Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

Video Event Recognition Using Conditional Random Fields

R. Kavitha[1],D. Chitra[2], N. K. Priyadharsini[3]
[1]Assistant Professor, Department of CSE, P. A. College of Engineering and Technology, rkavitha.pacet@gmail.com
[2]Professor and Head, Department of CSE, P. A. College of Engineering and Technology, chitrapacet@gmail.com
[3]Assistant Professor, Department of CSE, P. A. College of Engineering and Technology,
priyadharsini.pacet@gmail.com

ABSTRACT

Event Classification in videos is a challenging task in computer vision based systems. The Crowd Event Classification system recognizes a large number of video events. The decisive of the model is a difficult task in the event classification. a more important role in the various research fields particularly in surveillance detection system. In the existing system it is done by using Deep Hierarchical Context Model which utilizes the contextual information from the feature extraction and prior level recognition of event in video. However, this research method might perform low with increased volume of videos and might failed to predict the events accurately with less interrelation contextual features. The new method namely Improved Hybridized Deep Structured Model (IHDSM) resolve the above problem. Here,introduce three different context features that describe neighborhood event. Here the Hybrid textual perceptual descriptor and concept based attribute extraction is performed for accurate recognition of video events. These extracted interaction context features are grouped by using improved k means algorithm. And then utilize the proposed improved deep structured model that combines convolutional neural networks (CNNs) and Conditional Random Fields (CRFs) to learn the middle level representationsand combine the bottom feature level, middle semanticlevel and top prior level contexts together for event recognition. This proposed research method is evaluated by using VIRAT data set whose simulation analysis is performed using matlab simulation toolkit. The overall evaluation of the proposed research method proves that the proposed method can provide better performance in terms of accurate recognition of events.

Keywords: Video event recognition, Deep structured mode, Convolution neural network, conditional random field, semantic level

I. Introduction

Automatic event detection in video streams is gaining attention in the computer vision research community due to the needs of many applications such as surveillance for security, video content understanding, and human–computer interaction[1]. The type of events to be recognized can vary from a small-scale action such as facial expressions, hand gestures, and human poses to a large-scale activity that may involve a physical interaction among locomotory objects moving around in the scene for a long period of time[2]. There also may be interactions between moving objects and other objects in the scene, requiring static scene understanding. Addressing all the issues in event detection is thus enormously challenging and a major undertaking[3].

Although progress has been made in the past few years, the current video event recognition systems often involve modules that are extremely expensive to compute, such as the extraction of spatial-temporal interest points [4]. Different from the previous works which focused mostly on recognition accuracy, to improve recognition speed while still maintain a good accuracy.

In this work, focus on the detection of large-scale activities where some knowledge of the scene (e.g., the characteristics of the objects in the environment) is known[5]. One characteristic of activities of interest is that exhibit some specific patterns of whole-body motion[6]. For example, consider a group of people stealing luggage left unattended by the owners. One particular pattern of the ‘‘stealing’’ event may be: two persons approach the owners and obstruct the view of the luggage, while another person takes the luggage. In the following, the words ‘‘event’’ and ‘‘activity’’ are used to refer to a large-scale activity[7].

From shape and trajectory featuresmodel scenario events using a hierarchical activity representation, where events are organized into several layers of abstraction, providing flexibility and modularity in modeling scheme[8]. The event recognition methods are based on a heuristic method and could not handle multiple-actor events[9]. In this work, an event is considered to be composed of action threads, each thread being executed by a single actor. A single-thread action is represented by a stochastic finite automaton of event states, which are recognized from the characteristics of the trajectory and shape of the moving blob of the actor[10]. A multi-agent event is represented by an

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

event graph composed of several action threads related by logical and temporal constraints. Multiagent events are recognized by propagating the constraints and the likelihood of event threads in the event graph. Various event recognition approaches were discussed in section II. In section III three levels of context and deep structured model configurations were discussed. Section IV described about the dataset and simulation results. Conclusion about the work was discussed in section V.

II. RELATED WORKS

Techniques for recognizing complex events in diverse Internet videos are important in many applications. Stateof-the-art video event recognition approaches normally involve modules that demand extensive computation, which prevents their application to large scale problems. In this section various related research methodologies has been discussed in detailed.

