International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Hand-Held Object with Action Recognition Based On Convolutional Neural Network in Spatio Temporal Domain

R.Rajitha Jasmine, K.K.Thyagharajan

Abstract: ***Several applications such as object recognition and face recognition are established with the progress of smart devices and computer technology, to assist human-computer interaction (HCI). In HCI, Hand-held object recognition hasamain role. This approach helps the computer to realise the user's intentions and also meetsthe user requirements. Hand as an organ which is considered as a direct and natural way of communication for humans. The Hand-held Object Recognition (HHOR) assigns a label for the object which is heldin hand this could help machines in understanding the environment and the intention of the people. However, it has not been well studied in the community.So, in this paper, we proposed system for recognizing such activities happening between hands and faces in real time. The interaction events (e.g. eating, phoning and smoking) between hands and faces are analysed using the event analysis approach. Ratio histogram is used for obtaining the essential colour bins for detecting the desired objects via re-projection method. For object tracking and feature extraction, a code book method is used. To recognize various human-object interaction events, the dynamic and multiplicity contexts of event are modelled together. Finally, atwo stage cascaded CNN classifiers for the recognition is implemented as this technology improves the performance of object recognition.To make fair comparisons, six methods were compared in this paper based on the HMDB dataset. This system is effective and can be performed in real time because an exhaustive search process to find possible interaction pairs in the huge space of all possible event parameters is not involved. Experimental results have proved the superiority of our proposed system to analyse different human behaviours and events between hands and a face.***

Keywords: ***Computer technology, Hand-held object, convolutional neural network,***

I. INTRODUCTION

In computer vision, Human action recognition is a well-known subject with several vital applications,for example video surveillance and video retrieval. The intention of Human action recognition is to classify the actions in a video automatically.As stated in [1], the performance of action recognition has an impact in the quality of action representations.Recently,Human action recognition in videoshas been developed to be afamiliar area of investigationowing to its applications in sports analysis, retrieval of video, interaction ofhuman and computer,monitoring of health and video surveillance.In the literature,a lot of papershave been published, each one highlighting a specific attribute of recognition. In [2, 3] many methods for action recognitionaccomplished by a single person isfocused which classify the entire motion of bodybyclassifying them into spatial and temporal structures. Regarding the human action recognition in [4] a discussion is conducted about multi-view 2D and 3D approaches. For human action recognition in literature [5] severaldatasets arepresented to address several problems like realistic activity recognition, source based interaction and multi-view analysis.

Thetechniques for action recognition mostly rely on the descriptors or some of the features extractedfordiscriminateinformation for classification. In human action recognition, normally used features/descriptorsare histograms oriented gradient (HOG), histograms of optical flow (HOF) [7], bag-of-visual-words (BoVW) [6], motion boundary histograms (MBH) [8], densetrajectories [9] and action bank features [10].

Object recognition is the commonly used technology in HCI and it is considered as a standard task in multimedia and computer vision fields. However, in reality, this task still face many complications due to the impactof some factors such as deformation, illumination, occlusion, etc. For object recognition, certain methods are implemented with the aim toachieve more robust hand-craft feature representation such as SIFT [11], SURF [12], etc.

Recently, there is an emerging growth in Deep Learning (DL) technologies. In 2012 Krizhevsky et al proposed AlexNet [13] which is a classic Convolutional Neural Network (CNN)model. Later, thisCNN model becomes the frequently used method for scene and object recognition [14] [15].In most of the computer vision tasks DLapproaches have attainedboundlessachievements. However, these models use the supervised learning techniques which requires many iterations andlarge volume of labelled data to train their huge parameters.

Flexibility of deep learning models are restricted by thedescription of labelled data towards new classes and limitstheir applicationsto cope withrecentlyevolving objects or raretypes where the annotated imagesare limited.A method for hand-held object detection have been presented in this paper, moreoverseveral actions (e.g., eating, phoning and smoking)related to hands and faces in reality has been analysed.It is known that, a handheld object differsconsiderablyin its shapes, size, textures and colours invariedlighting conditions and viewpoints.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
R.Rajitha Jasmine, Department of Information Technology, RMK Engineering College, Chennai
K.K.Thyagharajan, Department of Electronics &Communication Engineering, RMD Engineering College, Anna University, Chennai.

For handheld object detection, we present a ratio histogram to equate the colour variationsconcerning the frames at the time of interaction process. Then to locate the object of interest a projection technique is used, whichextracts the significant colour bins .The Classifier is built using CNN whichclassifies the hand held objects along with the actions performed. The work flow of this proposed system is illustrated in Fig. 1. In a collection of action videos, the performance of the proposed method isexaminedto verify its competence in event analysis. This paper's remaining part is structured as follows. Section 2 presents the related works. Section 3 presents the details on feature extraction. Section 4 presents the Object Tracking using Code Book approach. Section 5 presents the process of Interaction Events representation.Section 5 shows the experimental results. Finally, Section 6 concludes the work.

