Multiple Action Recognition for Human Object with Motion Video Sequence using the Properties of HSV Color Space Applying with Region of Interest

N. Kumaran, U. Srinivasulu Reddy, S. Saravana Kumar

Abstract: ***Recognizing the activities of human beings from sequence of images is an on-going area of research in computer vision. Deep Convolutional Neural Network (CNN) models, recently started studies in recognising action of human together with categorization of images. However, these models are excessively encouraged by the background of the image, as well as understanding the existence of clues in standard computer vision datasets. For different robotics practice, the amount of innovation and variation activity backgrounds is too larger than PC vision datasets. To tackle this issue, we present the approach called "Action Region Proposal (ARP)". In this approach image regions and optical flow are acting as the informing agents which are probably to contain movements for giving contribution to the system at the training just as testing phase. The proposed sub-activity descriptor comprises of three stages: the posture, the locomotion, and the gesture. The proposed activity detection classic localizes and perceives the activities of various persons at the same time in video surveillance by utilizing appearance-based temporal features with multiple CNN. Through a series of examinations, we have demonstrated that the manual background segmentation is insufficient; However the active ARP at the time of training and testing enhances the execution on specific spatial and temporal video constituents. Finally, the study indicates that by focusing attention through ARP, the performance of the present up-to-date spatio-temporal action recognition can be improved.***

Index Terms: ***sub-action descriptor, action detection, video surveillance, convolutional neural network, multi-CNN.***

I. INTRODUCTION

One of the active researches carried out in the branch of PC vision is the automatic recognition of individual activities, since it has several real time applications in different fields. In the real world scenarios human action recognition by mobile robots has become a challenge and has been recently focussed in the robotics domain. Recognizing the human action by the mobile robots could **increase** the standard of these robots in identifying their subsequent job in hand or make the robots aware of the suspicious actions happening around the environment. It also, enables the automatic recognition of actions in surveillance systems to help alert, retrieve, and summarize the data [1], [2], [3]. Vision-based action recognition- recognition of semantic spatial–temporal visual patterns like *walking*, *running*, *texting*, and *smoking*, etc., are core computer vision issues in video surveillance [4]. Number of the progresses in surveillance have been possible owing to the accessibility of open datasets like KTH [5], Weizmann [6], VIRAT [7], and TRECVID [8] datasets. However, these existing datasets are well suited for the existing state-of-the-art surveillance systems. where actions are in restricted scenes and some unexpected surveillance record tends to be repetitive, more often dominated by the scenes of walking people. There is a need for a new video surveillance dataset for the stimulation of the development. In this study, we have used a large-scale video surveillance dataset created to evaluate the efficiency of recognizing an action and localizing the equivalent space–time capacity from a lengthy and a continuous video which includes detection, tracking and action recognition of human beings. One of the difficulties that robotics encounter in using these systems on independent real-time robots is the practical imageries that are naturally more varied and impartial than computer vision datasets [9]. This phenomenon can be observed in movement recognition, that comprises of traditional datasets with situation based informative backgrounds (refer figure 1. For standarized shot angles). The study is induced by the necessity to develop a commonly implementable action recognition system that works irrespective of context, platform, and background. The ultimate goal of this research is to create a systematized methodology whose target is on the parts of the image where actions are probable to occur in the training as well as in the testing phase.The first phase employs optical flow to recognize areas in which activity occurs and to give ARP inside the spatial images. In the second phase, Instead of considering the full images, the up-to-date network architecture can be trained only on those proposed regions. Finally, action recognition is done by producing the final classifier through merging the learnt spatial and temporal characteristics. In this paper, a number of experiments were conducted by comparing the performance with the present up-to-date systems. Experiments were conducted with the region proposal method including the control case for manually removing the background. The aim of this paper is to create a real-time high performance action identification algorithm on the basis of CNN. This is a big challenge, since tracking by detection and action recognition are computationally expensive and it is difficult to estimate together in practical. There are many works to evaluate human posture [10], [11], [12] and analyse gesture data [13] in real-time. [14] proposed a real-time CNN based action recognition method.

Revised Manuscript Received on April 07, 2019.

