Salient Object based Action Recognition using Histogram of Changing Edge Orientation (HCEO)

Hnin Mya Aye
Image Processing Lab
University of Computer Studies, Mandalay
Mandalay, Myanmar
hninmyaaye26@gmail.com

Sai Maung Maung Zaw
Faculty of Computer System and Technology
University of Computer Studies, Mandalay
Mandalay, Myanmar
saisaimmz@gmail.com

***Abstract*—Action recognition has been a growing research topic in computer vision due to its great potentials for real-world applications. In this paper, we develop an effective action recognition approach based on salient object detection and propose a new feature descriptor to represent the changes of edge orientation. Firstly, we detect salient objects from each frame of a video sequence and generate edge maps for those detected salient objects. Then, we extract features on developed edge maps, using a combination of proposed Histogram of Changing Edge Orientation (HCEO) feature descriptor and existing Histogram of Optical Flow (HOF) feature descriptor. Finally, supervised multi-class support vector machine (SVM) classifier is used for recognizing various actions. The experiments were carried out on the standard UCF-Sports action dataset. As experimental results, our proposed action recognition approach is achieved with a significant improvement in recognition accuracy.**

Keywords—Action Recognition; HCEO; HOF; SVM

I. INTRODUCTION

In current years, action recognition has become an attractive research field in computer vision community because of increasingly demands from a variety of domains such as human-computer interfaces, video surveillance, entertainment environments and healthcare systems. Many researchers have allotted large eagerness on it, and accomplished outcomes were achieved gradually. However, action recognition becomes a complicated problem since there are different action classes, various action features and cluttered background. Although a number of progresses have been done in the recognition of actions, it still remains challenges making the recognition process to be immensely difficult. These challenges are illumination changes, camera motion, viewpoint variation, inter and intra class variations, etc [12].

The efficient representation of video information is primarily essential part in action recognition. There are chiefly two types of features for action representation: global features and local features. Actually, the features should be vigorousness to appearance variations, background motions, viewpoint changes, partial occlusions and action execution. In global representation, a preprocessing step is needed to identify the region of interest or detect the intended foreground object from the background. The common global representations are in the form of optical flow, silhouettes or edges. They are sensitive to partial occlusion and viewpoint changes. In local representation, the action is described as an assemblage of independent patches. Local features are quite invariant to changes in viewpoint, person appearance and partial occlusions [12] comparing with the global representation. Because of their benefit, interest points based local spatial-temporal features are more and more popular in action recognition [5, 30].

There are widely used feature detectors like Harris and DoG, in which the first one detects corner-like structures and the second one, detect mostly blobs. Corner detection can be considered as point detection where lines bend very deeply with high curvature. The more modern approaches are localized regions or patches of interest detection. Finally, we can also investigate an approach called optical flow describing motion, in which the bits which are fastest moving are the brightest points. All of these can serve a set of points, albeit points with various characteristics, but all are appropriate to assemble in the processing shape extraction. A square box moving is considered through a sequence of images. The edges are the perimeter of the box; the corners are the apices; the flow is how the box moves. All these can be grouped together to find the moving box [26].

As scrutinized outcomes of the action recognition problem in training and testing on distinct datasets, recognition researches endeavor to invent progressively generalized methods which are tough to intra-class variation and inter-class uncertainty. Thus, meaningful and discriminative feature extraction from video is becoming a challenging in processing of action recognition. Throughout history, several number of interest point detector were established such as Dollar interest points, space-time interest points and Hessian detectors, etc. The prominent idea was to conjecture local descriptors only at those salient locations and eliminate the rest of the locations [4]. To extract and describe powerful video information, a number of feature descriptors have been introduced such as Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) and Scale-Invariant Feature Transform (SIFT) descriptors [7]. For action recognition, various statistical models eg. HMM model, linear and nonlinear classification models eg. SVM and neuron networks eg. SOM [25] have been applied.

The goal of this paper is to present more efficient action recognition approach by using a combination of proposed Histogram of Changing Edge Orientation (HCEO) and existing Histogram of Optical Flow (HOF) feature descriptors. The

978-1-5090-5756-6/17/$31.00 ©2017 IEEE
SERA 2017, June 7-9, 2017, London, UK

main contribution is to introduce an efficient salient object detection based action recognition approach and to propose a new feature descriptor extracting discriminative changing edge orientation features. At first, we create edge maps for detected salient foreground objects from each video frame. And then, we extract discriminative features using a combination of proposed HCEO and existing HOF feature descriptors. The resulted features are utilized with supervised multi-class SVM classifier to accurately recognize actions.

The remaining parts of the paper are organized as follows. Section II reviews the methods related to our research work. The explanation of the main structure of our approach is presented in Section III. Section IV presents the results of our experiments conducted on standard UCF Sports dataset. Section V concludes the research work.