II. RELATED WORK

A. RGB-D object recognition

Object recognition in computer vision is a standard task. Generally, object recognition using RGB images is a difficult task owing to the variable illumination caused by distressing real-word incidents.

Recent developments of the commodity depth cameras produce agrowingvolume of visual data that has colourand depth measurements.The depth information improves the performance of object recognition because of its robustness to the depth measurements for light and colour variation.Depth image correcting errors was used by Yanhua Cheng et al [16] in indoor semantic segmentation of images in object classification.

To clearly encode the 3D shape information from depth mapsa novel and effective descriptor was proposed by C Zhang et al [17]. An object detection method in 3D was presented by W Kehl et al [18] for 6D vote casting that usedreduced descriptors of locally-sampled RGB-D patches.A convolutional auto-encoder was employed for regression which was trained on a huge volume of random local patches. To show the dominance of this approach evaluation was done in three datasets. Besides, object recognition and RGB-D images were used by Xinhang Song et al [19] [20] to improve scene recognition act.

B. Hand-held object recognition

HHORbased studies arecategorised into two types i.e. interface of firstperson and interface of second-person. Literature in [21] [22] [23] focusedon first-person interface. Hand-held object is divided by the Authors in [21] using motion and fuses the objectsmotion and background movement.

Authors in [22] estimated the region of interest for the HHO and the object at distance using laser pointer and head pose estimation.

ROI is found foridentifying the HHO along with SIFT. In [23] the study is based onHHOR. The data set used by them isCORe50, which has 50 household objects belonging to 10 classes. Using CNN theyoffer a benchmark for this task.Furthermore, based on second-person interfacesome of the works has been done. In which,using RGB-D devices the main task is done, user'sskeletal information is used to detect the position of hand and object segmentation is done by attaining the depth information. RGB, depth image features and 3D point cloud based hand-craft features are extracted using Fine-tuned AlexNet [26] used by Lv et al [24] [25], finally threetypes offeatures arefused andan optimal result is obtained.

III. HAND-HELD OBJECT RECOGNITION

Face detection from videos is done using the Adaboostsystem [28].The colour difference of the faces are compared using ratio histogram in the interaction process. For finding the location of the HHO, a re-projection approach is used.For event representation three visual features i.e. the objects size,distance between the object and mouth along with density of the smog are obtained.Then, each event can be recognised using a two stage cascadedCNN [40] classifiers.The work flow is presented in Fig. 1. In the recognition phase some of the approaches such as feature extraction, feature fusion and classification are involved. The subsequent sections provides more details about the dissimilar features and the methods used.

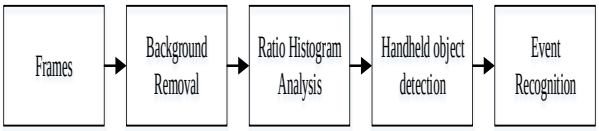


Fig. 1. Proposed Work flow

In the input video, the activities occurring between the face and hands is analysed by extracting the HHO.Identifying the object is a difficult task owing to unknown features such as shapes, sizes, colours, and textures. TheHHOcan also be detected from the face using the histogram-based approach.In this paper weadoptaskilledAdaboost-based detector [28] to identifythe frontal and the non-frontal faces. Then the ratio histogram is computed to represent the possible frames in which the HHO appears or disappears.

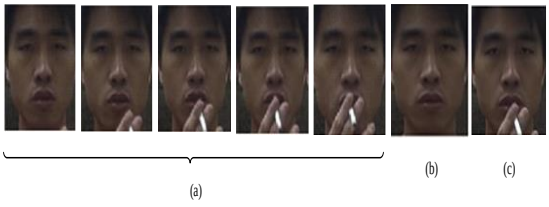


Fig. 2. Classification of Action for smoking events

Fig. 2(a) illustrates the classification of action for smoking event. In Fig. 2(b),a frame is shown before smoking i.e. when the cigarette is not present and in Fig. 2(c) the frame in which cigarette appearing is shown. To identify the object (i.e. cigarette) based on frame (b) and (c) the ratio histogram is used. The identifiedface is denoted as F and the observation captured at time*t*is denoted as F_t. Moreover, H_t (i) is represented as the colour histogram of Ft. The ratio histograms of F_t in the forward and backward states at the r^{th} order isdescribed below:

$$RatioH_t^{r+}(i)=\frac{H_t(i)}{1+H_{t-r}(i)} \text{ and } RatioH_t^{r-}(i)=\frac{H_{t-r}(i)}{1+H_t(i)} \quad (1)$$

From Eq. (1), $RatioH_t^r$ power is defined below:

$$\left\|RatioH_t^r(.)\right\|=\sum_{i=0}^{n_{bin}-1} RatioH_t^r(i) \quad (2)$$

The number of colour bins used in H_t is represented as n_{bin}.The $\left\|RatioH_t^{r+}(.)\right\|$ will be higher than $\left\|RatioH_t^{r-}(.)\right\|$ if an object is appeared in F_t. Conversely, $\left\|RatioH_t^{r-}(.)\right\|$ will be largerif the object disappears in F_t. To conformthe appearance or the disappearance of anobject in F_t, $\left\|RatioH_t^r(.)\right\|$ is a helpful factor. The rule in Eqn. 3 will be satisfied, if $F_t$contains a new object.