N. Kumaran, Research Scholar, Dept. Computer Applications, N.I.T, Tiruchirappalli-15.Tamil Nadu, India.

Dr. U. Srinivasulu Reddy, Assistant Professor, Dept. of Computer Applications , N.I.T, Tiruchirappalli-15. Tamil Nadu, India.

Dr. S. Saravana Kumar, Professor, Dept. of CSE, Shanmuganathan College of Engineering, Pudukkoattai. Tamil Nadu, India.

Figure 1. Some challenging video snapshots from UCF101 dataset to *recognize human actions.*

Figure 2: Complete structure of the proposed real-time activity detection model. Appearance basis transient features of the regions of interest (ROIs) are nourished into three CNNs using movement detection, human detection, and multiple tracking procedure. These CNN create forecasts by means of shape, movement history, and their consolidated cues. Each action of an individual has three sub-action categories, under the sub-activity descriptor, which conveys a total arrangement of data about a human activity.

The objective of this work is to construct a unique model that can simultaneously localize and perceive numerous activities of individuals in a video surveillance system. The full view of the proposed real-time action detection scheme is illustrated in figure 2. In the training phase, the Region of Interest (ROI) and the sub-activity interpretations are first determined physically in every frame of the training videos. Every single level of the sub-action descriptor has one CNN classifier. In the multi-CNN model, each prediction of a CNN is equal to one sub-action. In the testing phase, a movement saliency field is created utilizing an optical flow to eliminate regions which are improbable to have the action that further reduces the processed regions. The traditional sliding window based pattern is utilized as a mask in the movement saliency field. A human detection histogram of oriented gradient (HOG) descriptor [15] with a latent support vector machine (SVM) [16] is utilized in the sliding window to distinguish people in an underlying activity in the ROIs. At that time, Kalman filter is used to refine the activities in the region. Three sub-activity classes are predicted from the given refined action in the regions of interest using shape, movement history, and their consolidated cues with the assistance of the CNNs. At last, the post-processing phase examines the clashes in the arrangement of the sub-activity descriptor and performs a temporal smoothening as indicated by the historical activities of every person to reduce noise. The rest of the paper is organized as follows: Section 2, reviewed the existing research work related to action recognition. Next, in section 3, the outline of our approach and the framed network training details are described.

Implementation outputs are reported in Section 4 and Section 5 finishes up the paper with future guidelines.

II. RELATED WORK

Numerous applications like video surveillance, human-PC cooperation and video content retrieval require identifying and detecting human action from videos.

In human action recognition, extracting image features from a video clip and issuing the corresponding label for that action class for them is considered to be a common approach. There exist several researches on recognising human action [17],[18],[19],[20],[21] which can be sorted into two principle collections: 1) Handmade local features and a number of visual words representation and 2) Well-trained feature descriptors. Both groups are confirmed for providing excellent results in recognition of human actions.

Various researches have been conducted on activity analysis, and recent surveys related to the area are found in [22], [23]. Estimating a human pose using a predefined model (e.g., pictorial structures) in each frame is the most familiar technique for generic human model recovery. The model is driven by an attempt to minimize the cost function between the collection of parts organized in a deformable pattern and human contours [24], [25]. Because of the truncation and occlusion of body parts, pose estimation from a 2D image is a complicated process. Major enhancements in depth imaging and 3D input data have recently been developed.[26], [27] estimated parts of the human body and the 3D positions of every skeletal joint directly from a single depth image using a 3D sensor. However, a human action goes further than a human pose, and after human pose information is obtained, specific classifiers or decision rules are needed to recognize actions.

Many of the approaches introduce an appearance-based method where an action comprises a sequence of human silhouettes or shapes. In contrast to human model recovery, this method uses appearance-based temporal representations of static cues in a multitude of frames, where an action is expressed by a series of two-dimensional shapes. Motion History Image (MHI) and Motion Energy Image (MEI) [28], [29] are the most pervasive appearance dependent temporal features. The benefits of the approaches are: these methods are modest, rapid, and work efficiently in well-ordered locations, e.g., the surveillance video background obtained through a top vision camera is the ground at all times. The critical defect in MHI is that it is impossible to detect inner most movements; it can simply catch the person silhouettes [30]. However, the effect of shape and motion history cues with CNN for action recognition has not been examined carefully. In this work, a new technique for encoding those temporal characteristics is proposed, and a review of how different appearance dependent temporal characteristics influence execution is given. There are some other appearance dependent temporal strategies such as the learned dynamic prior classic, the active shape classic and the prior movement classic. Furthermore, movement is reliable and simply described by a well-defined space–time track in few characteristics locations. Certain methods utilize movement paths such as, conventional and parametric optical flow of previously mentioned human locales or body target points to perceive activities by means of visual tracking [31], [32].