II. Related Work

In recent years, many various action recognition techniques have been introduced. Among these, the first category of techniques utilizes contour or silhouette information to represent an action [34, 39]. This type of technique always needs accurate segmentation that is hard to achieve, especially in real-world videos. The second one uses local spatio-temporal features [2, 16, 20, 23], which are capable to catch both visual and motion appearance. This category does not need object localization and additionally is robust to background clutter. Many various local interest point detectors such as Harris3D [16], Cuboid [30], Hessian [5] or Dense sampling [11] and numerous spatio-temporal descriptors such as HOG [17], HOF [17], Cuboid [30] or SURF [13] have been introduced. The third type includes procedures which analyzing motion trajectories [24, 33, 22, and 10] by tracking of feature points. Different types of trackers have also been established in the action recognition tasks such as the KLT tracker [31, 33], the SIFT tracker [18], and dense sampling tracker [9].

Moreover, the local spatiotemporal features with bag of visual words (BoVWs) framework have significant performance for action recognition. Li [29] described video as spatiotemporal features using BoVWs and then modeled actions by utilizing a probabilistic latent semantic analysis (pLSA) to confine and classify human actions. Wang et al. [35] recognized actions using a BoVW framework by a combination of several descriptors and also proposed a motion boundary histograms (MBH) descriptor. Bobick and Davis [1] introduced a combined approach using Motion Energy Images (MEI) and Motion History Images (MHI) to develop a temporal template that was applied for representing human actions. The MEI pointed out the location of motion in a video sequence, and the MHI guided the temporal history of motion.

Orrite et.al [6] developed an approach in which a set of MHIs, captured in various viewpoints, were be put to each action-specific SOM to make projection of the motion templates into a new subspace. A Maximum Likelihood classifier was employed over all action specific SOMs for action recognition. Huang et.al [36] also used SOM for human action sequences recognition. In which, action poses were mapped in a SOM and action sequences were represented as a trajectory of map units. Their approach showed that the potential of SOM on action recognition by achieving satisfying results.

Oikonomopouls et al. [3] introduced the concept of saliency region selection form spatial images to spatiotemporal video space. Salient points are detected by measuring the information content changes of the set of pixels in cylindrical spatiotemporal neighbourhoods at different scales. Ashwan Abdulmunem et al. [4] introduced an approach considering saliency guided feature. With saliency guidance, they extracted local and global features for encoding video information.

III. Main Structure of Action Recognition System

The proposed action recognition system mainly consists of four steps. The first step is salient object detection, in which the salient foreground objects are detected. Applying salient object detection makes reducing the number of feature descriptors, suppressing the background interference and also helps making the method to be more robust to background fluctuations. The second step is edge map generation to capture the possible edge appearances. The third step is feature extraction which is transforming of input data into distinguishing characteristics of input patterns that aid in recognizing among the categories of input patterns. Finally, multi-class support vector machine (SVM) classifier is employed for achieving action recognition. In this section, we will explain detailed descriptions of each step. Fig. 1 shows the main structure of the proposed action recognition system.

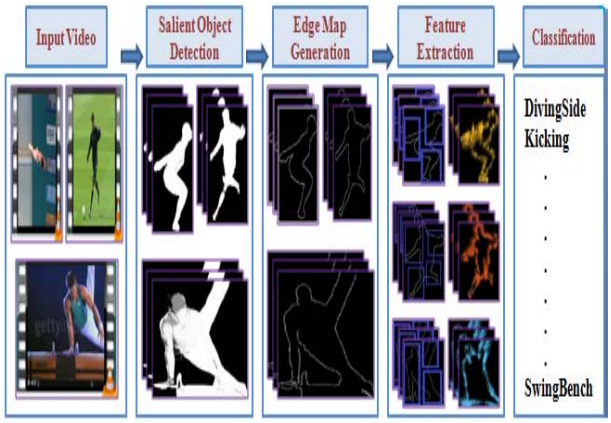


Fig. 1. Main structure of the proposed action recognition system (UCF Sports Dataset)

A. Salient object Detection

For salient object detection, we use inner and inter label propagation based detection algorithm proposed by Hongyang Li et al., [15]. This algorithm estimates saliency in an image by propagating the labels resulted from the most certain background and object regions extractions. To estimate the background appearance, the boundary cues are used because they are good indicators in distinguishing salient objects from the background. The objectness cues are also used to emphasize on the salient object characteristics. Fig. 2 shows

results of salient object detection for diving and swing-bench actions in UCF Sports dataset.