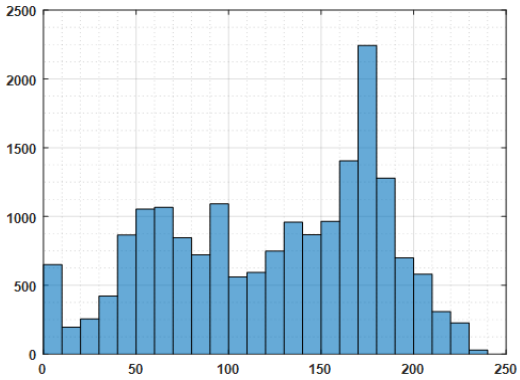
$$\left\|RatioH_t^r(.)\right\|>\tau_{ratio} \quad (3)$$

The HHO's position can be located using $RatioH_t^{r+}$. In $RatioH_t^{r+}$ the average values ofall binsis $T^H_{Ratio_t^r}$. If $RatioH_t^{r+}(i)>1.5\,T^H_{Ratio_t^r}$ for a colour bin i, then for highlighting the object the colouriisasignificantcolour bin. The re-projection result of important colour bins found from $RatioH_t^{r+}$ is illustrated in Fig. 4(c).

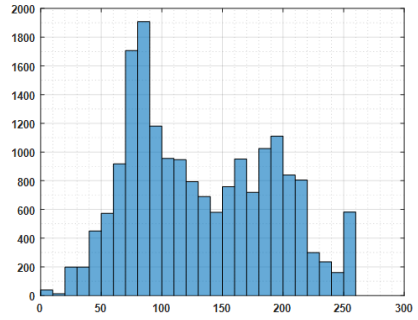
Certainly, most of the prominent colours are detected centrally around the cigarette area .Then the location is found using a connected component analysis. The region having maximum area is selected as the object candidate.

(a) (b) (c) (d)

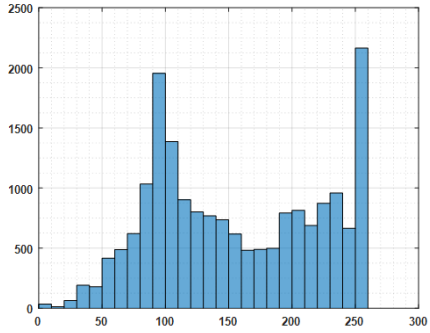
Fig. 3. Identifying the object using colour re-projection approach



(a)



(b)



(c)

Fig. 4. Ratio histograms

The Colour histograms of Fig. 3(a) and (b) is shown in Fig 4 (a) and (b) and the Ratio histogram between (a) and (b) is shown in Fig 4c.

Ratio histogram $RatioH_t^{r+}$ between Fig. 4(a) and 4(b) are shown in Fig. 3(c). From the training video set the average value of $\left\|RatioH_t^r(.)\right\|$ calculation is denoted as τ_{ratio}.

A. Tracking theObject using Code Book

TheHHO's trajectory is tracked using a codebook scheme. The particle filters [29] and mean-shift [29] cannot be applied here because of the issuesregarding the occlusion of hand and direction of the object.

Using the ratio histogram if a HHO is identified, thissystem will examineifit is recorded in codebook listor not. If sure, its visual features will berecorded and updated into the codebook. If no, it will be added asa new object. For recording the colours, a Gaussian modellingapproach is adopted.Theobject detection is denoted as O with variance (ρ_R,ρ_G,ρ_B) and colour means (μ_R,μ_G,μ_B).

Based on O, the probability of a pixel p is modelled using the below condition Eqn. 4,

$$G_{object}(p)=\exp\left[-\frac{(r_p-\mu_R)^2}{\rho_R^2}-\frac{(g_p-\mu_G)^2}{\rho_G^2}-\frac{(b_p-\mu_B)^2}{\rho_B^2}\right] \quad (4)$$

The colour channels of p is denoted as (r_p, g_p, b_p). Between two adjacent frames, the visual properties of O in real cases doesn'tchange. Colour is considered as a useful featurefor tracking.For every pixel p, whether p belongs to O can be determined using Eqn 5.

$$G_{object}(p)<0.85 \quad (5)$$

Later, I_t will betransformed to a binary map and the position of O will be extracted by applying the connected component analysis.O can be found and tracked whenthe object listed in the codebook is conformed. Output of tracking the HHO using the proposed approach is illustrated in Fig. 5.After tracking O, the distance is found for representing an action.

For a detected face F_t with dimension $w_{Ft} \times h_{Ft}$ andcenter (c_{Ft}^x, c_{Ft}^y), using the Adaboost approach the mouth region can be identified [30]. Though, due to the hands the mouth is always occluded. Hence, to prevent the issues of occlusion and also considering the efficiency, the mouth is expected to be fixed in F_t with the position $\left(c_{Ft}^x, c_{Ft}^y - h_{Ft}/3\right)$.

