International Journal of Research and Engineering
ISSN: 2348-7860 (O) | 2348-7852 (P) | Vol. 6 No. 2 | March 2019 | PP. 580 - 588
Digital Object Identifier DOI® http://dx.doi.org/10.21276/ijre.2019.6.2.2

Action Recognition Framework using Saliency Detection and Random Subspace Ensemble Classifier

ORIGINAL ARTICLE

Author(s): *[1]Hnin Mya Aye, [2]Sai Maung Maung Zaw
Affiliation(s): [1]Image Processing Lab, University of Computer Studies, Mandalay
[2]Faculty of Computer Systems and Technologies,
University of Computer Studies, Mandalay
*Corresponding Author: hninmyaaye26@gmail.com

***Abstract* - Action recognition can be defined as a problem to determine what kind of action is happening in a video. It is a process of matching the observation with the previously labelled samples and assigning label to that observation. In this paper, a framework of the action recognition system based on saliency detection and random subspace ensemble classifier, is introduced in order to increase the performance of the action recognition. The proposed action recognition framework can be partitioned into three main processing phases. The first processing phase is detecting salient foreground objects by considering pattern and color distinctness of a set of pixels in each video frame. In the second processing phase, changing gradient orientation features are used as a useful feature representation. The third processing phase is recognizing actions using random subspace ensemble classifier with discriminant learner. Experimental results are evaluated on the UIUC action dataset. The proposed action recognition framework achieved satisfying action recognition accuracy.**

***Keywords*:** *Action Recognition; Saliency Detection; Random Subspace Ensemble Classifier; UIUC Dataset*

I. INTRODUCTION

Due to the growing interest in human interaction and computation, research on action recognition and further kinds of research reflecting the intelligent interaction environment has been studied. It has been a growing research topic in computer vision due to progressively needs from a diversity of areas such as human-computer interfaces, video surveillance, sports video analysis entertainment environments and healthcare system [1]. Even though a number of evolutions have been achieved in action recognition, it still has challenges such as illumination fluctuations, camera motion, inter and intra class variations, etc. With successful action recognition, surveillance videos can command warnings to the public when predefined hazardous actions happen in the range of surveillance; human-computer interaction can be used in both medicine to save lives, and console video games to entertain people; and sports annotation system can execute complex player motion analysis and obtain play strategy information from live video of sport games in real-time.

Action is a sequence of movements of the human body, and may contain numerous portions of the body at the same time. Recognition of action is to match the observation with the defined early pattern and then give the action label [2]. For detecting and recognizing the different actions of people, the variety of the human body, appearance, posture and change of movement and lighting are difficult works in the area of research for recognizing human action. The same person will perform the same way differently at a separate time and different people will behave with the same actions. The same type of action can have massive variances in their visual appearance, variations in performing speed, clothing, and viewpoints [3].

Over the last era, spatial-temporal volume based holistic methods and local spatial-temporal feature representations have been succeeded good performance on some action datasets. But, these methods are until now difficult to extract the efficient visual information. Understanding human actions by tracking body parts, is also a solution following the way of human visual perception.

Action recognition is taken into consideration of a multi-class classification task in which each action type is an individual target class. In this work, it is aimed to improve the effectiveness of recognizing human actions by enhancing the classification phase. It can be argued that the discriminative ability of encoded information cannot be wholly used by single recognition methods. The weak point of single recognition methods comes to be more obvious when the complication of the recognition problem increases, mostly when having various action types and similarity of actions. So, an ensemble classification is used for improving the efficiency, compensating for an insufficiency in one learner. The experimental outcomes show that the recognition accuracy can significantly be enhanced [4].

II. RELATED WORK

Various action recognition approaches have been proposed and these approaches showed the significant progress towards action recognition in realistic and challenging videos. There are three main categories of the existing action recognition methodologies: Methods based on human body model, Holistic methods and Local feature methods.

A. Reviews of Human Body Model based Methods

The human body model based action recognition method extracts 2D or 3D information about the part of the human body, such as position and movement of the body part. In this approach, recognition of action is based on estimates of poses, parts of the human body, joint position trajectory or reference point.

Rohr, K. represented human body movements (walking action) through 3D models based on cylinders. The author used a kinematic modeling approach to represent the moving body and Kalman filter to estimate the model parameters in successive images [5]. Yilmaz, A. and Shah, M. have introduced human trajectory based action recognition method in which a video sequence captured by an uncalibrated non-stationary camera. To handle the movement of cameras and various homogeneous angles in a variety of environments, a variety of epipolar dynamic geometry scenes were modelled with the help of a temporal base matrix [6].