Izadinia et al. [11] fused six different low-level features, such as SIFT, STIP, GIST, together with 62 activity concepts ashigh-level features.Ramanathan et al. [12] used SIFT, MFCC and other low-level features together with 13 roles and 46 actions.Sun et al. [13] fused the motion feature with 60 activity concepts. It seems that dense trajectory feature is the single best feature, and other visual features complement each other.Wang et al. [14] propose a contextual feature capturing interactions between interest points in spatio-temporal domains from both local and neighborhood. Also, Zhu et al. [15] propose both the intra-activity and inter-activity context feature descriptors for activity recognition. At semantic level, context captures interactions among event and its components.

Gupta et al. [16] present a BN based approachfor joint action understanding and object perception. Yao et al. [17] utilize an MRF model to capture mutual context of activities, objects and humans poses. At prior level, the context captures the prior information of events. Here, the scene prior information is widely used for event recognition.Sun et al. [18] extract the point-level context feature, the intra-trajectory context feature and the inter-trajectory context feature, and combine the features using a multiple kernel learning model. These multiple level contexts are all in the feature level.Li et al. [19] build a Bayesian topic model to capture the semantic relationships among event, scene and objects. This model essentially captures the semantic level context, and incorporates the hierarchical priors in the model

Zhu et al. [20] exploit feature level contexts and semantic level contexts among events simultaneously through the structural linear model.Zeng et al. [21] build a multistage contextual deep model that uses the score map outputs from multi-stage classifiers as contextual information for the pedestrian detection deep model. However, both these two models are not designed to capture three levels of contexts, and are not for event recognition. As far as concerned, there is no existing event recognition research that simultaneously utilizes three levels of contexts through a deep probabilistic model.

III. DEEP STRUCTURED MODEL FOR ACCURATE VIDEO EVENT RECOGNITION

In the proposed research method improved Hybridized Deep Structured Model (DSM) is introduced. Here, first introduce three types of context features describingthe event neighborhood. Here the Hybrid textual perceptual descriptor and concept based attribute extraction is performed for accurate recognition of video events. These extracted interaction context features are grouped by using improved k means algorithm. And then utilize the proposedimproved deep structured model that combines convolutional neural networks (CNNs) and Conditional Random Fields (CRFs)to learn the middle level representationsand combine the bottom feature level, middle semanticlevel and top prior level contexts together for event recognition.

Here considered contexts in three levels. Those are feature level contexts, semantic level contexts, and prior level contexts.

3.1. Feature Level Contexts

Develop two types of context features including theappearance context feature and the interaction context featureextracted from the event neighborhood. This is done as like in the existing research method Deep Hierarchical

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

Context Model in [22].Suppose the event bounding box can be denoted as $\{(x_t,y_t,w_t,h_t)^T_{t=1}\}$ from frame 1 to T. (x_t,y_t) representsthe upper-left corner point. wt and ht denote the widthand height.

3.1.1. Appearance Context Feature

The appearance context feature captures the appearance ofcontextual objects, which are defined as nearby non-targetobjects located within the event neighborhood. Since event neighborhood is a direct spatial extension of the eventbounding box, it would naturally contain both the contextualobjects and the background.To efficiently extract andcapture the contextual objects from the background, utilizeHybrid textual perceptual descriptor.

In this context,propose a new descriptor describing the spatial frequency property ofsome perceptual features in the image. This descriptor has the advantageof being lower dimension vs. traditional descriptors as SIFT (60 vs 128),thus computationally more efficient, with only 5% loss in performance. Usually, spatial frequency is analyzed using spatial frequency descriptors. Thesedescriptors are based on image transform matrix as Fourier. Fourier transformis one of the most powerful descriptor in texture analyzing domain. For descriptor, a transform closely related to Fourier called Hartley transform $H_{k,l}$ isused: it contains the same information that Fourier does. In addition, contrary toFourier, it has the advantage of being a real function and this offer computationaladvantages in signal processing application.Hartley transformation matrix of size M ×M is computedas given:

$$H_{k,l} = \sum_{l,k=0}^{M-1} \left(\cos\left(\frac{2\pi lk}{M}\right) + \sin\left(\frac{2\pi lk}{M}\right)\right) (1)$$