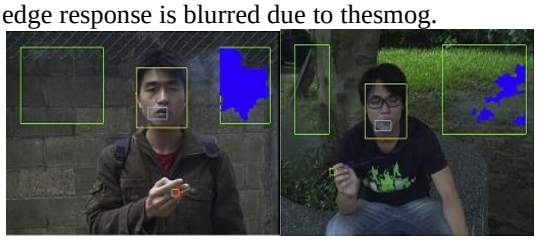
(a)(b)

Fig. 5. Results of Colour slicing based Object tracking. For event representation thedistance of object to the mouth is considered as a visual cue.

In Fig. 5, two examples are shown in which the mouth region is represented as blue squares shown in the face.

B. Smog density estimation

In this paper three features are used to represent an interaction event.i.e., size of object, distance between hands and the HHO, and the density of smog. In these, the smog's visiblefeaturesareunclear and inexact in texture and shape.Thus, the density of smog is termed ''inferior'' because compared with the other features it is less reliable in representing an event.As the smog flutters in the air isrepresented as a moving block.Using the background subtraction or optical flow approach ''moving'' feature can be extracted. In this approachbackground subtraction is done using codebook technique [31]. To calculate the motion ofpixels, optical flow isused. Then, for event analysis the real smog blocks are isolated from the non-smog ones. For the verification task varioussmog features will be extracted from the HIS colour space. The smog colour is distributed in a specific area after analysing a set of training videos. A pixels edge response is blurred due to thesmog.



(a) (b)

Fig. 6. Smog detection results. Blue colour denotes the smog region

In this paper, ten sub-band featuresare extracted from each moving block using three levels of wavelet decomposition. The results of smog detection are shown in Fig. 6. Bounding boxes B_{left} and B_{right} are set besides the face region F_t, to calculate the density of smog.

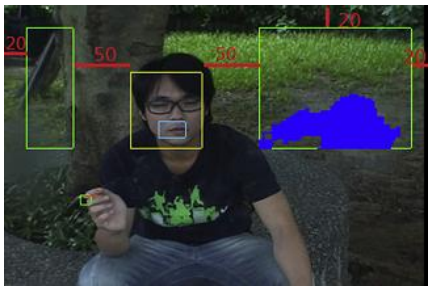


Fig. 7. Bounding box settings

Bounding box's geometry settings are shown in Fig. 7. Both the base of B_{left} and $F_t$are similar and the boundariesof B_{right} is alsoset similarly. The pixels of the smogidentified within B_{left} and B_{right} is denoted as n_{smog}. Then, den $_{smog}$ (i.e. the density of smog) is described as: $den_{smog} = n_{smog} / n_{box}$. Eight action primitives with this density for event analysis isspecified in Section 4.

IV. INTERACTING EVENTS BETWEEN HAND AND FACE

Visual features arerequired to analyse the interactive behaviours between faces and handsafter that each frame is convertedinto different action primitives. The event representation and theframe coding details are discussed in this section.

A. Frame Coding

In event analysis, each frame is converted into a symbol m_i using three bins, weights are added to each bin. The weight for the k^{th} bit is denoted as w_{b_k} and $\oplus$ is the *X*-OR operator.

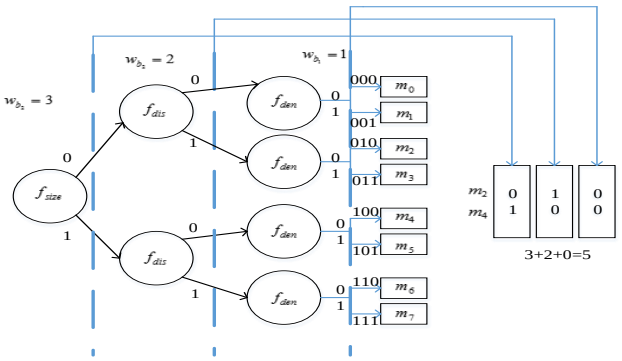


Fig. 8: Different weights for symbol comparisons

In the above fig. 8 shows different weights for the symbols comparison. We assign 3, 2, 1 as the values of w_{b_2} , w_{b_2} , and w_{b_0} . The bit weight w_{b_2} of the object size is the highest, because than other features it intensely identifies the cigarettes.

The distance between m_i and m_j is calculated as follows:

$$\beta[m_i, m_j] = \sum_{k=0}^{K} w_{b_k} \left(b_k^{m_i} \oplus b_k^{m_i} \right) (6)$$

Where,the k[th] bit of the symbol *m*is denoted as b_m^k.The smog density's bit weight w_{b_0} is easily affected by noise.