The affinity matrix construction is vital importance in the label propagation. It is constructed among superpixels by calculating the similarity of two image regions called superpixels (generated by SLIC algorithm [32]), in which image border regions become a set of boundary nodes, defined as B. The similarity is measured by a defined distance of the mean features in each region. The affinity entry w_{ij} of superpixel i (image region i) to a certain node j is defined as:

$$w_{ij} = \begin{cases} \exp\left(-\frac{D(f_i, f_j)}{\sigma^2}\right) & j \in N(i) \text{ or } i, j \in B \\ 0 & i=j \text{ or otherwise} \end{cases} \quad (1)$$

where f_i, f_j denotes the mean feature vectors of pixels inside node i,j, σ is a turning parameter to control strength of the similarity, N(i) indicates the set of the direct neighboring nodes of superpixel i. A degree matrix D = diag $\{d_1,...,d_N\}$ where $d_i = \sum_j w_{ij}$ is sum of the total entries of each node to other nodes. As an affinity matrix, the information of the background labels is propagated to predict saliency measure of other superpixels. Given a dataset $R = \{r_1, \ldots r_l, r_{l+1}, \ldots, r_N\} \in R^{D \times N}$, where the former l regions serve as query labels and D denotes the feature dimension, a function $V=[V(r_1), \ldots, V(r_N)]^T$ indicates the possibility of how similar each data point is to the labels. The similarity measure V (r_i) satisfies:

$$V_{t+1} = \sum_{j=1}^{N} a_{ij} V_t(r_j) \quad (2)$$

where a_{ij} is the affinity entry and t is the recursion step. For a given region, the similarity [37] is learned iteratively via the similarity measures propagation of its neighbors such that a region's final similarity to the labels is effectively affected by the features of its surroundings.

(a) (b)

Fig. 2. Results of salient object detection: (a) diving (b) swing-bench

In most cases, the inner propagation works well in determining saliency of superpixels by ranking similarities of superpixels to the boundary labels. But, in some complex scenes, using only boundary prior may lead to high saliency assignment to the background regions. To improve the results further, some foreground priors such as multi-scale saliency (MS), color-contrast (CC) and edge density (ED) are used. MS, introduced by [37], measures the uniqueness of objects according to the spectral residual of the image's FFT. CC [31] considers the distinct appearance of objects via a center-surround histogram of color distribution. ED computes the density of edges near window.

In some cases, the inner propagation via boundary labels alone has more accuracy results than a fusion of boundary and objectness labels due to the small interference of objectness measures near the salient object. So, a compactness score is evaluated to determine the quality of the regional saliency map. Only the score of saliency maps score lesser than a compactness measure will be adapted by the inter propagation via a co-transduction algorithm.

Thus, to ensure high quality of the saliency maps and improve the computational efficiency, a co-transduction algorithm is formulated to combine both boundary and objectness labels based on an inter propagation scheme. The inter propagation algorithm can distinguish the foreground finer from the background by making larger the set of boundary labels from objectness cues.

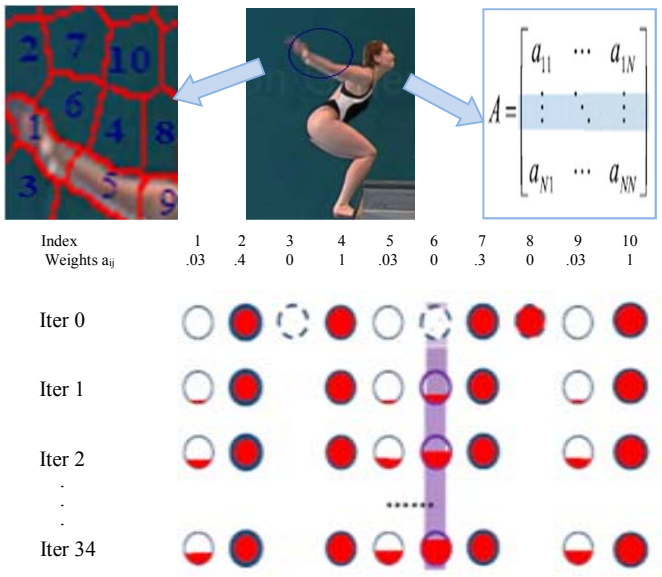


Fig. 3. An example of illustration how the inner propagation algorithm works

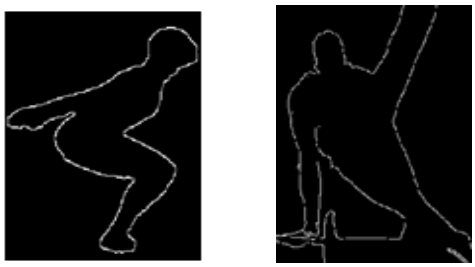
Fig. 3 shows how the inner propagation algorithm works. In which, one superpixel region 6 is investigated and it can be seen that how its value V(r) changes during each iteration. It is assumed that there has 10 regions and the dash-outline regions (3, 6, and 8) are not neighbours of region 6 and thus they are not thought in the propagation. The outline weight of each circle points out the weight of affinity. The red area within each circle deals with the value of V(r). Since region 2, 4, 7 and 10 are assumed as background labels, fill absolutely red colour within circles in each iteration.

B. Edge Map Generation

In our work, edge map generation is an essential step because edge features are good indicators of the salient foreground object orientation estimation. It is well known that edge features can change their appearance at different scales

due to blurring, so new edges can appear at different scales [19]. Therefore, it is important to create a good edge map to capture the possible edge appearances.

To model the edge map, we use canny edge detector. It was implemented with three main objectives: optimal detection with no spurious responses which reduce the response to noise, good localization with minimal distance between detected and true edge position, and single response which eliminate multiple responses to a single edge [26]. Fig.4 shows results of edge map generation for diving and swing-bench actions in UCF Sports dataset.