where k, l ∈ {0,…,M-1}

In general, Fourier descriptor is blamed for not efficient in capturing a local features. Several researchers have proposed methods attempting to overcomethis drawback. According to Unser, the local texture property of an image regioncan be characterized by a set of energy measures computed at the output of a filter bank. In this context, Unser proposed an interesting way to exploit thespatial dependencies that characterize the texture of a region, more computationally efficient called local linear transform: it consists in computing foreach point of interest $x_{k,l}$, a local linear property $y_{k,l}$ as given in equation:

$$y_{k,l} = \sum_{l,k=0,..,M=1} T_M . x_{k,l} (2)$$

In equation 2 T_M represents an image transform matrix of size M ×M. In case, $T_M$represents Hartley transform $H_{k,l}$.Extending this method, each point of interest $x_{k,l}$ is tracked back in the Perreira'ssystem to multi-resolution feature pyramids computing phase. For each pyramidlevel $P_{t,\sigma}$, a neighborhood window $w_{xk,l}$ of the same size M × M as Hartleymatrix, is centered around $x_{k,l}$. In this method, $y_{k,l}$ represents texture energymeasures and it is defined by:

$$y^c_{k,l} = y_{k,l} = \sum_{l,k=0,...,M-1} \left[w_{x_{k,l}} . T^c_M \right]^2 (3)$$

where T^c_M is the convolution of each column with each rowin the matrix $H_{k,l}$ and c ∈ {1,…,M^2}. By combining the (M^2) channels as given in the equation 3, obtained $\left(M \times \frac{M+1}{2}\right)$ channels(noted TR) invariant to some rotation transformations:

$$TR_{k,l} = \frac{y_{k,l}+y_{l,k}}{2} (4)$$

To provide better visual perception, a contrast enhancement is applied usingequation 4, resulting an histogram f^t_l of dimension$\left(M \times \frac{M+1}{2}\right)$computedfor each multi-resolution feature pyramid level $P_{t,\sigma}$:

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

$$f_t = \frac{\log(\epsilon) - \log(TR_{k,l} + \epsilon)}{\log(\epsilon) - \log(\epsilon + 1)} (5)$$

where $i \in \left\{1, \ldots, M \times \frac{M+1}{2}\right\}$ and $\epsilon > 0$ is a suitably small value(here use $\epsilon = 0:05$).Concatenating these histograms for each point of interest $x_{k,l}$, the results ina descriptor of dimension($P \times M \times \frac{M+1}{2}$) with P=number of multi-resolutionpyramids. In evaluation, M = 3 and P = 10, resulting a descriptor of dimension 60.

3.1.2. Interaction Context Feature

The interaction context feature captures the interactions betweenevent objects and contextual objects as well as amongcontextual objects. The contextual objects are representedby the SIFT key points extracted in the event neighborhoodas discussed in Section 3.1.2. SIFT key points detectedwithin the event bounding box used to representthe event objects.Then, the Modified k-means clustering is applied to the 128 dimensionalfeatures of key points in both within the eventbounding box and event neighborhood of all training sequencesto generate a joint dictionary matrix D_l with K'words.

Modified K means clustering method is better in terms of efficiency and effectiveness. This algorithm works very well with large dataset images. It is based on iterative process. Cluster analysis is one of the major tools for exploring the underlying structure of a given data and is being applied in wide variety of engineering and scientific disciples such as medicine, psychology, biology, sociology, pattern recognition and image processing. Ahead of the performance of modified K means clustering, the properties of the clusters have to be recognized. This modified version of clustering overcomes the problem of parameter evaluation. This algorithm has certain additional properties than conventional clustering methods like ability to deal with noise, insensitive to the order of input records, capability to pact with variety of image types, scalability in case of both time and space.

The algorithm has the following steps.

1. Read the context features in to the MATLAB environment using the imread function.

2. Calculate the mean in every step.

3. Calculate the co-occurrence frequencies of words

4.Classifying the features using k means clustering label

5. Every pixel in the image using the results from k mean.

6. Create that feature groups based on cooccurrence values using cluster.

3.2. Semantic Level Contexts

The semantic level contexts stand for the semantic interactionsamong event entities. Since both the person andobject are two important entities of an event, the semanticlevel contexts for this work capture the interactions betweenevent, person and object.