Recent researches applied shape-based features such as HOG [33], SIFT [34] and action reliant qualities like optical flow, MBH [35] using trained classifiers (e.g. SVM, decision forests) and also concentrated on higher level coding like Bag of Words, Fischer vectors to predict actions. Hand-made descriptors can't be optimized for visually representing actions and lacks definite capacity for recognizing action even though they have shown good performance in some cases.

Deep learning systems are a division of machine learning procedures which can be trained by a group of characters through the creation of top-level characters from low-level characters. These inspiring benefits of CNN model on image classification chore [36], researchers focus more of their effort on CNN models that solves action recognition [37],[38],[39],[40],[41]. Recently proposed techniques such as Convolutional RBMs [42], 3D CNNs [43], RNN [44],[45], CNNs [46] and Two-Stream CNNs [47] have proposed significant evidences to this area.

III. OVERVIEW OF THE APPROACH

A. Action region segmentation of video frames from the HSV Color space

3.1 Properties of HSV color space

A hexagon is 3-D illustration in the HSV color space, with dominant color and the sensitivities of our eyes are low to its deviations when comparing it with deviation in intensity or hue. The pixel point saturation determines whether the intensity or else the hue is progressively relevant to our ocular sensitivity of that pixel color and ignores the real incentive for saturation. For minor saturation, a color can be estimated through a grey esteem indicated by the intensity vertical axis. Hue can be stated as an angle between 0 and 2π in the red axis with red at point 0, green at $2\pi/3$, blue at $4\pi/3$ and red again at 2π. The clarity of color is called as saturation and it is calculated by the outspread space measuring between 0 and 1 assuming 0 at the central axis and 1 at the external surface respectively. The HSV space convert any color into grey-color by appropriately diminishing the saturation. The particular gray shade to which the conversion has to be converged is determined by the intensity value. Saturation defines the level of depth for greater saturation; its hue is used to estimate the color. The saturation threshold which decides this change is one more time reliant on the intensity. Smaller intensities for a higher saturation indicates that a color is nearer to the grey color and the other way around. The differences in-between the hue and intensity strength determined by a saturation is 0.2 at a larger intensity value. The most extreme intensity esteem is assumed as 255, the threshold function given below is utilized to choose whether a pixel is signified by hue or intensity similar to its principal characteristics.

$$th_{sat}(V) = 1.0 - 0.8V / 255 \quad (1)$$

Thus, if the pixel in an image has higher saturation than th_{sat} (V) and hue as the prevailing constituent then it is preserved as real color pixel. Similarly, if the pixel has smaller saturation and has intensity as the prevailing constituent then they are considered as a grey-shade pixel. This novel concept of splitting pixels of real-color from grey-color utilizing saturation achieves the image separation which is useful for tracking of objects. The big limitation of a large portion of the pixel space object tracking strategies is its sensitiveness to force changes that is decreased to greater degree. Similarly, video frames that are transiently closer to each other have great object stage likeness if it does not have an interceding snap limit. Thereby, an object level depiction of the video planes is found to give more strong method for comparison of objects for tracing [48]. Hence, this technique is comparable to the way a person sees the available object and captures variations in video. However, person eyes see a variation in object motion simply if the current object in a frame differs extensively from its past edge

B. Selecting Action Region Proposals

Choosing an ARP is a difficult chore when matched to object proposals. This is because of both the look and gesture cues require to have effective action region proposals. In contrast, object proposals are purely relied on ocular look information. In addition, thinking about the assortment of human activities, human actions cannot be easily differentiated from background and other dynamic motions [49].

Our proposed method isolates parts of interest and ARP at the frame level. Figure 3 illustrates the produced action region proposals of the applied technique on specimens of UCF101 dataset video frames. Our study focuses on Edge Boxes technique [50] which is already proved to perform well for detecting objects [51] and thereby a strategy is proposed in order to select a greatest activity proposal. Here, we have somewhat modified the Edge Boxes technique that detects appropriate action regions. Moreover, video frames are first extracted and motion is represented by using optical flow signals [52].

