information persuades us a fuze through global features in order to constitute a classifier of LDCRFs. The features of motion are extracted and relied on computing centroid $\vec{z}(t)$, which carries motion center. Furthermore, the feature $\vec{v}(t)$ that explaining the motion distribution is generally provided by;

$$\vec{v}(t) = \lim_{n \to \infty} \frac{\Delta \vec{z}(t)}{\Delta t} \quad (13)$$

Such that $\frac{1}{2}(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i)$ represents the spatial coordinate of $\vec{z}(t)$ in current frame with respect to the total number n of moving pixels. By using these features, it would be eligible to differentiate, i.e., among human actions such that the motion occurs along a comparatively considerable area (i.e., running). Additionally, the action is localised in a minimal region, such that the slight parts of human body are in activity (i.e., moving one or two hands).

- *Optic Flow*: The activity of body parts includes an optical flow velocity. We can say that the person who manage the action (i.e., hand waving), the motion can appear on the hand only. When the user carries out the walking action, the motion will squeeze the full body. Moreover, the pruning value of estimated flow seems to be a suspicion to truthful the flow field and tolerates the best estimation for motion. The pruning method of optic flow contains two phases, each of which relied on Euclidean length of optic flow vector to isolate appropriate from inappropriate flow vector [5]. For first phase, we remove each flow vector with relatively small or very large in magnitudes. So, we identify two thresholds; minimum and maximum to control and filter the flow vectors for that purpose. Momentarily talking, by the given two thresholds ρ_1 and ρ_2, the flow vector $\vec{v}_{op} = [x, y]^T$ is acknowledged as true when it verifies valid condition: $\rho_1 < \|\vec{v}_{op}\| < \rho_2$ in which $\|.\|$ refers to the flow vector magnitude in regard to Euclidean metric. Or else, it is supposed as a noisy flow element and then detached. In the second phase, the vector $\vec{v}_{op}$ is pickled as a true flow element when the distance between flow center and analysed vector does not overdo the definite threshold τ. Correctly, it is stated by;

$$\left\| \vec{v}_{op} - \vec{z} \right\| < \tau \quad (14)$$

Such that, $\vec{z}$ represents the centroid of human motion region. Experimentally, the best performance of pruning is verified with the setting values of ρ_1= 5, ρ_2= 20. In addition, the average of image frames height h and width w (i.e., $l = (w+h)/2$) is verified when the τ equal 25%.

2.3. Classification

In this stage, the activities of human are classified based on the label number of LDCRFs. In general, LDCRFs are treated as undirected graphical models, which sophisticated for labelling sequential data [6, 8]. Here, every label belongs to a specific human action (Figure 4). The models of LDCRFs are normally considered to classify un-segmented sequence since it includes a class state (label) for each observation. Additionally, they can capably model and inference the human action sequence within the testing and learning processes.

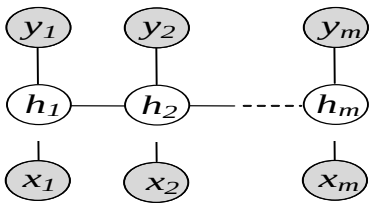


Figure 4. LDCRFs approaches, such that x_j is to the j^{th} equivalent emission value, h_j is to hidden states assigned to x_j . y_j refers to x_j label in which the grey circles signify the observed vector.

With respect to every sequence, we suppose that the vector of sub-structure with hidden variables is $h = h_1, h_2,\ldots, h_m$. So, the probability of state action y for a specified an emission sequence x is estimated by;

$$p(y \mid x,\theta) = \sum_{h:\forall h_i \in H_{yi}} p(h \mid x,\theta) \quad (15)$$

Where h_i represents one member of the set H_{yi} for the likely hidden states of the label action y_j

$$p(h \mid x,\theta) = \frac{1}{Z(x,\theta)} \exp\left(\sum_{i=1}^{n} F_\theta(h_{i-1}, h_i, x, i)\right) \quad (16)$$

Where parameter $\theta = (\lambda_1, \lambda_2, \ldots \lambda_{Nf}; \mu_1, \mu_2, \ldots \mu_{Ng})$, N_f is to the number of transition feature, N_g is to state feature function number and n refers to the length of emission action vector x. Thus, the function F_θ is realized by;

$$F_\theta(h_{i-1}, h_i, x, i) = \sum_f \lambda_f t_f(h_{i-1}, h_i, x, i) + \sum_g \mu_f s_f(h_i, x, i) \quad (17)$$

Where $t_f\ (h_{i-1}, h_i, x, i)$ represents a function of transition feature at location i-1 and i. Also, $s_g(h_i, x, i)$ represents a function of state feature at location i. Here, μ_g and λ_f are to the weight function of the state and transition features, respectively. Moreover, $Z(x,\theta)$ is a normalized factor in which it is computed by;

$$Z(x,\theta) = \sum_h \exp\left(\sum_{i=1}^{n} F_\theta(h_{i-1}, h_i, x, i)\right) \quad (18)$$