(a) (b)

Fig. 4. Results of edge map generation: (a) diving (b)swing-bench

This algorithm runs in five steps [40]:

1) *Smoothing:*
   It is unavoidable the image may contain some amount of noise. Therefore, the image is firstly smoothed by using a Gaussian filter to minimize those noises.
2) *Finding gradients:*
   This algorithm fundamentally locates edges where the grayscale image intensity changes the greatest, by determining gradients of the image. Gradients at each pixel in the smoothed image are determined by applying the Sobel-operator.
3) *Non-maximum suppression:*
   This step is converting the "blurred" edges in the image of the gradient magnitudes to "sharp" edges. It is done by keeping all local maxima in the gradient image, and removing everything else.
4) *Double thresholding:*
   In this step, edge pixels stronger than the high threshold are identified as strong; edge pixels weaker than the low threshold are suppressed and edge pixels between the two thresholds are identified as weak.
5) *Edge tracking by hysteresis:*
   Strong edges are elucidated as "certain edges", and can be involved in the final edge image. Weak edges are contained if and only if they are linking to strong edges. Final edges are determined by suppressing all edges that are not linking to a very certain (strong) edge.

C. Feature Extraction

To extract features, we use a combination of feature descriptors: motion based descriptor called Histogram of Optical Flow and appearance based descriptor called Histogram of Changing Edge Orientation.

1) Histogram of Optical Flow

HOF is based on extracting motion features from consecutive video frames using optical flow. The significant benefit of this method is that the heavy load of accurately approximation motion in changing lighting conditions and clutter is absolutely confined to optical flow calculation. The optical flow can be computed with a variety of ways, in our work, we calculate optical flow by using Farneback algorithm [8]. Optical flow computes the absolute motion, which captures location motion information and makes feature computation process very efficient. The histogram of optical flow field includes big information about motions in a video. Let $\theta_i\,(x, y)$be the direction of optical flow vector at the pixel (x, y) in frame I; and θ_i is defined by Eq. (3):

$$\theta_i\,(x, y) = arctan\frac{d_y}{d_x} \tag{3}$$

where d_x and d_y are the displacements in the x and y directions, respectively. These orientations in the human region are accumulated into a histogram, which is then normalized with the l_2–norm normalization form as in Eq. (4) through the length of frame L.

$$f = \frac{v}{\sqrt{||v||_2^2 + e^2}} \tag{4}$$

where f is normalized vector of v which is histogram vector, and e is a constant of exponential number.

The HOF vector value is calculated by Eq. (5), where j is the current frame, L is the number of frames, and f is HOF value of j-th frame. The dimension of the HOF in one frame is 8.

$$Desc_{HOF_j} = \sum_{j=1}^{L} f_j \tag{5}$$

2) Histogram of Changing Edge Orientation

In order to distinguish actions more accurately, we propose a new descriptor named HCEO to represent the appearance changes of salient object in each video frame. As shown in Fig. 5, the changes of edge orientations were highlighted on the blue regions which indicate much different in 'Diving' action and 'Swing-Bench' action. Therefore, the HCEO vector is an informative descriptor in predicting action labels for actions with orientation changes of body parts.

To compute the HCEO descriptor, edge maps are created for salient objects to accurately extract motion shape inside video frames and the x and y derivatives (I_x, I_y) of edge map I are computed. After calculating I_x and I_y, we compute orientation of edge maps as:

$$O = arctan\,\frac{I_y}{I_x} \tag{6}$$

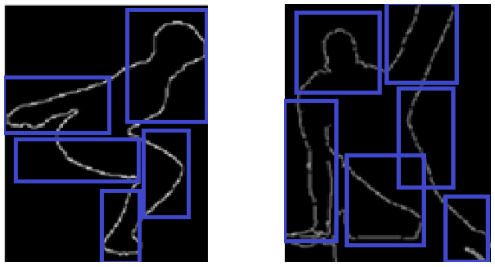
If the orientation of edge map for a given frame i is O_i, the change of edge orientation in each frame EO_i is calculated using the Eq. (7):

$$EO_i = |O_i - O_{i-1}| \quad (7)$$

These changes of edge orientations in the edge maps are accumulated into a histogram, which is then normalized with the l_2–norm normalization form as in Eq. (4) through the length of frame L. Then, HCEO vector value is calculated using the Eq. (8), where i is the current frame, L is the number of frames, and EO is HCEO value of i-th frame. The dimension of the HCEO in one frame is 8.

$$Desc_{HCEO_i} = \sum_{i=1}^{L} EO_i \quad (8)$$

The HCEO descriptor is an informative descriptor for recognizing actions in videos by representing the changing appearance in each edge map of a salient object inside video frame. Consequently, the HCEO vector serves meaningful information to describe human actions in a video, which is proven as an effective descriptor for increasing the recognition rate in action recognition problem.