Applying Edge Boxes on optical flow pictures, a huge amount of potential bouncing boxes in an image are observed that must be provided with a score for the particular process of activity recognition. As the outcome of this approach, we provide result for each of the bouncing box dependent on the greatness of optical flow motion inside the box. That is, the result of each box is computed by the standardized size of the optical flow waveform, which might be taken as a temperature map at the pixel level [53]. The score function is:

$$OF(S) = \frac{1}{S}\sum_{i \in \mu} OF(i) \qquad (2)$$

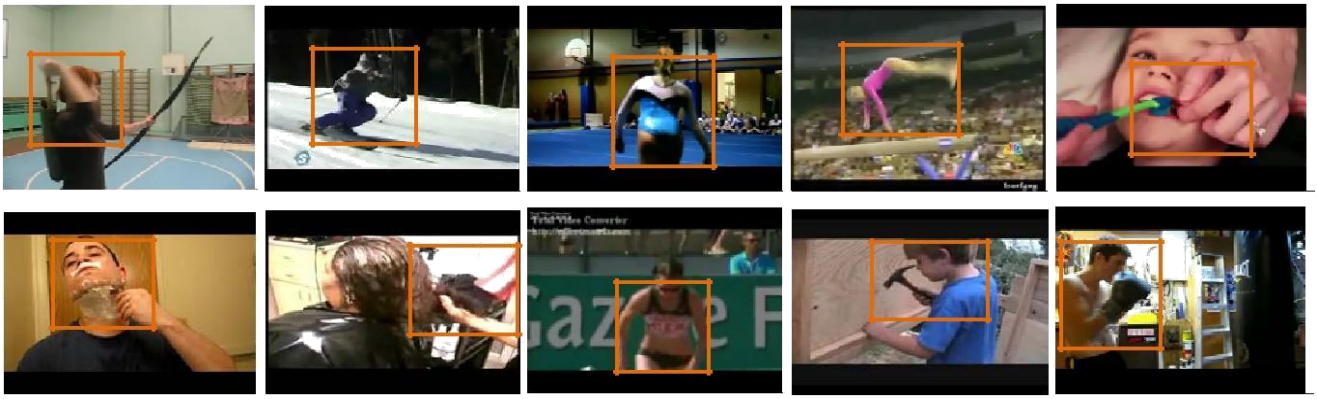

Figure 3. Snapshots of generated bounding boxes using Edge boxes method to detect action area.

Where S represents a bouncing box. S can be discarded when $OF(S) \leq \delta$, this δ can be provisionally obtained. For the experiment purpose, the value of δ can be selected as 0.32 and Edge Boxes parameters are set to the default values.

C. Human Activity Detection

The important objective of the proposed approach is to detect the actions of multiple individuals for real-time surveillance applications. The human action regions are detected by a frame-based human detector and a Kalman tracking algorithm. The action classifier is composed of three CNNs that operate on the shape, movement history and their cues combination. According to the sub-action descriptor, the classifier predicts the regions to produce three outputs for each action. The outputs of the classifiers go through a post-processing step to render the final decisions.

3.3.1 Sub-Action Descriptor

The problem of representing an action is not well-defined as a measurement problem of geometry (e.g., measurement of an image or camera motion). Intra-class changes of the activity class is equivocal, as illustrated in Figure 4 (a) and (b). Despite the fact that the activities of the three people are *texting* in Figure 4(a), would you be able to tell if what they are doing is exactly the same? The first person is *texting while sitting*, the second person is *texting while standing* and is *stationary*,

and the third person is *texting while standing* and *walking*. For the above three persons, giving the same action representation (*texting*) is often confusing—all of them have different postures and locomotion states for the similar activities. These are the equivalent problems for the activities of *smoking* showed in Figure 4(c).

To deliver complete information about human actions, and to clarify action information, the proposed approach in this paper models an action with a sub-action descriptor. A depiction of the sub-action descriptor is displayed in Figure 4(c). The descriptor includes three stages: the posture level, the locomotion level, and the gesture level. The posture level comprises the following two sub-actions: *sitting* and *standing*. The locomotion level comprises *stationary*, *walking*, and *running*. The gesture level comprises *nothing*, *texting*, *smoking*, and *others* (e.g., *phoning*, *pointing*, or *stretching*, which are not considered).