Ali, S., Basharat, A. and Shah, M. have also introduced an action recognition approach using the idea of chaotic theory for modeling and analyzing nonlinear dynamic schemes. Trajectories of reference joints were utilized as the illustration of the non-linear dynamical scheme producing the action. The reference joint points' trajectories were also used to rebuild phase space using an integrated scheme. To find optimum dimensions for embedding, the closest fake neighboring method was used to display the observed orbit from automated overlaps arising from the projection of dynamically dynamic systems to lower dimension space [7].

B. Reviews of Holistic Methods

Holistic methods represent actions using the appearance and movement of the entire human body without any detection and labeling of individual body parts. In general, this method can be classified into three fundamental categories: mask-based methods, methods based on optical flow form and template-based methods.

Typically, the shape mask based methods use silhouettes or contours of the whole human body from the image sequences. Wang, L. and Suter, D. proposed a transformation of a set of human silhouettes into two compact description forms, average motion energy (AME) and mean motion shape (MMS). The AME was computed with a set of moving binary silhouettes based on periodical detection of motions. The MMS was calculated with the single-connectivity binary silhouette applying an image boundary [8].

Methods based on optical flow form are not based on background segmentation. In Efros, A.A. et al. introduced motion descriptors focused on the optical flow to recognize human action. The flow fields were divided into four different channels related to positive and negative components, as well as horizontal and vertical components. The blurring process was followed to prevent the noisy shift [9].

Blank et al. built space-time shapes that contain spatial information on human poses at any time (torso and extreme location and orientation, different aspects of body parts), likewise dynamic information (body movements). This form of time space was generated by composing sequences of silhouettes, calculated by background rejection [10].

C. Reviews of Local Feature Methods

Local features catch the characteristic shape and information movement of local video regions. These features are generally extracted precisely from the video and, thus, escape from the failure of pre-processing approaches like motion segmentation or human detection.

Laptev and Lindeberg introduced space-time interest point Harris3D detector by expanding 2D Harris-Laplace detector. A spatiotemporal second-moment matrix at each spatio-temporal point with different dimensional and progressive time scale, a separable Gaussian smoothing function and space-time gradients were computed [11].

To produce more dense functional space points, Dollar et al., observed that sometimes angles of real-time space are rare, though interesting movements occur, and may be extremely sparse in convinced circumstances. Thus, the Gabor detector offering a more solid decision than Harris3D, was introduced. Gabor detectors used a group of Gaussian spatial kernels and temporal Gabor filters [12].

Messing, R., Pal, C. and Kautz, H. extracted feature trajectories using KLT tracker. For representing feature trajectories with variable length, the authors applied a uniform quantization in the log polar coordinates, along 8 bins for direction, and 5 for weight. By comparing SIFT descriptors between two successive frames, Sun et al. computed trajectories with a hierarchic structure to construct spatio-temporal contextual knowledge. Actions were classified with intra- and inter-trajectory statistics. The random sampling points were detected in the long-term trajectory region extracted through KLT tracker and SIFT descriptor match [13].

III. THE PROPOSED ACTION RECOGNITION FRAMEWORK

The proposed action recognition framework primarily involves three stages. The earliest stage is saliency detection, in which the salient foreground objects are detected. Employing saliency detection can serve to decrease the background intervention and also aid to create the technique to be stronger to background variations. The second step is feature extraction in which input data is transformed into distinctive features of input patterns which help in recognizing among the kinds of input patterns. Lastly, the random subspace ensemble classifier with discriminant learner, is used for succeeding action recognition.

A. *Saliency Detection*

The foremost phase in the proposed action recognition framework is the saliency (salient object) detection. In this paper, a saliency detection algorithm developed in [14], is used. In this algorithm, a salient object is taken into account consisting of pixels whose local neighborhood (region or patch) is distinct in both pattern and color.

Pattern distinctness is assessed by way of computing the inner statistics of the patches in the image. The distinct patch is identified by measuring the distance to the average patch. The length to the average patch is measured with the patch distribution in the image, by calculating the L_1 distance between the patch and the average patch in PCA coordinates.

To capture dominant variations among image patches, the principal components of each patch are pulled out as patch attributes based on PCA. As PCA is arithmetically identified as a technique for transforming correlated variables to linearly uncorrelated variables called principal components, it can produce great result for finding the preferred distinct patch [15]. To find principal components, n × n covariance matrix is initially constructed. Covariance is a measure to find out how much the pixel neighborhoods vary from the mean with respect to each other. After constructing n × n covariance matrix, eigenvalues and eigenvectors are calculated from the n × n covariance matrix. The eigenvectors of the covariance matrix make over the random vector into statistically uncorrelated random variables. These eigenvectors can provide information about the patterns in the data.