A concept space C^K as an K-dimensional semanticspace, in which each dimension encodes the valueof a semantic property. This space is spanned by K conceptsC = $\{C_1, C_2, \ldots, C_K\}$. In order to embed a video xinto the K-dimensional space, Define a set of functions$\Phi = \{\Phi_1, \ldots, \Phi_k\}$, where Φ_i assigns a value $c_i \in [0, 1]$ to avideo indicating the confidence of the ith concept presencein it. The definition of Φ_i depends on the application. Notethat Φ_i is not necessary the concept detector 'i. If the conceptdetector 'i take the whole video as one single input,then can treat Φ_i andφ_ii same.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

Max Concept Detection Score(Max): This method selectsthe maximum detection score C_i^{max} over all slidingwindows as the detection confidence of detector i. Since the maximumdetection score provides information on the presence of aconcept, this feature is useful for some applications such asnovel event recognition.

Statistics of Concept Score(SCS): For some application,knowing the maximum detection score is not enough. Also need the distribution of the scores to model a specificevent.

Bag of Concepts(BoC): Akin to the bag of words descriptorsused for visual word like features, a bag of conceptsfeature measures the frequency of occurrence of each conceptover the whole video clip.

Co-occurrence Matrix(CoMat): A histogram of pair wiseco-occurrences is used to represent the pair wise presence ofconcepts independent of their temporal distance.

Max Outer Product(MOP): Since concepts represent semanticcontent in a video, the max value of each conceptacross the whole video represents the confidence in the presenceof a concept in a video.

3.3. Prior Level Contexts

The prior level contexts capture the prior information ofevents. Here utilize two types of prior contexts: the scenepriming and the dynamic cueing. The model can alsobe applied to other prior level contexts.

Scene priming. The scene priming context refers to thescene information obtained from the global image. It reflectsthe environment such as location (e.g. parking lot,shop entrance) and time (e.g. noon, dark) that can serve asprior to dictate whether certain events would occur.

Dynamic cueing. The dynamic cueing context providestemporal support for the prediction of the current eventgiven previous event. In this work, the previous event isrepresented by the K dimensional binary vector y_{-1} in the1-of-K coding scheme. Moreover, y_{-1} is further connectedto previous event measurement vector m_{-1} which denotesthe recognition measurement of the previous event.

3.4. Improved Deep Structured Models

Given the contexts in three levels as introduced previously,now discuss about the formulation of the proposedImproved Deep Structured Modelsthat combines convolutional neural networks (CNNs) and Conditional Random Fields (CRFs)for integrating them.

Here present the details of deep CRF model. One input image denotedby $x \in X$ and $y \in Y$ the labeling maskwhich describes the label configuration of each node in theCRF graph. The energy function is denoted by E(y, x,θ)which models the compatibility of the input-output pair,with a small output value indicating high confidence inthe prediction y. All network parameters are denoted byθ which need to learn. The conditional likelihood forone image is formulated as follows:

$$P(y|x) = \frac{1}{Z(x)} \exp[-E(y,x)] \quad (6)$$

Here Z is the partition function, defined as: $Z(x) = \sum_y \exp[-E(y,x)]$. The energy function is typically formulatedby a set of unary and pairwise potentials:

$$E(y,x) = \sum_{U \in u} \sum_{p \in N_U} U(y_p, x_p) + \sum_{V \in v} \sum_{(p,q) \in S_V} V(y_p, y_q, x_{pq}) \quad (7)$$

Here U is a unary potential function. To make the expositionmore general, consider multiple types of unarypotentials with U the set of all such unary potentials. N_U isa set of nodes for the potential U. Likewise, V is a pairwisepotential function with V the set of all types of pairwisepotential. S_V is the set of edges for the potential V. $x_p$and x_{pq} indicates the corresponding image regions whichassociate to the specified node and edge.The potential