The connecting line between two sub-actions at different levels indicates that the two sub-actions are independent of each other. No connection indicates an incompatible relation where the two sub-actions cannot happen together. Each level has one CNN; therefore, for one action of an individual, three CNNs work simultaneously. The first network, BDI-CNN, functions on a stationary cue and snaps catches the person's silhouette. The second system, MHI-CNN, functions on a cue movements and catches the historical backdrop of the person's movements. The third system, WAI-CNN, functions on a combination of stationary and movement cues and catches the inconspicuous activity forms. The designed descriptor of actions can transform the difficult problem of action recognition into many easier problems of multi-level sub-action recognition. It is inspired by huge quantities of activities that can be constructed by not many autonomous sub-activities. Here, three stages are consolidated to denote numerous kinds of activities with an enormous measure of
opportunity.

3.3.2 Failure Detection and Recovery

In Figure 4(c), posture level objects are static, but camera motion are leading to action region are occluded, when it reaches the angle from 1810 to 3600, during that period, object action may disappear. Object occlusion or failure may occur. To overcomes this failure and recovers the object action recognition that we will maintained for the

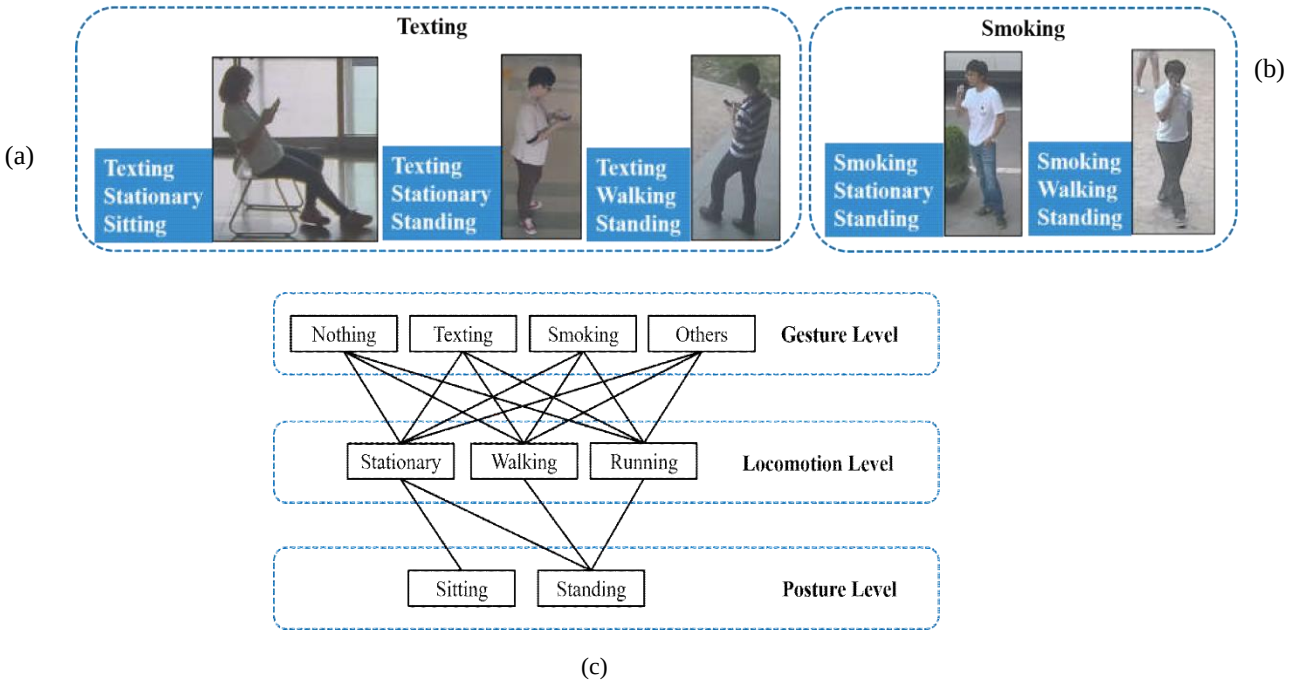

Figure 4: Conventional representation problems and sub-action descriptor. (a) Conventional representation problem: *texting*, (b) conventional representation problem: *smoking*, and (c) structure of the sub-action descriptor. A sub-activity descriptor includes three stages: the posture level, the locomotion level, and the gesture level. Each level has one CNN; therefore, for one action by an individual, three CNNs work simultaneously.