Colour distinctiveness is calculated by segmenting a video frame into regions using the SLIC superpixels to create the PCA basis and then approximating which region is distinctive in colour. SLIC (Simple Linear Iterating Clustering) is easy to use and understand. The colour distinctness of a region is found out by the summation of L_2 distances from all other regions in the colourspace.

To compute the salient region which are distinct in both color and pattern, it is basically calculated the product of the pattern and color distinctness maps. As objects have a tendency to be in the center of the frame, a Gaussian map encompassing the centre of the frame is also made. The last saliency space map is the product of the colour distinctness map, patch distinctness map, and the Gaussian map.

B. *Feature Extraction*

The second phase in the proposed action recognition framework is the extraction of features. It is the procedure of producing the information from the raw input data that is most relevant for discrimination between the classes.

To make a distinction of actions more precisely, changing gradient orientation feature is used for representing the appearance variations of salient object in each video frame. It is an informative descriptor in predicting action labels for actions with orientation changes of body parts. For computing the feature descriptor, the x and y derivatives of salient objects in each video frame are computed. After computing derivatives, orientations of salient objects are calculated and changes of orientations in each frame are computed. These orientation changes are added for all frames in a video sequence and then the orientation histogram is normalized. This feature vector gives out important information for describing human actions in a video [16].

C. *Classification*

In this system, random subspace ensemble classifier is used to recognition actions. It is a type of ensemble classifiers that involve many classifiers in a subspace of data feature space. Classification outcomes are based on these individual classifiers result by majority voting. When the total of training objects is quite smaller than the data dimension, building classifiers in random subspaces can settle the lesser training size problem. The subspace dimension is smaller than in the original feature space, while the number of training objects remains the same. Thus, the relative training sample size rises.

When data have several redundant features, one may take superior classifiers in random subspaces than in the original feature space. The combined result of such classifiers may be better to a single classifier created on the original training set in the whole feature space [17]. As base classification learners, discriminant analysis (learner) is used. Discriminant analysis is a classification problem in which two or more groups or clusters or populations are identified a priori and one or more new observations are categorized into one of the known populations based on the measured characteristics.

Random subspace models data from the original feature set and constructs one base classifier on each subset. The ensemble gives a class label by either majority voting or averaging of output probabilities. Let f={x_1,..,x_n} be the feature set with n dimensionality. For constructing a random subspace ensemble with L classifiers, L samples are collected with each of size M, drawn without replacement from a uniform distribution over X. Each feature subset describes a subspace of X of cardinality M, and a classifier is trained by base classifiers like support vector machine, k-nearest neighbor and discriminant analysis [18].

IV. System Evaluation

This section presents the evaluation of the proposed action recognition framework. The characteristics of the dataset and parameters of classifiers are explained. The evaluation of salient object detection is described for evaluating the detected salient object is completely detectable or not. The evaluation of classifier performance is also described.

A. *Experimental Setup*

The experiments are conducted on UIUC action dataset. The University of Illinois at Urbana-Champaign (UIUC) created the UIUC Action Dataset [19]. The dataset consists of 14 actions: walking, running, jumping, waving, jumping jacks, clapping, jumping from sit-up, raise one hand, stretching out, turning, sitting to standing, crawling, pushing up and standing to sitting. Fig 1 describes sample frames of each action class in UIUC dataset. In this paper, total number of observations is 154 sequences and each action class contains eleven action sequences performed by three actors.

The classification accuracy is evaluated using SVM, KNN and Ensemble classifiers with 10-fold cross-validation scheme. Cross-Validation is a statistical technique of estimating and matching learning algorithms by separating data into two parts: one part is used for learning or training a model and the other part is used for validating the model. The classifier is executed in twenty times and the recognition accuracy of the proposed action recognition system is calculated by averaging accuracy results of each execution round. The sensitivity and specificity are also computed by averaging sensitivity and specificity results of each execution round of classification.