B. Event Representation

A set of action primitives used to convert an event to a string is denoted as *M*. Observation of an event is denoted as O,a series of symbols are convertedinto *O*using *M*i.e. $O = \{A_1, A_2, ...A_n\}$, where $A_t \in M$. Many repeated actions will be included in an event in a real cases. A new symbol S_i is used to represent *O*, whichimplies the recurrentstate where S_i is the*i*[th] action in *O* which is repeated r_i times; i.e., $S_i = (A_i, r_i)$ and $A_i \in M$. Hence, more efficiently *O* can berepresented as:

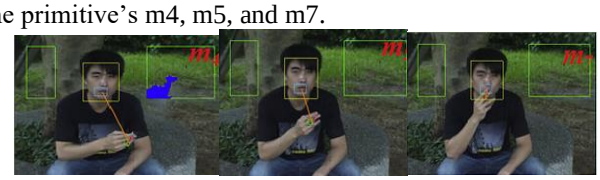
$$O = \{(A_1, r_1), (A_2, r_2), ..., (A_n, r_n)\} = (A, r) \quad (7)$$

Where $A = (A_1, A_2, ..., A_n)$ and $r = (r_1, r_2, ..., r_n)$. Eight action primitives generated toencode I_t is show in Fig 10.

Table 1. Physical meanings of eight action primitives.

Primitives	R	L	S	Meaning
m_0	0	0	0	Non-cigarette, leaving, escalating
m_1	0	0	1	Non-cigarette, leaving, declining
m_2	0	1	0	Non-cigarette, approaching, escalating
m_3	0	1	1	Non-cigarette, approaching, declining
m_4	1	0	0	Cigarette, leaving, escalating
m_5	1	0	1	Cigarette, leaving, declining
m_6	1	1	0	Cigarette, approaching, escalating
m_7	1	1	1	Cigarette, approaching, declining

Eight action primitives' physical meaningsare illustrated above in table 1. Examples are presentedin Fig 9 to clarify the primitive's m4, m5, and m7.



(a) m_4 (b) m_5 (c) m_7

Fig. 9. Visual interpretations of the Primitives m4, m5 and m7. (a) Cigarette, leaving, escalating. (b) Cigarette, leaving, declining. (c) Cigarette, approaching, declining.

V. EVENT RECOGNITION USING TWO CASCADED STAGES CNN CLASSIFIERS

The design of the CNN classifier-1 having 5 layers with weights are shown in Fig 10. The first four layers are the convolutional layers and the last layer is the fully-connected layer. Thepresence of hand-held object is rapidly verified at this stage, so this classifier is constructed with edge map input. With the similar architecture, theclassifier 2 is built. In the training step of CNN, the over fitting problem is common. To deal with this, the dataset is artificially enlarged for CNN training particularly in small dataset. Two distinct forms of data augmentation are done. The first is mirror image,using this the size of the training data gets increased and also reduces over fitting and creates more training samples.

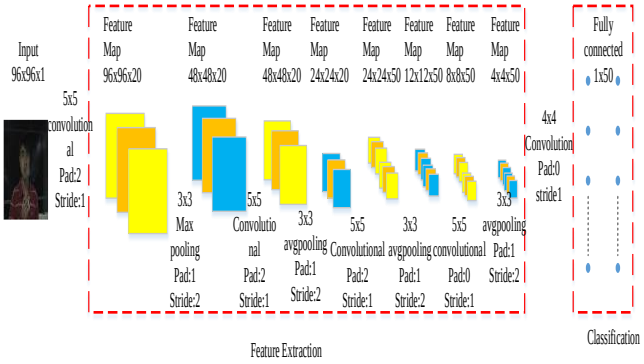


Fig. 10. The architecture of our CNN classifier

The second is initiating PCA [39] (Principal Components Analysis) in Alex-Net [32],which uses three quantities to alter the ratio histogram channels intensity. They are defined by:

$$[E_1, E_2, E_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T \quad (8)$$

Where λ_k and E_k are the eigenvalue and the eigenvector of the 3 × 3 covariance matrix of histogram values, the random variable are denoted as α_k.

VI. EXPERIMENTAL RESULTS

In this paper, threeevents such as drinking, phoning and smoking are collected to evaluate the performance of the proposed approach. Twenty actors performed each action certain times inseveral lighting conditions; as a result,100 sequences were collected for each event. Other kinds of events can also be analysed by using our proposed scheme. Frame rate set for the system is 15fps.Simulation is conducted in Mat lab (R2014a) on a personal computer, configured with Intel 4720HQ 2.6GHz processor and 8 GB memory. For performance evaluation the first set of experiments are used.The outcome of detecting a cigarette is shown in Fig. 11 which is displayed in the open environment. Fig 11b and d shows the detection results from Fig 11 a and c.

(a)(b) (c) (d)

Fig. 11. Detection results of Cigarette. The original frame is shown in (a) and (c). The ratio histogram based detection results are presented in (b) and (d).

Table 2 lists the accuracy analysis of the detection of phone and cigarette. No tracking technique was used here to make a fair performance evaluation. The number of frames used for testing is 2200. The number of ‘‘false-positive’’ is 32, ‘‘true-positive” is 2732and ‘‘false-negative’’ is 249. Accuracy for detecting cigarettes using the proposed approach is found to be 91.7%.