same object action of previous frame before lost object until action may reappear, within 10 sec

should be 20-30 *mms* for every frame, a static bounding box for the person activity location, and a small false detection

3.3.3 Tracking by Detection

The main goal of this paper is real-life activity recognition in surveillance video. For the human activity discovery and tracking algorithm, we adopt current approaches so as to give a steady human activity space for consequent activity recognition. The essential factors of person identification and tracking algorithm are: the operating period which

rate.

Figure 5: Mini motion map for minimizing the undesirable calculation in HOG dependent person detection. (a) Original image with a size of 640 × 360, and (b) mini motion map with a size of 77 × 34, which was estimated from GMM-base movement detection

Computational efficiency of the surveillance application is critical. Generally, several windows does not encompass any object and therefore, the operating time of the object detection is restricted by sliding window. Hence, the movement detection is employed prior to object detection procedure in order to dispose of locales which have null movement. A mini motion map is generated [54] by utilizing a Gaussian mixture model–based motion detection procedure [55]. The mini motion map is rescaled to the same number of width × height sliding windows, such that minimum computation is required when determining whether a sliding window is a foreground or a background. The mini movement map dimension is calculated using the given expression:

$$\boldsymbol{size}_{mni-map} = \frac{\boldsymbol{size}_{original} - \boldsymbol{size}_{detection}}{\boldsymbol{stride}} \quad (3)$$

In HOG, the default esteem of $\mathbf{size}_{\text{detection}}$ and **stride** are (64, 128) and (8, 8) respectively [15]. Figure 5 illustrates the mini movement map. For example, when the input image size is 640 × 360, then the mini movement map dimension is 77 × 34.

For the classification, latent SVM [16] with a fixed C = 100, as recommended elsewhere, is used to classify the HOG to detect humans. The next step is to associate the resulting detections with the corresponding tracks [54]. Humans appear and disappear at random times; thus, three types of situations present in the information affiliation issue: 1) add another track, 2) update the current track, and 3) delete a path [55]. The system for dealing with different detection and tracks is appeared in Figure 6. At the point when another track is included, it begins to calculate the quantity of frames in which track has refreshed without identification. In this way, even when a motion is not

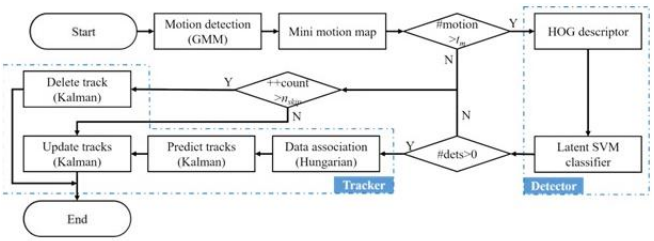

Figure 6: Procedure for multiple detections and tracks.

actor in 2D frames, referred as $b(x, y, t)$, and it can be detected in some frames, the track still updates according to the previous prediction. When this value is bigger than that of the threshold n_{skip}, the track is to be vanished or erased.

3.3.4 Feature extraction and post-processing

Next to the activity detection, feature extraction is intended with in this proposed system. Each frame is characterised as a group of pixel characteristics as given below:

$$F \equiv \{(pos, [r/g], val)\} \quad (4)$$

Here every pixel is depicted as a trio where "pos" means the spatial position (x, y) of the pixel, "[r|g]" indicates whether the pixel is a "real" or "grey" in shade and "Val" indicates real or grey esteem. Hence, val ∈ [0, 2π], when [r|g] is given 'r' and val ∈ [0, 1] when [r|g] is given g.