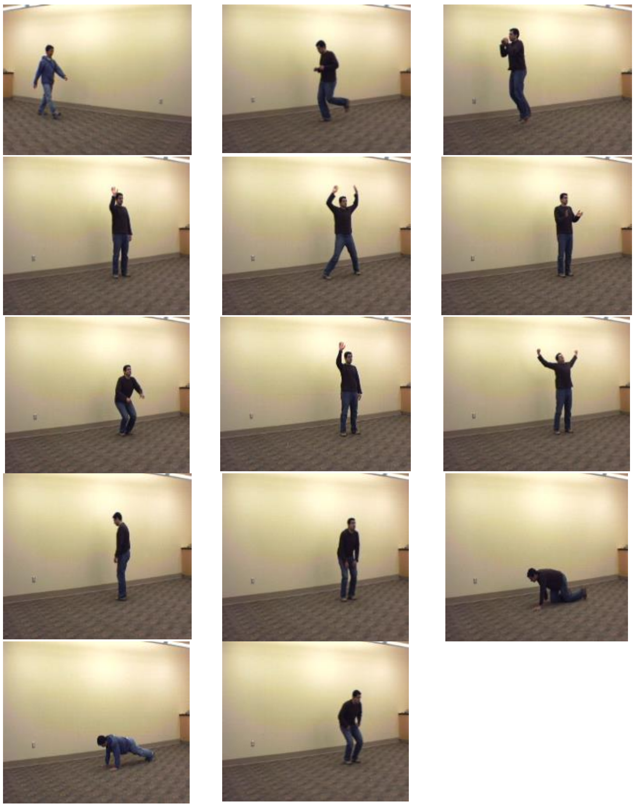


Fig. 1 Sample frames of each action class in UIUC action dataset.

B. *Evaluation of Salient Object*

To evaluate the quality of salient objects detected from each video frame, an evaluation metric called Area under the ROC curve (AUC) is used. AUC measures how well the salient object of an image can be predicted by comparing with the ground truth human fixations on the image. The perfect prediction score corresponds to an AUC score of value 1. With AUC, human fixations are considered as the positive set and some points from the image are sampled to form the negative set. TABLE I shows average AUC score of each action class in UIUC dataset. The average AUC score of all action classes is 0.96.

TABLE I. Avergae AUC Score of each action class

Action Class	Average AUC Score
Clap	0.96
Crawl	0.98
Jump	0.95
Jump from Sit	0.97
Jumping Jack	0.95
Push up	0.98
Raise one hand	0.95
Run	0.94
Sit to Stand	0.97
Stand to Sit	0.97
Stretch out	0.94
Turn	0.95
Walk	0.94
Wave	0.96

D. Evaluation of Classifier Performance

There are assessment methods to evaluate the classifier performance. Classifiers are generally assessed using evaluation metrics, such as accuracy. Classification metrics are computed from true positives (TPs), false positives (FPs), false negatives (FNs) and true negatives (TNs), all of which are tabularized in the confusion matrix.

Accuracy can be identified as a relation between the accurately classified samples to the whole number of samples. A valued measure for understanding FNs is sensitivity (also named recall, hit rate or the true positive rate), which represents the positive correctly classified samples to the total number of positive samples. Specificity or true negative rate, is expressed as the proportion of the correctly classified negative samples to the total number of negative sample [20].

The accuracies and training times of KNN, SVM and Ensemble classifiers are compared as shown in TABLE II. The highest recognition accuracy of 94.45% is achieved with random subspace ensemble classifier based on the discriminant learner and the lowest training time of 2.33 seconds is obtained with KNN classifier. Because of having the highest accuracy, random subspace ensemble classifier based on the discriminant learner is chosen as the classifier for recognizing actions.

The accuracy, sensitivity and the specificity values in each execution round using random subspace ensemble classifier with the discriminant learner, are described in TABLE III. The average sensitivity and specificity values are 0.945 and 0.995, respectively. TABLE IV shows confusion matrix of the execution round (R10) and Fig 2 shows ROC curves of each action class related with the confusion matrix of the execution round (R10).

TABLE II. COMPARISON OF ACCURACIES AND TRAINING TIMES OF KNN, SVM, RANDOM SUBSPACE ENSEMBLE CLASSIFIERS

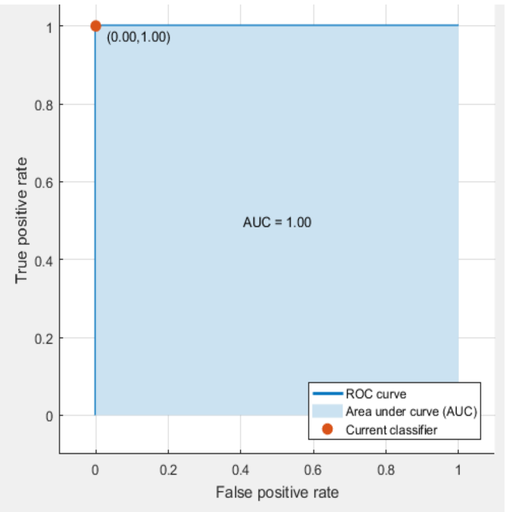
	KNN	**SVM**	**Ensemble**	
			Discriminant	**Nearest Neighbor**
Training Time (second)	2.33	6.59	8.13	3.71
Accuracy (%)	89.0	92.20	94.45	76.07