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

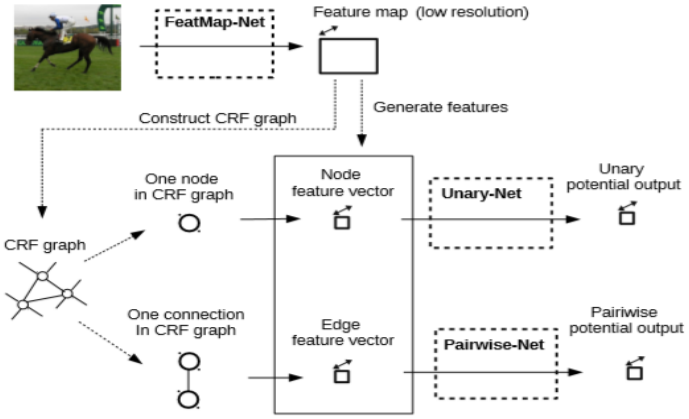
function is constructed by a deep networkfor generating feature map (FeatMap-Net) and a shallownetwork (Unary-Net or Pairwise-Net) to generate the outputof the potential function.

The unary potential function formulated by stacking theFeatMap-Net for generating feature maps and a shallowerfully connected network (referred to as Unary-Net) to generatethe final output of the unary potential function. Theunary potential function is written as follows:

$$U(y_p, x_p;\ \theta_U) = -z_{p,y_p}(x;\ \theta_U) \quad (8)$$

Here $z_{p,yp}$ is the output value of Unary-Net, which correspondsto the p-th node and the y_p-th class.Fig. 1 shows an illustration of the Unary-Net and howit corporates with FeatMap-Net. Fig. 2 demonstrates theprocess for generating the feature vector for one node. Theinput of the Unary-Net is the node feature vector extractedfrom the feature map which is generated by FeatMap-Net.The feature vector for one CRF node is simply the correspondingfeature vector in the feature map. The dimensionof the Unary-Net output vector for one node is K, which isthe same as the number of classes.

Figure1– An overview of the proposed contextual deep structured model. Unary-Net and Pairwise-Net are shown here forgenerating potential function outputs



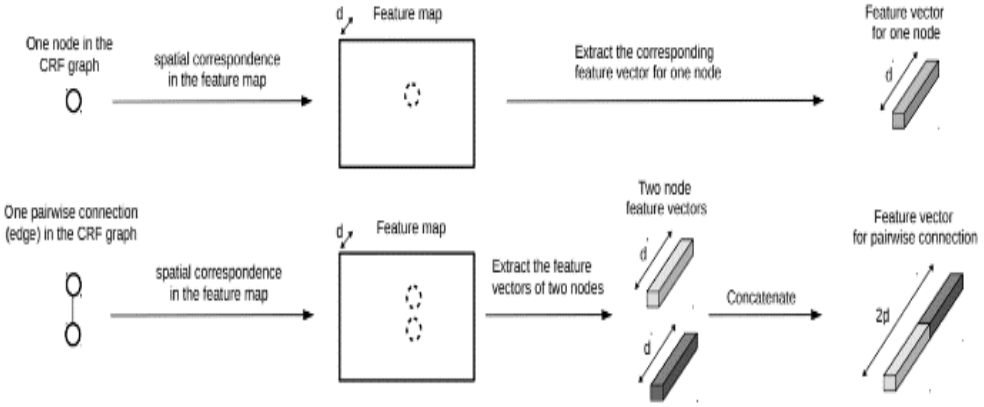


Figure 2– An illustration of generating feature vectors for CRF nodes and pairwise connections from the feature map outputby FeatMap-Net. The symbol d denotes the feature dimension. The corresponding features of two connectednodes in the feature map are concatenatedto obtain the CRF edge features.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

IV. RESULTS AND DISCUSSION

The effectiveness of the proposed appearanceand interaction context features are evaluated. The experiment is performed on the VIRAT 2.0Ground Dataset with the six person-vehicle interactionevents. The baseline event feature is the STIP extractedfrom event bounding box.

4.1. DescriptionDataset

The first portion of the video dataset consists of stationaryground camera data. Collected approximately25 hours of stationary ground videos across 16 differentscenes, amounting to approximate average of 1.6 hours ofvideo per scene. The snapshots of these scenes are shown inFig. 3, which include parking lots, construction sites, openoutdoor spaces, and streets. These scenes were selectedbased on the observation that human and vehicle events occurfrequently in these areas. Multiple models of HD videocameras recorded scenes at 1080p or 720p to ensure that obtain appearance information from objects at distance, andframe rates range 25~30 Hz. The view angles of camerastowards dominant ground planes ranged between 20 and 50degrees by stationing cameras mostly at the top of buildingsto record large number of event instances across area whileavoiding occlusion as much as possible. Heights of humanswithin videos range 25~200 pixels, constituting 2.3~20% ofthe heights of recorded videos with average being about 7%.