2.3.1. Training LDCRFs Model

The parameter $\theta = (\lambda_1, \lambda_2, \ldots \lambda_{Nf}; \mu_1, \mu_2, \ldots \mu_{Ng})$ is derived by the learning action data $D = \{x^{(i)}, y^{(i)}\}_{j=1}^{Td}$, in which $x^{(j)}$ represents an emission sequence of learning action set. $y^{(j)}$ is to correspondent label sequence of an emission sequence action $x^{(i)}$. Here, Td represents the learning sequence number. The objective function which maximizes log-likelihood to learn parameter θ, is estimated using Equation (19).

$$L(\theta) = \sum_{j=1}^{Td} \log p(y^j \mid x^j, \theta) = \sum_{j=1}^{Td} \left(\sum_{i=1}^{n} F_\theta(h_{i-1}^{(j)}, h_i^{(j)}, x^{(i)}, i) - \log Z(x^{(i)}, \theta)\right) \quad (19)$$

Furthermore, *BFGS* optimization technique of 300 iterations is to maximize the Likelihood using a gradient ascent technique for converging [8];

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{j=1}^{Td} \left(\sum_{i=1}^{n} \frac{\partial F_\theta(h_{i-1}^{(j)}, h_i^{(j)}, x^{(i)}, i)}{\partial \theta} - \sum_x p(h \mid x^{(i)}) \sum_{i=1}^{n} \frac{\partial F_\theta(h_{i-1}^{(j)}, h_i^{(j)}, x^{(i)}, i)}{\partial \theta}\right) \quad (20)$$

2.3.2. Inferencing LDCRFs Model

To obtain the probability $p(y|x,\theta)$ of the action labels y for the given specific emission features x [9], a matrix set is calculated. To facilitate some expressions, a start h_0 and stop h_{n+1} labels are added that so-called dummy labels.

Using the $p(h|x,\theta)$ which given using Equation (16) and for every location i in an emission sequence, $M_i(x)$ is built as a $|S \times S|$ matrix Equation (21);

$$M_i(h', h \mid x) = \exp(F_\theta(h', h, x, i)) \quad (21)$$

Such that $S = h_1, h_2, \ldots h_l$ represents a set of action label of training dataset. h' and h refer to the human action labels of S at time i. For that, the probability of action labels sequence y for given an emission sequence and model parameters is estimated by;

$$p(y \mid x,\theta) = \frac{\prod_i^{n+1} M_i(h_{i-1}, h_i, x)}{Z(x,\theta)} \quad (22)$$

Hence, the normalized $Z(x,\theta)$ represents the product access of this matrix:

$$Z(x,\theta) = \left(\prod_{i=1}^{n+1} M_i(i)\right)_{start,stop} \quad (23)$$

2. Experiments and Results

Weizmann dataset is one of the utmost generally used in human activity field [7]. This dataset is publicly available to the researchers in the action recognition field with no access charge. Furthermore, it includes ten various action categories so-called 'walk', 'jump forward on two legs' (i.e., jump), 'side', 'skip', 'jumping via two legs' (i.e., p-jump), 'run', 'wave one hand' (i.e., wave1), 'bend', 'jack', and 'wave two hands' (i.e., wave2). Every action is carried out using ten subjects by a static camera for given still background at the rate of 25FPS, and 180 × 144 spatial resolution with 24 bits per pixel. The captured sequences are very short time (i.e., each sequence only about a few seconds) (Figure 5-a). We split this action dataset into learning set with two thirds and testing set with one third to hoard balanced estimator of LDCRFs ability of the recognition process. We take in consideration the testing samples are completely different than training samples.

Table 1. The confusion matrix for per-frame classification with respect to weizmann dataset.

Action	Walk	Run	Jump	p-jump	Jack	side	Bend	Skip	Wave1	Wave2
Walk	0.93	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Run	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.00	0.08	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P-jump	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Jack	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Side	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.07	0.00	0.00
Bend	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	0.00	0.00	0.00	1.00	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

As a result, Table 1 brief in a confusion matrix for per frame classification, in which the valid echoes specify the main diagonal. Here, it is being notice that there is an obvious variance between leg and arm actions. The inaccuracies wherever confusions place is only between run action and walk human action.

In similar, it also happened between run and jump, and between sides and skip human actions. This is due to the rise proximity in each action paid (Table 2).

Table 2. Comparison with those previously reported on weizmann dataset.

Method	Recognition rate
Our Method	97.80%
Zhang *et al.* [18]	92.80%
Sadek *et al.* [15]	98.00%
Fathi and Mori [5]	100.00%
Niebles *et al.* [11]	90.00%

a) Running action due to their silhouette image, spatial boundary and corresponding elliptic Fourier Descriptors (FDs).

b) The optical flow result of walk and wave2 actions form left to right.

c) Temporal evolution of the probabilities of running action.

Figure 5. System outputs.