TABLE III. ACCURACY, SENSITIVITY AND SPECIFICITY VALUES IN EACH EXECUTION ROUND

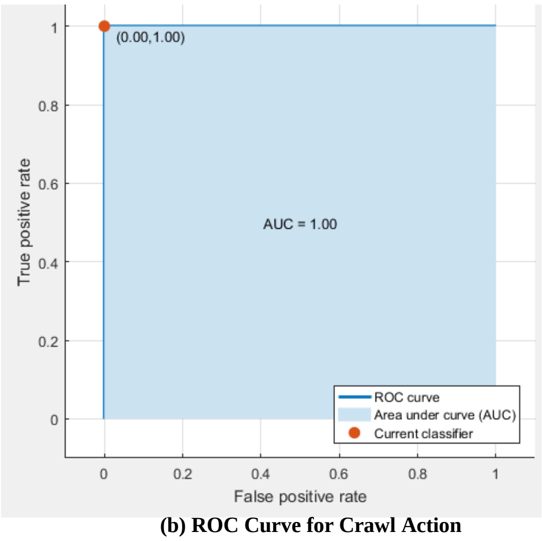
Execution Round	**Accuracy (%)**	**Sensitivity**	**Specificity**
R1	96	0.96	0.996
R2	95	0.95	0.995
R3	94	0.94	0.995
R4	94	0.94	0.995
R5	94	0.94	0.994
R6	94	0.94	0.995
R7	93	0.93	0.996
R8	94	0.94	0.995
R9	93	0.93	0.994
R10	95	0.95	0.996
R11	95	0.95	0.995
R12	92	0.92	0.994
R13	96	0.96	0.996
R14	94	0.94	0.994
R15	93	0.94	0.994
R16	95	0.95	0.996
R17	96	0.96	0.996
R18	96	0.96	0.996
R19	96	0.96	0.996
R20	94	0.94	0.996

TABLE IV. CONFUSION MATRIX OF THE EXECUTION ROUND 10

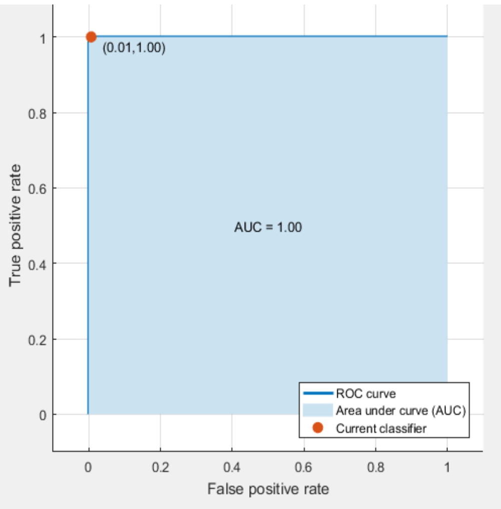
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	TPR	FPR
1	11														1	0
2		11													1	0
3			11												1	0.01
4				10								1			0.91	0
5					11										1	0
6						11									1	0
7							11								1	0.01
8								11							1	0
9									11						1	0.01
10										11					1	0
11							1				8	2			0.73	0.01
12							1				1	9			0.82	0.02
13									1				10		0.91	0
14			1											10	0.91	0