In terms of scene diversity, only two pairs of scenes (total4) among 16 scenes had FOV overlap, with substantial outdoorillumination changes captured over days. In addition,virat dataset includes approximate homography estimates forall scenes, which can be useful for functions such as trackingwhich needs ground coordinate information.Most importantly, most of this stationary ground videodata captured natural events by monitoring scenes overtime, rather than relying on recruited actors. Recruitedmulti-actor acting of both people and vehicles was involvedin the limited subset of 4 scenes only: total acted scenes areapproximately 4 hours in total and the remaining 21 hoursof data was captured simply by watching real-world events.Originally, more than 100 hours of videos were recorded inmonitoring mode during peak activity hours which includemorning rush hour, lunch time, and afternoon rush hour,from which 25 hours of quality portions were manually selectedbased on the density of activities in the scenes.

4.2. Simulation Comparison

In this research, proposed method has been implemented and evaluated in the matlab simulation environment. Here varying set of training videos are taken out which would learned together to learn the different feature variation present among the videos of different kinds. In this work, the training videos taken virat dataset. These videos would be learned accurately for their features presence based on which final outcome would be made. The proposed system is implemented using MATLAB 2013a and the experimentation is performed with i5 processor of 3GB RAM. The performance metrics that are considered in this research method for the efficient implementation of the proposed and existing research methodologies are listed as follows:

- Accuracy
- Sensitivity
- Specificity
- Precision
- Recall
- F-Measure

The evaluation of the proposed method Improved Hybridized Deep Structured Model (IHDSM) based on these performance metrics is done by comparing it with the existing research method namelyDeep Hierarchical

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

Context Model (DHCM). The numerical evaluation of the proposed research method is conducted by comparing it with the existing research method which is shown in the following figure 3.

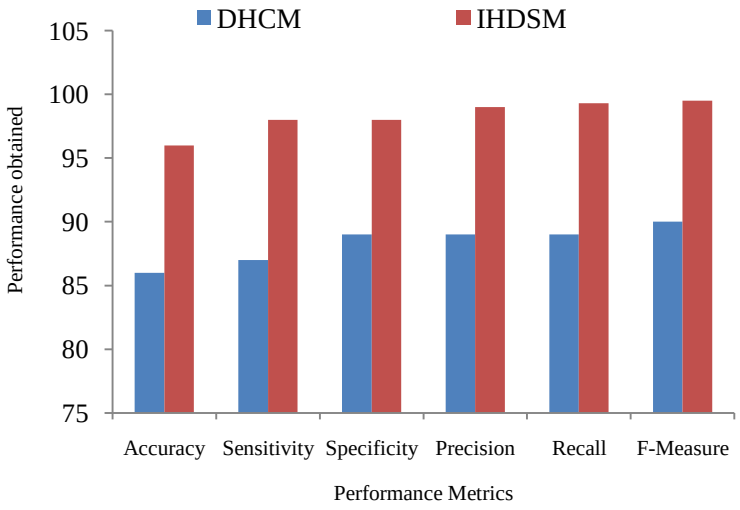


Figure 3. Numerical comparison outcome

From the figure 3 it can be concluded that the proposed research method leads to provide the improved performance than the existing research method by accurately retrieving the similar videos from the training database. From this outcome it can learnt that the proposed method IHDSM shows 11% improved performance ratio than the existing research methodologies in terms of accurate retrieval of videos.

V. CONCLUSION

In the proposed research method improved Hybridized Deep Structured Model (IHDSM) is introduced. Here, first introduce three types of context features describingthe event neighborhood. Here the Hybrid textual perceptual descriptor and concept based attribute extraction is performed for accurate recognition of video events. These extracted interaction context features are grouped by using improved k means algorithm. And then utilize the proposedimproved deep structured model that combines convolutional neural networks (CNNs) and Conditional Random Fields (CRFs)to learn the middle level representationsand combine the bottom feature level, middle semanticlevel and top prior level contexts together for event recognition. This proposed research method is evaluated by using VIRAT data set whose simulation analysis is performed using matlab simulation toolkit. The overall evaluation of the proposed research method proves that the proposed method can provide better performance in terms of accurate recognition of events.