(a) (b)

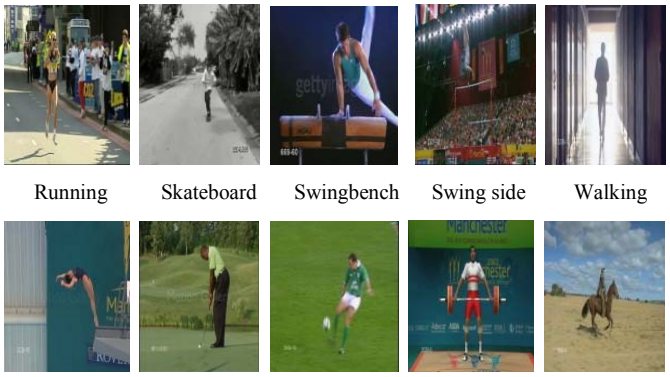
Fig. 5. An example of difference HCEO vector (changing edge orientation) of diving and swing-bench action

D. Classification

The final remaining step in the proposed action recognition pipeline is classification. There are many ways to assign a category of classifiers based on various properties. We use SVM introduced in [28], which is supervised learning models which analyze data and recognize patterns, is used for classification and regression analysis. A multi-class SVM model is utilized for training with the RBF kernel. Given a set of training examples, each is marked as belonging to one of the categories.

IV. Experimental Setup

In this section, we describe the dataset used and the experimental results computing recognition accuracy and processing time. To evaluate the performance, we carried out experiments on the UCF-Sports dataset [21,27]. It is one of the earliest action recognition datasets which includes realistic actions in unconstrained environment. It contains 10 sport actions which are diving, golf-swing, kicking, lifting, riding horse, running, skateboard, swinging-bench (on the pommel horse and floor), swinging-side (at the high bars), and walking, with a total of 150 videos. These videos have different frame rate and high image resolution. This dataset is one of the most challenging datasets because of having complicated background and large intra-class variations.



Running Skateboard Swingbench Swing side Walking

Diving Golf-swing Kicking Lifting Ridinghorse

Fig. 6 Sample frames from video sequences from UCF sports dataset

The experiments were conducted on UCF Sports dataset using multi-class SVM classifier with leave-one-out testing. We carried out three experiments: experiment on only proposed Histogram of Changing Edge Orientation (HCEO) feature descriptor and experiment on only existing Histogram of Optical Flow (HOF) feature descriptor, and experiment on combined HCEO and HOF feature descriptors.

A. Accuracy Evaluation

In the first experiment, the only proposed feature descriptor called HCEO was utilized to evaluate. The total results were achieved correct recognition accuracy rates of 74% with the proposed approach and 70% without salient object as shown in Table I in which S means salient object and NS means non-salient object.

TABLE I. Recognition Accuracy Results of HCEO

Action names	Recognition accuracy results						
	Total videos	Correctly labeled videos		Incorrectly labeled videos		Correct rate (%)	
		S	NS	S	NS	S	NS
Diving	14	12	10	2	4	85.7	71.4
Golf-Swing	18	13	12	5	6	72.2	66.7
Kicking	20	16	13	4	7	80.0	65.0
Lifting	6	4	4	2	2	66.7	66.7
Riding horse	12	7	8	5	4	58.3	66.7
Running	13	8	8	5	5	61.5	61.5
Skateboard	12	8	9	4	3	66.7	75.0
Swing-bench	20	18	17	2	3	90.0	85.0
Swing-side	13	10	11	3	2	76.9	84.6
Walking	22	15	13	7	9	68.2	59.1
Total result	150	111	105	39	45	74	70.0

Table II shows the confusion matrix results of HCEO based on salient object. The confusion matrix describes that 39 videos were wrongly recognized out of 150 videos. As shown in Table II, at most seven videos were confused in walking

action. In addition to, there was also the confusion relating with five videos occurred in golf-swing, riding horse and running actions, four videos occurred in kicking and skateboard actions, three videos occurred in swing-side action, two videos in diving, lifting and swing-bench actions.

TABLE II. CONFUSION MATRIX RESULTS OF HCEO ON SALIENT OBJECT

Action names	D	G	K	L	Rh	R	Sk	Sb	Ss	W
Diving	12	2	0	0	0	0	0	0	0	0
Golf-Swing	2	13	0	0	2	0	1	0	0	0
Kicking	0	0	16	0	0	2	1	0	0	1
Lifting	0	0	0	4	0	0	0	1	1	0
Riding horse	0	2	0	0	7	2	1	0	0	0
Running	0	0	1	0	2	8	0	0	0	2
Skateboard	0	0	0	0	0	2	8	0	2	0
Swing-bench	1	0	0	1	0	0	0	18	0	0
Swing-side	0	0	0	1	0	0	1	1	10	0
Walking	0	0	2	0	0	3	2	0	0	15

Table III shows the confusion matrix results of HCEO without salient object. The confusion matrix describes that 45 videos were wrongly recognized out of 150 videos. As shown in Table III, at most nine videos were confused in walking action. In addition to, there was also the confusion relating with seven videos occurred in kicking action, six videos in golf-swing action, five videos in running action, four videos in diving and riding horse actions, three videos in skateboard and swing-bench actions, two videos in lifting and swing-side actions.