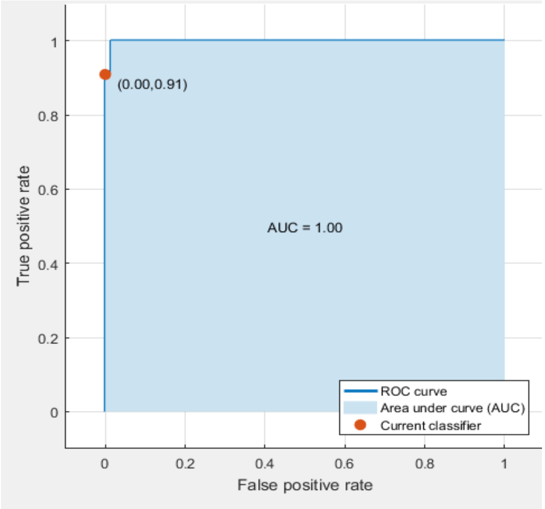
(a) ROC Curve for Clap Action



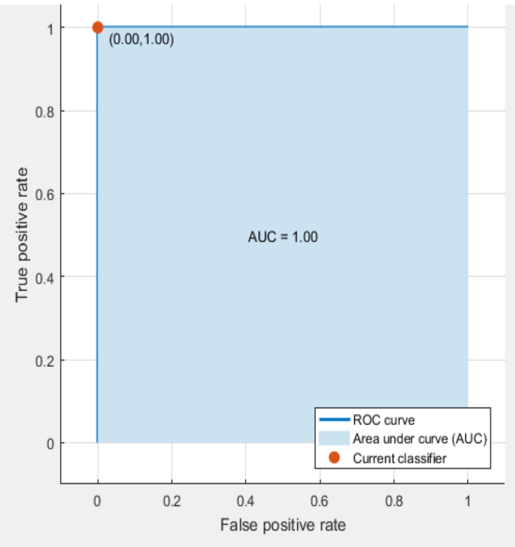
(b) ROC Curve for Crawl Action



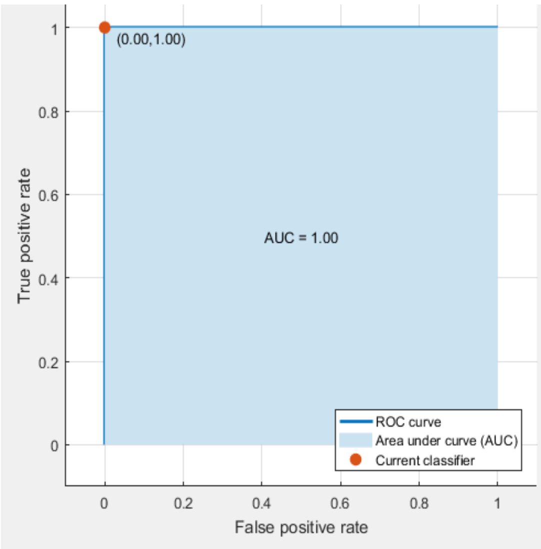
(c) ROC Curve for Jump Action



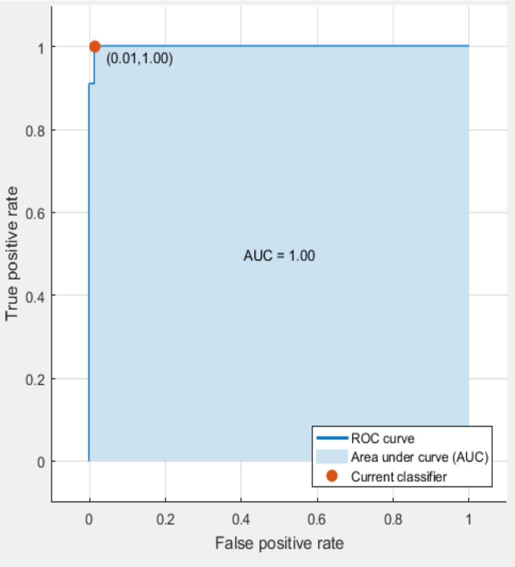
(d) ROC Curve for Jump from Sit Action



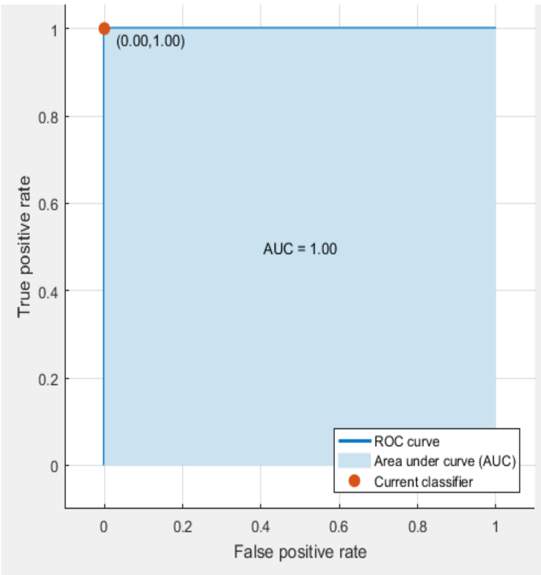
(e) ROC Curve for Jumping Jack Action



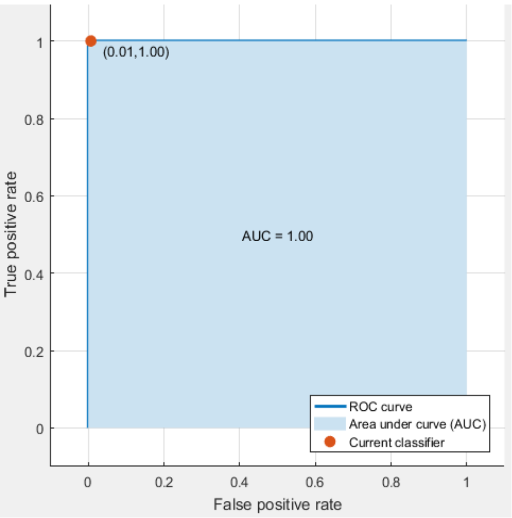
(f) ROC Curve for Push up Action



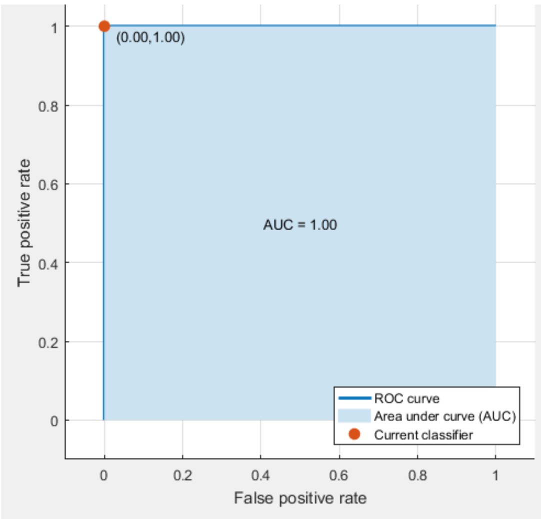
(g) ROC Curve for Raise one Hand Action



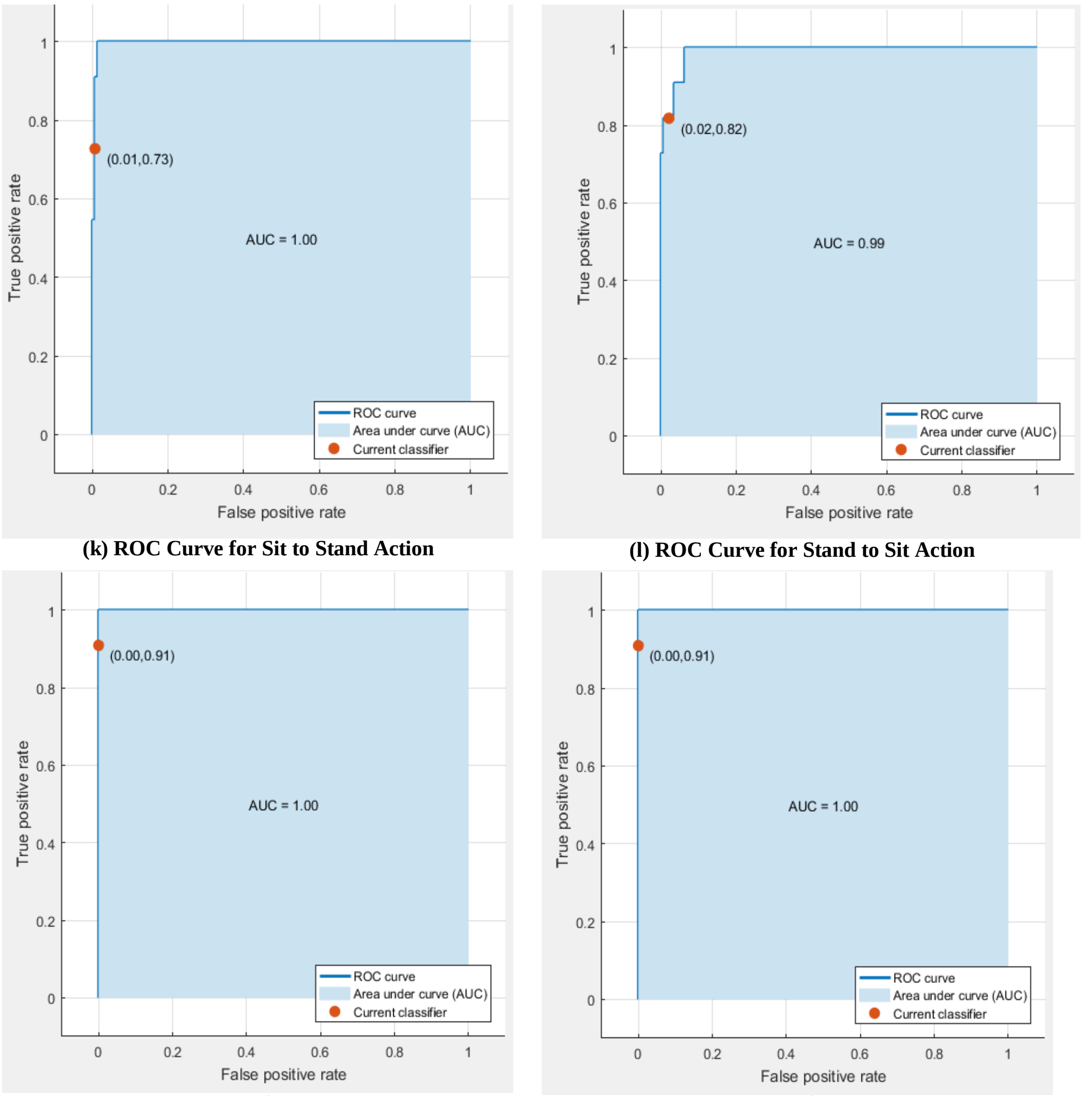
(h) ROC Curve for Run Action



(i) ROC Curve for Stretch out Action



(j) ROC Curve for Wave Action



(k) ROC Curve for Sit to Stand Action

(l) ROC Curve for Stand to Sit Action

(m) ROC Curve for Turn Action

(n) ROC Curve for Walk Action

Fig. 2 ROC curve for the execution round 10

V. CONCLUSIONS

In this paper, an efficient action recognition framework is introduced to improve the action recognition performance. In the proposed action recognition framework, salient object detection used to get the intended action region (foreground object), can reduce background interventions in processing. The changing gradient orientation feature descriptor can offer important and valued information to recognize different actions. With the help of ensemble classifier, the proposed framework can lead to having a stronger action recognition framework. According to experimental results, the proposed action recognition framework achieved satisfying action recognition accuracy. As a future work, the appearance and motion features should be combined to obtain better recognition accuracy.

DECLARATION

The authors have disclosed no conflicts of interests and the project was self-funded.

REFERENCES