Table 2 Accuracy analyses of cigarette, can and phone.

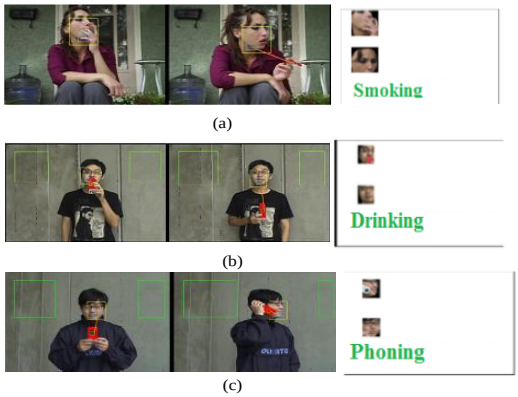
object	frames	True positive	False negative	False positive	False alarm %	Miss rate	Accuracy
Can	9228	8923	40	150	1.72	0.52	96.2
Phone	3687	3244	312	131	3.55	8.46	88.9
Cigarette	3012	2732	248	32	1.08	9.23	91.7

Because of the hand occlusion caused in the phoning actionsthe phone is notseen clearly. Therefore, identifying the phone in the front view is not possible compared to the sided view. Accuracy for phone detection is 88.9%.

(a) (b) (c) (d) (e) (f)

Fig. 12. Results of phone and can. (a) (c) and (e) shows the original image. (b) (d) and (f) shows the detected images.

The can which is held in hand is also detected for detecting the drinking events from videos. The results of detecting the can and phone is seen in Fig. 12, the phone is detected in the side view.



(a) (b) (c)

Fig 13(a) Results of smoking event detection, (b) Results of drinking event detection, (c) Results of phoning event detection.

The results of smoking, drinkingand phoning event analysis from YouTube videos is presented in Fig. 13. The experiments done above have confirmed the dominance of proposed approach in HHOR.

Table 3. Accuracy comparisons of event analysis

Approach	Action categories			
	Smoking (%)	Eating (%)	Drinking (%)	AVG (%)
HOG	35.49	34.23	34.88	34.87
GIST	31.65	28.81	29.37	29.94
HOF	37.19	34.68	35.85	35.91
Trajectory	34.65	32.23	33.11	33.33
Action bank	38.46	36.35	36.84	37.22
Trajectory+ HOG + HOF	41.55	39.43	39.87	40.28
Our method	43.67	40.21	41.95	41.94

Based on the HMDB [33] dataset,six methods areused for comparisonare,histogram of optical flow (HOF)[34], histogram of oriented gradients (HOG)[35], dense trajectory, action bank[36], GIST[37], HOF and HOG based dense trajectory.

The collected dataset comprises of human–human communications, facial and general body movements. From the YouTube videos and movies many video clips are collected. Three action categories (i.e. including, smoking, phoning and eating) are selected from the dataset for comparison. Spatial–temporal interest points (STIP) for every clips were initially calculated in HOG and HOF methods, then for action classification their 3D HOGs and HOFs were extracted.Action bank [36] represents the Video contents. GIST [37] is suggested in [33] to represent video frames. Dense trajectories and motion flows are used by the dense sampling scheme for action classification [38] which providesintense local structural information. Table 3 displays the comparison of accuracy with the methods mentioned above. Less feature points are extracted by the STIP approach [41] which does not fit for action recognition. Among other descriptors used for comparison, the dense sampling scheme [38] performed well. The proposed method outperforms the existing approaches in classification. Obviously, eachaction in real time can be analysed by the proposedmethod.

VII. CONCLUSIONS

As it is known that a HHO’s can vary considerably in its sizes, shapes, colours, and textures under distinct viewpoints and lighting conditions. The detection task is made very challenging due to the uncertainty. Another challenge along with the uncertainty, is the multiplicity of event representation. In real time, an event is rarely performed with the same, repetitions, beginning states and ending states.

The multiplicity of an event can vary significantly in its representation. Number of possible combinations and relative configurations between the analysed object and body parts makes the multiplicity to increase in order to make the interaction event analysis more complicated.Ratio histogram is used in this approach for detecting the object held in hand and it is applied to analyse severalactions (e.g. smoking, phoning and eating). The colour difference between frames are compared using the ratio histogram during the interaction process. The object of interest is located using a projection technique by extracting significant colour bins. Three visual features are extracted for event representation i.e. the size of object, the distance concerning mouth and the object held in hand, and the density of smog. To recognize the interaction events between human and object, the dynamic and multiplicity contexts of events are modelled together. CNN classifiers are used to recognise the events. For event analysis, the average accuracy comparisons among different methods on the HMDB dataset is found to be 41.94%. This approach can be performed in real time because it doesn't have an in-depth search process to find possible interactive pairs in the huge space of all possible event parameters. Our proposed method have been rigorously validated on a considerable event videos to prove its dominance in event analysis from the prospects of efficiency and robustness.