TABLE III. CONFUSION MATRIX RESULTS OF HCEO ON NON-SALIENT OBJECT

Action names	D	G	K	L	Rh	R	Sk	Sb	Ss	W
Diving	10	2	0	0	0	0	2	0	0	0
Golf-Swing	2	12	0	0	1	1	2	0	0	0
Kicking	0	0	13	0	0	4	1	0	0	2
Lifting	0	0	0	4	0	0	0	1	1	0
Riding horse	0	1	0	0	8	1	2	0	0	0
Running	0	0	2	0	0	8	0	0	0	3
Skateboard	0	1	0	0	0	2	9	0	0	0
Swing-bench	1	0	0	0	0	0	0	17	2	0
Swing-side	0	0	0	0	0	0	0	2	11	0
Walking	0	1	3	0	0	4	1	0	0	13

In the second experiment, the only existing feature descriptor called HOF was used to evaluate. The total results were achieved correct recognition accuracy rates of 79.3% with the proposed approach and 74.7% without salient object as shown in Table IV.

Table V describes the confusion matrix results of only existing HOF descriptor based on salient object. It shows that 31 videos were incorrectly recognized out of 150 videos. As shown in Table V, five videos were mostly confused in riding horse action and walking action. And also, there was the confusion associating with four videos found in running action and skateboard action, three videos found in kicking action and golf-swing action, two videos found in lifting action and swing-bench action and swing-side action, and one video found in diving action.

TABLE IV. RECOGNITION ACCURACY RESULTS OF HOF

Action names	Recognition accuracy results						
	Total videos	Correctly labeled videos		Incorrectly labeled videos		Correct rate (%)	
		S	NS	S	NS	S	NS
Diving	14	13	10	1	4	92.9	71.4
Golf-Swing	18	15	13	3	5	83.3	72.2
Kicking	20	17	18	3	2	85.0	90.0
Lifting	6	4	4	2	2	66.7	66.7
Riding horse	12	7	9	5	3	58.3	75.0
Running	13	9	10	4	3	69.2	76.9
Skateboard	12	8	7	4	5	66.7	58.3
Swing-bench	20	18	16	2	4	90.0	80.0
Swing-side	13	11	10	2	3	84.6	76.9
Walking	22	17	15	5	7	77.3	68.2
Total result	150	119	112	31	38	79.3	74.7

TABLE V. CONFUSION MATRIX RESULTS OF HOF ON SALIENT OBJECT

Action names	D	G	K	L	Rh	R	Sk	Sb	Ss	W
Diving	13	1	0	0	0	0	0	0	0	0
Golf-Swing	1	15	0	0	0	1	1	0	0	0
Kicking	0	0	17	0	0	2	1	0	0	0
Lifting	0	0	0	4	0	0	0	0	2	0
Riding horse	0	2	0	0	7	2	1	0	0	0
Running	0	1	1	0	0	9	1	0	0	1
Skateboard	0	2	0	0	0	1	8	0	0	1
Swing-bench	0	0	0	0	0	0	0	18	2	0
Swing-side	1	0	0	0	0	0	0	1	11	0
Walking	0	0	2	0	0	2	1	0	0	17

Table VI describes the confusion matrix results of only existing HOF descriptor without salient object. The confusion matrix shows that 38 videos were incorrectly recognized out of 150 videos. As shown in Table VI, seven videos were mostly confused in walking action. And also, there was the confusion associating with five videos found in golf-swing and skateboard actions, four videos found in diving and swing-bench actions, three videos found in riding horse, running and swing-side actions, two videos found in lifting and kicking actions.

TABLE VI. CONFUSION MATRIX RESULTS OF HOF ON NON-SALIENT OBJECT

Action names	D	G	K	L	Rh	R	Sk	Sb	Ss	W
Diving	10	2	0	0	0	0	2	0	0	0
Golf-Swing	1	13	0	0	0	2	2	0	0	0
Kicking	0	0	18	0	0	2	0	0	0	0
Lifting	0	1	0	4	0	0	0	0	1	0
Riding horse	0	1	0	0	9	0	2	0	0	0
Running	0	0	1	0	0	10	1	0	0	1
Skateboard	0	2	0	0	0	3	7	0	0	0
Swing-bench	0	0	0	0	1	0	1	16	2	0
Swing-side	0	0	0	1	0	0	0	2	10	0
Walking	0	0	2	0	0	4	1	0	0	15

In the third experiment, the two combined feature descriptors called proposed HCEO and existing HOF were employed. The total results were achieved correct recognition accuracy rates of 86% with the proposed approach and 77.3% without salient object as shown in Table VII. According to results, our proposed action recognition approach achieved a significant improvement in recognition accuracy.