[1] Kwak, N.J. and Song, T.S., “Human Action Recognition Using Accumulated Moving Information,” International Journal of Multimedia and Ubiquitous Engineering, 10(10), pp.211-222, 2015.

[2] Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K. and Buckles, B.P., “Advances in human action recognition: A survey,” arXiv preprint arXiv:1501.05964, 2015.

[3] Wang, J., Chen, Z. and Wu, Y., “Action recognition with multiscale spatio-temporal contexts,” In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 3185-3192), June 2011.

[4] Bagheri, M., Gao, Q., Escalera, S., Clapes, A., Nasrollahi, K., Holte, M.B. and Moeslund, T.B., “Keep it accurate and diverse: Enhancing action recognition performance by ensemble learning,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 22-29), 2015.

[5] Rohr, K., “Towards model-based recognition of human movements in image sequences,” CVGIP: Image understanding, 59(1), pp.94-115, 1994.

[6] Yilmaz, A. and Shah, M., “Recognizing human actions in videos acquired by uncalibrated moving cameras”, In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on (Vol. 1, pp. 150-157), October, 2005.

[7] Ali, S., Basharat, A. and Shah, M., “Chaotic invariants for human action recognition,” In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (pp. 1-8), October, 2007.

[8] Wang, L. and Suter, D., “Informative shape representations for human action recognition,” In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on (Vol. 2, pp. 1266-1269), August, 2006.

[9] Efros, A.A., Berg, A.C., Mori, G. and Malik, J., “Recognizing action at a distance,” In null (p. 726), October, 2003.

[10] Blank, M., Gorelick, L., Shechtman, E., Irani, M. and Basri, R., “Actions as space-time shapes,” In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on (Vol. 2, pp. 1395-1402), October, 2005.