REFERENCES

1. D. Weinland, R. Ronfard, E. Boyer, "A Survey of Vision-based Methods for Action Representation,"Segmentation and Recognition, Elsevier Science Inc, 2011.
2. J. Aggarwal, M. Ryoo, "Human activity analysis: a review," ACM Comput. Surv, vol. 43(3), 2011, pp. 1–43. http://dx.doi.org/10.1145/1922649.1922653.
3. J. Lei, G. Li, J. Zhang, Q. Guo, and D. Tu, "Continuous action segmentation and recognition using hybrid convolutional neural network-hidden Markov model model." IET Computer vision, vol. 10(6), 2016,pp. 537-544.
4. M.B. Holte, C. Tran, M.M. Trivedi, and T.B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments", IEEE Journal of selected topics in signal processing, vol. 6(5), 2012, pp.538-552.
5. S. Vishwakarma, and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," The Visual Computer, vol. 29(10), 2013, pp. 983-1009.
6. X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," Computer Vision and Image Understanding, vol.150, 2016, pp. 109-125.
7. R.V.H.M. Colque, C. Caetano, M.T.L. de Andrade, and W.R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos", IEEE Transactions on Circuits and Systems for Video Technology, vol.27(3), 2016, pp. 673-682.
8. I.C. Duta, J.R.R. Uijlings, T.A. Nguyen, K. Aizawa, A.G. Hauptmann, B. Ionescu, and N. Sebe, "Histograms of motion gradients for real-time video classification," In 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), 2016, pp. 1-6.
9. B. Fernando, P. Anderson, M. Hutter, and S. Gould, "Discriminative hierarchical rank pooling for activity recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1924-1932.
10. J. Yang, Z. Shi, and Z. Wu, "Vision-based action recognition of construction workers using dense trajectories," Advanced Engineering Informatics, vol. 30(3), 2016, pp. 327-336.
11. G.L. David Object recognition from localscale- invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conferenceon Computer vision, Vol. 2, 1999, pp. 1150–1157.
12. B.Herbert, A. Ess, T.Tuytelaars, and L.V. Gool, "Speeded-uprobust features (SURF),"Computer vision and image understanding, vol.110, 3(2008), pp. 346–359
13. A. Krizhevsky, I. Sutskever, and G.E. Hinton,"Imagenet classification with deep convolutional neural networks," In Advance sin neural informationprocessing systems, 2012, pp. 1097–1105.
14. C. Zheng, J. Chen, J. Kong, Y. Yi, Y. Lu, J. Wang, and C. Liu, "Scene Recognition via Semi-Supervised Multi-Feature Regression," IEEE Access, vol.7, 2019, pp. 121612-121628.
15. X.Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, "Scene recognition with objectness," Pattern Recognition, vol. 74, 2018, pp. 474-487.
16. Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Localitysensitivedeconvolution networks with gated fusion for rgb-d indoor semantic segmentation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 3.
17. C. Zhang, and Y. Tian, "Histogram of 3D facets: A depth descriptor for human action and hand gesture recognition," Computer Vision and Image Understanding, vol.139, 2015, pp. 29-39.
18. W. Kehl, F. Milletari, F. Tombari, S. Ilic, and N. Navab, "Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation," In European Conference on Computer Vision, pp. 205-220. Springer, Cham, 2016.
19. X. Song, L. Herranz, and S. Jiang. Depth CNNs for RGB-DScene Recognition: Learning from Scratch Better than Transferring from RGBCNNs. In AAAI, 2017, pp. 4271–4277
20. S. Max, H. Schulz, and S. Behnke, "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features," In 2015 IEEE international conference on robotics and automation (ICRA), pp. 1329-1335. IEEE, 2015.
21. X. Renand, C. Gu. "Figure-ground segmentation improve shandled object recognition in egocentric video," In IEEE Conference on ComputerVision and Pattern Recognition (CVPR), IEEE, 2010,pp. 3137–3144.
22. Y. Richard, D. Xu, and J.S. Jin, Individual Object Interaction for Camera Control and Multimedia Synchronization, vol. 5 (5), 2006, pp. V–V.
23. V.Lomonaco, and D. Maltoni, "Core50: a new dataset and benchmark for continuous object recognition,"arXiv preprint arXiv:1705.03550, 2017.
24. X. Lv, S.Q. Jiang, L.Herranz, and S. Wang,"Rgb-d handheld object recognition based on heterogeneous feature fusion," Journal of Computer Science and Technology, vol.30 (2), 2015, pp.340–352.
25. X. Lv, X. Liu, X. Li, Xue Li, S. Jiang, and Z. He, "Modality-specific and hierarchical feature learning for RGB-D hand-held object recognition,"Multimedia Tools and Applications, vol.76 (3), 2017, pp. 4273–4290.
26. K. Alex, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," In Advances in neural informationprocessing systems,2012, pp.1097–1105.
27. P. Viola, and M. Jones, "Robust Real-time Object Detection," International Journal of Computer Vision, vol.57(2), May2004, pp.137-154.
28. http://alereimondo.no-ip.org/OpenCV/34/profileFace10.zip
29. S.Guo, X. Shi, Y. Wang, and X. Zhou, "Non-rigid object tracking using modified mean-shift method," In Information Science and Applications (ICISA) 2016, pp. 451-458. Springer, Singapore, 2016.
30. R. Martin, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," In 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2018, pp. 10-20.
31. R. Liu, Y. Ruichek, and M.E. Bagdouri, "Enhanced Codebook Model and Fusion for Object Detection with Multispectral Images," In International Conference on Advanced Concepts for Intelligent Vision Systems,Springer, Cham, 2018, pp. 225-232.
32. A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks,"inProc. of Advances in Neural Information Processing Systems, 2012.
33. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, "HMDB: a large video database for human motion recognition,"in: IEEE International Conference on Computer Vision, 2011, pp. 2556–2563.
34. I. Laptev, P. Perez, "Retrieving actions in movies,"in: International Conference on Computer Vision, October 2007.
35. I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies," in: IEEE Conference on Computer Vision and Pattern Recognition, June 2008.