TABLE VII. RECOGNITION ACCURACY RESULTS OF COMBINED HCEO AND HOF

Action names	Recognition accuracy results						
	Total videos	Correctly labeled videos		Incorrectly labeled videos		Correct rate	
		S	NS	S	NS	S	NS
Diving	14	13	11	1	3	92.9	78.6
Golf-Swing	18	16	15	2	3	88.9	83.3
Kicking	20	18	16	2	4	90.0	80.0
Lifting	6	5	4	1	2	83.3	66.7
Riding horse	12	8	8	4	4	66.7	66.7
Running	13	11	9	2	4	84.6	69.2
Skateboard	12	9	8	3	4	66.7	66.7
Swing-bench	20	19	17	1	3	90.0	85.0
Swing-side	13	12	12	1	1	92.3	92.3
Walking	22	18	16	4	6	81.8	72.7
Total result	150	129	116	21	34	86.0	77.3

The confusion matrix for two combined feature descriptors called proposed HCEO and existing HOF based on salient object is shown in Table VIII. The confusion matrix shows that 21 videos were not correctly recognized out of 150. As shown in Table VIII, the confusion of four videos occurred in the following actions: riding horse and walking. There was also confusion concerning with three videos for skateboard action, two videos for golf-swing, kicking and running actions, as well as confusion of one video for diving, lifting, swing-bench and swing-side actions.

TABLE VIII. CONFUSION MATRIX RESULTS OF COMBINED HCEO AND HOF ON SALIENT OBJECT

Action names	D	G	K	L	Rh	R	Sk	Sb	Ss	W
Diving	13	0	0	1	0	0	0	0	0	0
Golf-Swing	1	16	0	0	1	0	0	0	0	0
Kicking	0	0	18	0	0	2	0	0	0	0
Lifting	0	0	0	5	0	0	0	0	1	0
Riding horse	0	1	0	0	8	1	2	0	0	0
Running	0	0	1	0	0	11	0	0	0	1
Skateboard	0	1	0	0	0	2	9	0	0	0
Swing-bench	0	0	0	0	0	0	0	19	1	0
Swing-side	0	0	0	1	0	0	0	0	12	0
Walking	0	0	1	0	0	3	0	0	0	18

The confusion matrix for for two combined feature descriptors called proposed HCEO and existing HOF without salient object is shown in Table IX. The confusion matrix shows that 34 videos were not correctly recognized out of 150. As shown in Table IX, the confusion of six videos occurred in the following action: walking. There was also confusion concerning with four videos for kicking, riding horse, running and skateboard actions, three videos for diving, golf-swing, and, swing-bench actions, two videos for lifting action, as well as confusion of one video for swing-side action.

TABLE IX. CONFUSION MATRIX RESULTS OF COMBINED HCEO AND HOF ON NON-SALIENT OBJECT

Action names	D	G	K	L	Rh	R	Sk	Sb	Ss	W
Diving	11	1	0	0	0	0	2	0	0	0
Golf-Swing	0	15	0	0	0	1	2	0	0	0
Kicking	0	0	16	0	0	3	1	0	0	0
Lifting	0	1	0	4	0	0	0	0	1	0
Riding horse	0	1	0	0	8	1	2	0	0	0
Running	0	0	1	0	1	9	0	0	0	2
Skateboard	0	1	0	0	2	1	8	0	0	0
Swing-bench	0	0	0	0	0	0	1	17	2	0
Swing-side	1	0	0	0	0	0	0	0	12	0
Walking	0	0	2	0	0	4	0	0	0	16

B. *Processing Time*

As a consideration of computational cost in non-salient object based action recognition scheme, HCEO takes 0.18 seconds and HOF takes 0.47 seconds for each frame of a video sequence. In salient object based scheme, HCEO takes 0.03 seconds and HOF takes 0.07 seconds for each video frame. Thus, salient object based feature extraction consumes lesser time and 6 - 7 times speedup is offered. But, if we consider as a total processing time by including preprocessing time, salient object based action recognition scheme requires more processing time in 1.4 - 2 times because of estimating time of observing foreground salient object in each video frame. Although our recognition approach takes more time, it can serve better recognition accuracy with a remarkable improvement. It is a trade-off between two desirable performances: better accuracy and lower processing time.

V. CONCLUSION

In this paper, we have presented an efficient action recognition approach based on salient object detection. The proposed salient object based action recognition approach contributes a way to reduce background interferences in processing and to make more robust recognition approach to background fluctuations. To achieve high accuracy in solving the action recognition problem, action representation method is very important. We used a new combination of the proposed HCEO and existing HOF feature descriptors to represent actions on a video frame. Because it is a combination of the changing appearance, and motion information of the video, it can provide meaningful and valuable information for recognizing various actions. According to experimental results, the proposed action recognition approach based on salient object detection achieved better recognition accuracy with a significant improvement than traditional action recognition approaches without salient object detection.

REFERENCES

[1] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.23, no.3, pp.257-267, Mar 2001.

[2] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients", In BMVC, 2008.

[3] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal Salient Points for Visual Recognition of Human Actions", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 36, No. 3, pp. 710–719, 2005.