[11] Laptev, I.,Lindeberg, T., “Space-time interest points,” International conferenceon computer vision, IEEE, 2003.

[12] Dollár, P., Rabaud, V., Cottrell, G. and Belongie, S., “Behavior recognition via sparse spatio-temporal features,” In Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on (pp. 65-72), October, 2005.

[13] Messing, R., Pal, C. and Kautz, H., “Activity recognition using the velocity histories of tracked keypoints,” In Computer Vision, 2009 IEEE 12th International Conference on (pp. 104-111), September, 2009.

[14] Margolin, R., Tal, A. and Zelnik-Manor, L., “What makes a patch distinct?,” In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on (pp. 1139-1146), June, 2013.

[15] Yang, B., Zhang, X., Liu, J., Chen, L. and Gao, Z., “Principal components analysis-based visual saliency detection,” In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 1936-1940), March, 2016.

[16] Hnin, M.A. and Sai, M.M.Z., “Histogram of Accumulated Changing Gradient Orientation (HACGO) for Saliency Navigated Action Recognition,” 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD, 2017), Kanazawa, Japan, pp. 225-230, ISBN: 978–1–5090–5504–3.Pattern Analysis & Applications, 5(2), pp.121-135, June 26-28, 2017.

[17] Skurichina, M. and Duin, R.P., “Bagging, boosting and the random subspace method for linear classifiers,” Pattern Analysis & Applications, 5(2), pp.121-135, 2002.

[18] Kayal, P. and Kannan, S., “An Ensemble Classifier Adopting Random Subspace Method based on Fuzzy Partial Mining,” Indian Journal of Science and Technology, 10(12), 2017.

[19] Tran, D. and Sorokin, A., “Human activity recognition with metric learning,” In European conference on computer vision (pp. 548-561). Springer, Berlin, Heidelberg, October 2008.

[20] Lever, J., Krzywinski, M. and Altman, “Points of significance: classification evaluation,”, 2016.