The characteristics of a pixel is the duo ([r|g], Val) to identify real-color or a grey shade pixel and the matching value of hue or intensity. Once every pixel is represented by trio form, K-means clustering method can be applied in order to group the target objects with identical feature values. It is a rapid process because the grouping is employed simply to an intent object. Here the intent object can be of any "real color" or "grey color" and it contains every pixels of the intent object with colors nearer to "Val"

3.3.5 Appearance-Based Temporal Features

Appearance basis temporal features extract static information from a multitude of frames, in which an activities are depicted through a sequences of two-dimensional shapes. The characteristics are extremely basic and quick, and they function very well in constrained conditions, like surveillance frameworks where the cameras are placed on housetops or tall shafts in order to make the cameras vision towards the overwhelming ground planes. Consider, a video F as a function of only three variables for the current analysis.

$$F = f(x, y, t) \quad (5)$$

Here, the coordinate $(x, y, t) \in R^3$ is the Cartesian coordinate of the video space. Every level of a sub-action descriptor contain single autonomous CNN which acquires various appearance basis temporal characteristics. The BDI feature accurately captures the still shape cue of the

calculated by Eq. 3:

$$(x, y, t) = \begin{cases} 255, if\ f(x, y, t) - f(x, y, t_0) > \xi \text{thr} \\ 0, \quad otherwise \end{cases} \quad (6)$$

where the values in the BDI are set to 255 when the change of the present frame $f(x, y, t)$ from the background frame $f(x, y, t0)$ of the input video is bigger than a threshold ξthr, and x and y are indices of the image domain. BDI is a binary image that indicates the silhouette of the posture.

Examples are given in Figure 7.

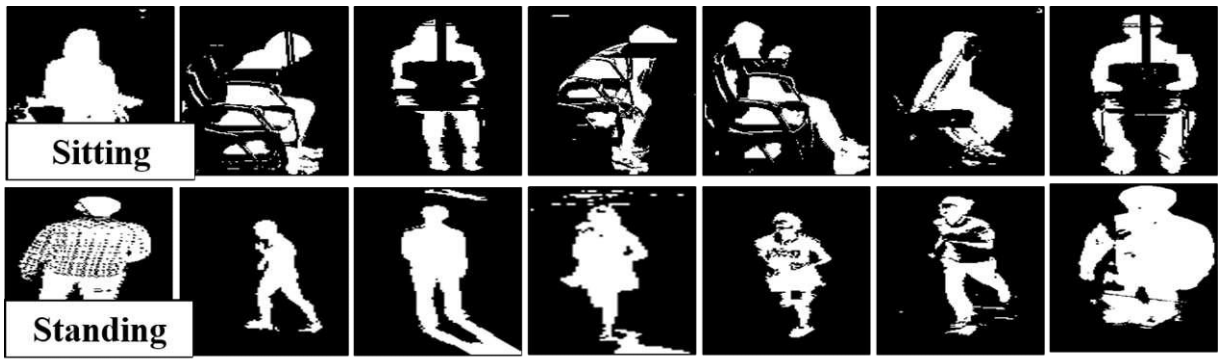

Figure 7: Samples of BDI for various sub-actions. BDIs are utilized for the posture level of the sub- activity descriptor that includes *sitting* and *standing*. BDI catches the static shape cue of the actor.

In a historical movement image, the operation of the past temporal movement is denoted as intensity of pixel at that location. MHI catches the past movement forms of the actor, represented as $h(x, y, t)$, and is characterized utilizing a straightforward substitution and decay operator in Eqs. 7-8 [28]

$$d(x,y,t) = \begin{cases} 255, if\ f(x,y,t) - f(x,y,t-1) \xi \text{thr} \\ 0, \quad otherwise \end{cases} \tag{7}$$

$$h(x,y,t) = \begin{cases} \tau_{max}\ \ if\ d(x,y,t) = 255 \\ \max\left(0, h(x,y,t-1) - \Delta\tau, otherwise\right. \end{cases} \tag{8}$$

$$\Delta\tau = \frac{\tau_{\max} - \tau_{min}}{n} \tag{9}$$

MHI is used for the locomotion level, which comprises stationary, walking, and running. It is produced due to the distinction of the present frame f(x, y, t) and the preceding frame f(x, y, t-1) in Eq. 7. The MHI at time t is estimated from the outcome of the preceding MHI for each frame. Thus, these temporal features need not be computed once more. MHI represents an image movement vector and the space which contains the most recent movements are illustrated as brighter in Figure 8. The number of frames to be taken as the activity history limit is denoted as n in Eq. 9. The hyper-parameter n is the basic parameter to characterize the temporal scope of an activity. The higher n value of an MHI describes a large scope of activity history; but, it is unresponsive to present activities. In the same way, smaller n of MHI places the attention on the latest activities and disregard the historical activities. Thus, picking a best n can be genuinely troublesome.