[4] Ashwan Abdulmunem, Yu-Kun Lai, and Xianfang Sun, "Saliency guided local and global descriptors for effective action recognition", Computational Visual Meida, Vol.2, No.1, pp. 97-106, March 2016.

[5] G. Willems, T. Tuytelaars, and L. Van Gool, "An Efficient Dense and scale-invariant spatio-temporal interest point Detector", Proceedings of 10th European Conference on Computer Vision (ECCV2008), Marseille, France, pp. 650-663, 2008.

[6] C. Orrite, F. Martinez, E. Herrero, H. Ragheb, and S. Velastin, "Independent viewpoint silhouette-based human action modelling and recognition", in Proc. of the 1st International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA'08), 2008.

[7] Feng Shi, Robert Lagani`ere and Emil Petriu, "Gradient Boundary Histograms for Action Recognition", IEEE Winter Conference on Applications of Computer Vision, 2015.

[8] G. Farneback, "Two-frame motion estimation based on polynomial expansion", in Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA), Halmstad, Sweden, pp. 363-370, 2003.

[9] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition", International Journal of Computer Vision, vol. 103, no. 1, pp. 60-79,2013.

[10] H. Wang, A. Klaser, C. Schmid, and L. Cheng-Lin. "Action recognition by dense trajectories", In CVPR, 2011.

[11] H.Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition", In BMVC, 2009.

[12] Haiam A. Abdul-Azim, Elsayed E. Hemayed, "Human action recognition using trajectory-based representation, Egyptian Informatics Journal, pp.187–198,16 January 2015.

[13] Herbert Bay, Speeded-Up Robust Features (SURF), 10 September 2008.

[14] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded Up Robust Features", European Conference on Computer Vision (ECCV2008).

[15] Hongyang Li, Huchuan Lu, Zhe Lin, Xiaohui Shen, and Brian Price, "Inner and Inter Label Propagation: Salient Object Detection in the Wild", IEEE Transaction Image Processing, Vol. 24, No. 10, October 2015.

[16] I. Laptev. On space-time interest points. IJCV, 2005.

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies", In CVPR, 2008.

[18] J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, 2009, pp. 2004-2011.

[19] K. Mikolajczyk A. Zisserman C. Schmid, "Shape recognition with edge-based features".

[20] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition", In CVPR, 2009.

[21] Khurram Soomro and Amir R. Zamir, "Action Recognition in Realistic Sports Videos", Computer Vision in Sports, Springer International Publishing, 2014.

[22] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis", In ECCV, 2010.

[23] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match:Video structure comparison for recognition of complex human activities", In ICCV, 2009.

[24] M. Kaaniche and F. Bremond, "Recognizing gestures by learning local motion signatures of hog descriptors", TPAMI, 2012.

[25] Markus Hagenbuchnera,Ah Chung Tsoib, "A Supervised Training Algorithm for Self-Organizing Maps for Structures".

[26] Mark Nixon & Alberto Aguado, "Feature Extraction and Image Processing Second Edition".

[27] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah, "Action MACH: A Spatio-temporal Maximum Average Correlation High Filter for Action Recognition", Computer Vision and Pattern Recognition, 2008.

[28] N. Deng, Y. Tian, and C. Zhang, "Support Vector Machines-Optimization Based Theory, Algorithms, and Extensions", Boca Raton, FL: CRC Press, 2013.

[29] Niebles, J. C.; Li, F.-F, "A hierarchical model of shape and appearance for human action classification", In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2007.

[30] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features", Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72, (2005).

[31] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: action recognition through the motion analysis of tracked features", Proceedings of IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshop), Kyoto, Japan, pp. 514-521, 2009.

[32] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels", EPFL, Lausanne, Switzerland, Tech. Rep. 149300, Jun. 2010.

[33] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked key points", Proceedings of IEEE 12th International Conference on Computer Vision, Kyoto, Japan, , pp. 104-111, 2009.

[34] T.-S. Kim and Z. Uddin, "Silhouette-based Human Activity Recognition Using Independent Component Analysis, Linear Discriminant Analysis and Hidden Markov Model", InTech, 2010.

[35] Wang, H.; Klaser, A.; Schmid, C.; Liu, C.-L, "Dense trajectories and motion boundary descriptors for action recognition", International Journal of Computer Vision, Vol. 103, No. 1, 60–79, 2013.

[36] Wei Huang, Q. M. Jonathan Wu, "Human Action Recognition based on Self Organizing Map", IEEE International Conference on Acoustics, Speech and Signal Processing, 2010.

[37] X. Bai, X. Yang, L. J. Latecki, W. Liu, and Z. Tu, "Learning contextsensitive shape similarity by graph transduction," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 5, pp. 861–874, May 2010.

[38] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach", Proc. IEEE CVPR, Jun. 2007, pp. 1–8.

[39] Z. Lin, Z. Jiang, and L. S. Davis", "Recognizing actions by shape-motion prototype trees", In ICCV, 2009.

[40] http://watkins.cs.queensu.ca/~jstewart/457/notes/24-canny.pdf, March 23 2009.