In the lighting of this differentiation, the proposed framework executes competitively with other previously reported as well as to proceed with no immolating real-time performance. But in Table 2, the mentioned methods in [5, 15] are more efficient than our method because the cost of used BFGS optimization technique of 300 iteration to verify the converging of learned parameters requires more time to reach the optimality. In future, LDCRFs will be employed with another gradient technique to alleviate this problem which in turn achieves promising results. Furthermore, our framework has realised a 97.80% accurateness per-frame classification. The outcome of the proposed framework is shown in Figure 5.

3. Conclusions

In this work, we propose a framework for human action recognition relied on an affine-invariant shape descriptor such as Zernike moments and elliptic Fourier, as well as to mass centre and optic flow features. These features are integrated and employed for LDCRFs to build actions models. Experiments on normal benchmark action Weizmann dataset showed that the projected framework can successfully classify per frame human action with 97.80% recognition rate.

References

[1] Ahmad M. and Lee S., "Human Action Recognition Using Shape and CLG-Motion Flow from Multi-View Image Sequences," *Pattern Recognition*, vol. 41, no. 7, pp. 2237-2252, 2008.

[2] Bar-Shalom Y. and Li X., *Estimation and Tracking: Principles, Techniques, and Software*, Artech House, 1993.

[3] Bobick A. and Davis J., "The Recognition of Human Movement Using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257-267, 2001.

[4] Efros A., Berg A., Mori G., and Malik J., "Recognizing Action at a Distance," *in Proceedings of 9th IEEE International Conference on Computer Vision*, Nice, pp. 726-733, 2003.

[5] Fathi A. and Mori G., "Action Recognition by Learning Mid-Level Motion Features," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, pp.

1-8, 2008.

[6] Ganapathy S., Vijayakumar P., Yogesh P., and Kannan A., "An Intelligent CRF Based Feature Selection for Effective Intrusion Detection," *The International Arab Journal of Information Technology*, vol. 13, no. 1, pp. 44-50, 2015.

[7] Gorelick L., Blank M., Shechtman E., Irani M., and Basri R., "Actions as Space Time Shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no.12, pp. 2247-2253, 2007.

[8] Lafferty J., McCallum A., and Pereira F., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *in Proceedings of the 8th International Conference on Machine Learning*, Williamstown, pp. 282-289, 2001.

[9] McCallum A., "Efficiently Inducing Features of Conditional Random Fields," *in Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, Acapulco, pp. 403-410, 2003.

[10] McKenna S., Raja Y., and Gong S., "Tracking Colour Objects Using Adaptive Mixture Models," *Image and Vision Computing*, vol. 17, no. 3, pp. 225-231, 1999.

[11] Niebles J., Wang H., and Fei-Fei L., "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Journal of Computer Vision*, vol. 79, no. 3, pp. 299-318, 2008.

[12] Nixon M. and Aguado A., *Feature Extraction and Image Processing*, Newnes, 2002.

[13] Poppe R., "A Survey on Vision-Based Human Action Recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976-990, 2010.

[14] Sadek S., Al-Hamadi A., Elmezain M., Michaelis B., and Sayed U., "Human Activity Recognition via Temporal Moment Invariants," *in Proceedings of 10th IEEE International Symposium on Signal Processing and Information Technology*, Luxor, pp. 79-84, 2010.

[15] Sadek S., AI-Hamadi A., Krell G., and Michaelis B., "Affine-Invariant Feature Extraction for Activity Recognition," *ISRN Machine Vision*, vol. 2013, pp.1-7, 2013.

[16] Shechtman E. and Irani M., "Space-Time Behavior Based Correlation," *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, pp. 405-412, 2005.

[17] Zelnik-Manor L. and Irani M., "Event-Based Analysis of Video," *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Kauai, pp. 1-8, 2001.

[18] Zhang Z., Hu Y., Chan S., and Chia L., "Motion Context: A new Representation for Human Action Recognition," *in Proceedings of European Conference on Computer Vision*, Marseille, pp. 817-829, 2008.

Mahmoud Elmezain was born in Egypt. Between1997 and 2004 he worked as demonstrator in Dept. of Statistic and Computer Science, Tanta University, Egypt. He received his Masters Degree in Computer Science from Helwan University, Egypt in 2004. He received PhD Degree in Computer Science from Institute for Electronics, Signal Processing and Communication sat Otto-von-Guericke-University of Magdeburg, Germany. His work focuses on image processing, pattern recognition, human-computer interaction and action recognition. Dr.-Ing. Elmezain is the author of more than 55 articles in peer-reviewed international journals and conferences.

Ayoub Al-Hamadi was born in Yemen. He obtained his master's degree and PhD degree from Otto-von-Guericke-University of Magdeburg, Germany between 1997 and 2001. He had been a Junior Research Group Leader in 2003 at Magdeburg University, Germany. He obtained a position of Junior Professor for Neuro Information Technology in 2008. He obtained the Habilitation degree in the fields of Image Processing, Artificial Intelligence and Pattern Recognition in 2010. He is the author of more than 325 papers in international conferences, international journals and books.