REFERENCES

1. Awad, G., Fiscus, J., Michel, M., Joy, D., Kraaij, W., Smeaton, A. F., &Ordelman, R. (2016). TRECVID 2016. Evaluating Video Search, Video Event Detection, Localization and Hyperlinking.
2. Edwards, M., Deng, J., &Xie, X. (2015). From Pose to Activity: Surveying Datasets and Introducing CONVERSE. *arXiv preprint arXiv:1511.05788*.
3. Battaglia, P., Pascanu, R., Lai, M., & Rezende, D. J. (2016). Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems* (pp. 4502-4510).
4. Jiang, Y. G., Dai, Q., Mei, T., Rui, Y., & Chang, S. F. (2015). Super fast event recognition in internet videos. *IEEE Transactions on Multimedia*, *17*(8), 1174-1186.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6565 – 6573
Received 05 March 2021; Accepted 01 April 2021.

5. Heilbron, F. C., Escorcia, V., Ghanem, B., &Niebles, J. C. (2015, June). Activitynet: A large-scale video benchmark for human activity understanding. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (pp. 961-970). IEEE.

6. Frost, D. M., Beach, T. A., Callaghan, J. P., & McGill, S. M. (2015). FMS Scores Change With Performers' Knowledge of the Grading Criteria—Are General Whole-Body Movement Screens Capturing "Dysfunction"?. *The Journal of Strength & Conditioning Research*, *29*(11), 3037-3044.

7. Kousalya, R., & Dharani, S. (2017). Multiple Video Instance Detection and Retrieval using Spatio-Temporal Analysis using Semi Supervised SVM Algorithm. *International Journal of Computer Applications*, *163*(4).

8. Gaidon, A., Harchaoui, Z., & Schmid, C. (2014). Activity representation with motion hierarchies. *International journal of computer vision*, *107*(3), 219-238.

9. Onofri, L., Soda, P., Pechenizkiy, M., &Iannello, G. (2016). A survey on using domain and contextual knowledge for human activity recognition in video streams. *Expert Systems with Applications*, *63*, 97-111.

10. Kale, G. V., & Patil, V. H. (2016). A study of vision based human motion recognition and analysis. *arXiv preprint arXiv:1608.06761*.

11. Izadinia H. and ShahM., "Recognizing complex events using large margin joint low-level event model," in Proc. ECCV, 2012, pp. 430–444.

12. RamanathanV., LiangP., and Fei-FeiL., "Video event understanding using natural language descriptions," in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 905–912.

13. SunC. and NevatiaR., "ACTIVE: Activity concept transitions in video event classification," in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 913–920.

14. WangJ., ChenZ., and WuY. Action recognition with multiscale spatio-temporal contexts. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3185– 3192, June 2011.

15. ZhuY., NayakN., and Roy-ChowdhuryA..Context-aware modeling and recognition of activities in video. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2491–2498, June 2013

16.Gupta A.and DavisL. Objects in action: An approach for combining action understanding and object perception. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, June 2007

17. Yao B. and Fei-FeiL. Modeling mutual context of object and human pose in human-object interaction activities. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 17–24, June 2010.

18. SunJ., WuX., YanS., CheongL.-F., ChuaT.-S. and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2004–2011, 2009.

19. LiL.-J. and Fei-FeiL.. What, where and who? classifying events by scene and object recognition. In IEEE International Conference on Computer Vision (ICCV), pages 1–8, Oct. 2007

20. ZhuY., NayakN. and Roy-ChowdhuryA. Context-aware modeling and recognition of activities in video. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2491–2498, June 2013.

21. ZengX., OuyangW., and WangX. Multi-stage contextual deep learning for pedestrian detection. In IEEE International Conference on Computer Vision (ICCV), pages 121– 128, 2013.

22. Wang X., & Ji, Q. (2015). Video event recognition with deep hierarchical context model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4418-4427).