36. S. Sadanand, J. Corso, “Action bank: a high-level representation of activity in video,”in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1234–1241.
37. A. Oliva, A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope,”Int. J. Comput. Vision, vol. 42, 2001, pp. 145–175.
38. H. Wang, A. Klaeser, C. Schmid, C.-L. Liu, “Dense trajectories and motion boundary descriptors for action recognition,”Int. J. Comput. Vision,vol. 103 (1), 2013, pp. 60–79.
39. H. Jegou, and O. Chum,“Negative evidences and cooccurrences in image retrieval: the benefit of PCA and whitening,”In Proc. of European Conference on ComputerVision, 2012.
40. D.Y. Chen, and H.S. Wang, "Handheld Food Localization and Food Recognition Using Convolutional Neural Network," In Proceedings of the 2018 International Conference on Digital Medicine and Image Processing, 2018, pp. 61-64. ACM.
41. J.C. Niebles, H.C. Wang, F.F. Li, “Unsupervised learning of human action categories using spatial-temporal words,” Int. J. Comput. Vision, vol. 79 (3), 2008, pp. 299–318.

AUTHORS PROFILE

Ms.R.RAJITHA JASMINE B.TECH, M.E,(Phd)., is an Assistant Professor in Department of Information Technology at RMK Engineering college, since June 2006. She obtained her B.Tech(IT) from CSI Institute of Technology (Anna university) and M.E (Computer Science) from RMK Engineering college (Anna University), Chennai. She is currently pursuing Ph.D in Anna University and her research area is in Computer Vision & Machine learning. She has 13 Years of working experience in the teaching profession and has handled UG program subjects such as Software Engineering, Computer Architecture, Object Oriented Analysis and Design, Information management. She has presented and published 4 papers in International/National journals. Her area of interest includes Machine Learning, Software Engineering and Computer Networks. She also attended workshops, seminars and faculty development programs. She is a Life member of professional societies like ISTE, ACM and IACSIT.

K.K. Thyagharajan received his B.Eng. degree in Electrical and Electronics Engineering from PSG College of Technology, Madras University, India and received his M.Eng. degree in Applied Electronics from Coimbatore Institute of Technology, India in 1988. He also possesses a Post Graduate Diploma in Computer Applications from Bharathiar University, India. He obtained his Ph.D. degree in Information and Communication Engineering from College of Engineering Guindy, Anna University, India in 2007. He is in teaching profession for more than three decades and served at various levels including Principal, Dean and Professor at various Engineering Colleges in Tamil Nadu-INDIA. During his tenure as a Principal and Dean he was a member of Governing Council of RMK Group of Institutions. He has written 5 books in Computing including “Flash MX 2004” published by McGraw Hill (India), which has served recommended as text and reference book by universities. He is a grant recipient of Tamil Nadu State Council for Science and Technology. He has been invited as chairperson and delivered special lectures in many National and International conferences and workshops. He is reviewer and editorial board member for many International Journals and Conferences. He is a recognized supervisor for Ph.D candidates and Master students at Anna University. He has published more than 100 papers in National & International Journals and Conferences. Seven candidates have completed PhD and nine more are doing PhD under his supervision. His research interests include Computer Vision, Semantic Web, Image & Video Processing, Multimedia Streaming, Video Coding, Content-based Information Retrieval, Microcontrollers and e-learning. He is a life member of ISTE, CSI (INDIA) and also senior member and invited member in many professional associations. He has been recognized as a Teacher Par Excellence twice by the management of SSN institutions. He received Distinguished Faculty (Multimedia and Image Processing) Award from Venus International Foundation Chennai, INDO GLOBAL EDUCATION EXCELLENCE AWARD from International Benevolent Research Foundation Kolkata,and Best Administrator award from PEARL Foundation Madurai. He has been recognized by Marquis Who's Who in the World for his contribution to the technical society and his biography has been published in its 25th Anniversary Edition