D. Training on Region Proposed Images

3.4.1 Network Training

In this section, we explain 'ActPropNet' the procedure to train our deep activity proposal system. Our approach is trained using UCF101 dataset [55]. UCF101 is an activity recognition dataset which contains real-time action videos, collected from YouTube that has hundred and one action categories[71]. This dataset is publicly available that contains 13320 video clips divided into three categories as training data, testing data and evaluation data. We presented the mean of acquired precisions on these three categories as the end exactness in tables.

An arbitrary frame is chosen for every video cut, and then the flat (flow_x) and perpendicular (flow_y) optical flow signals are calculated from two successive frames. The impact of universal movements among the frames is reduced by applying mean subtraction.

As specified before, the study employs an ARP to minimize the image background in which the activity is happening. The proposed activity region proposal approach is employed on optical flow images that further utilize the bounding boxes for spatial images also. Certain modern techniques, separate the portions of image by arbitrarily cropping the entire image [51,52, 53, 54] while we pick up portion of images more suitable for motion-based tasks which are notified by optical flow. Subsequently region suggested images can be given as input to Caffe framework after resized to 224×224 [70].

Both the optical flow signals (horizontal and vertical) are stored in HSV format and 3rd channel is created by the amount of optical flow signals and are linearly rearranged to 255 range.

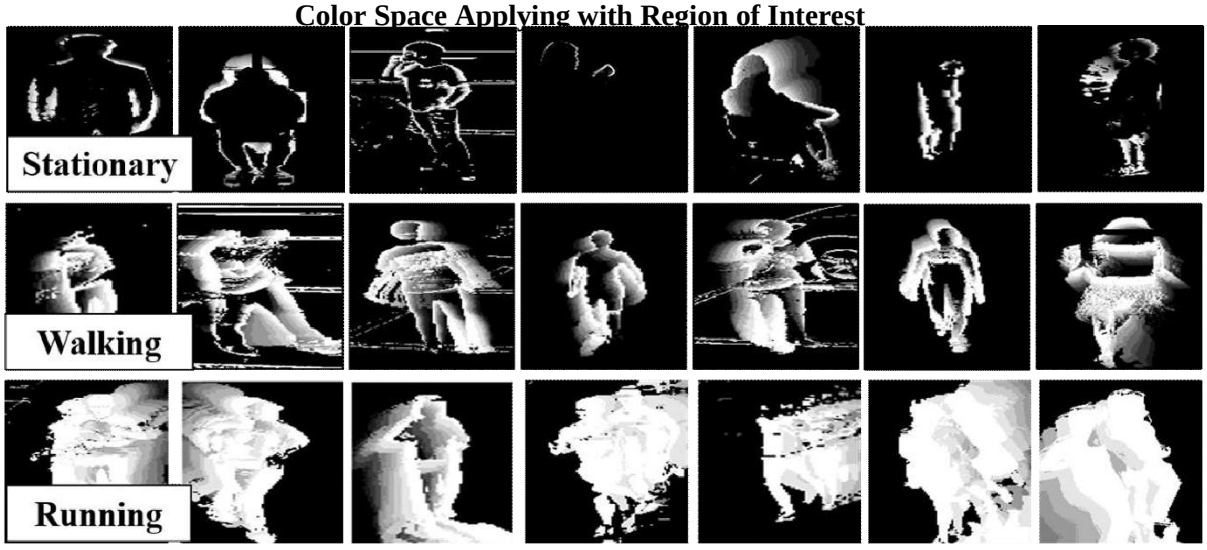

Figure 8: Examples of MHI for different sub-actions. MHIs are utilized for the locomotion level of the sub activity descriptor, which includes stationary, walking, and running. MHI catches the movement cue history of the actor, where the most recent pixel movement spaces are brighter. The human eye can easily distinguish stationary, walking, and running from MHIs

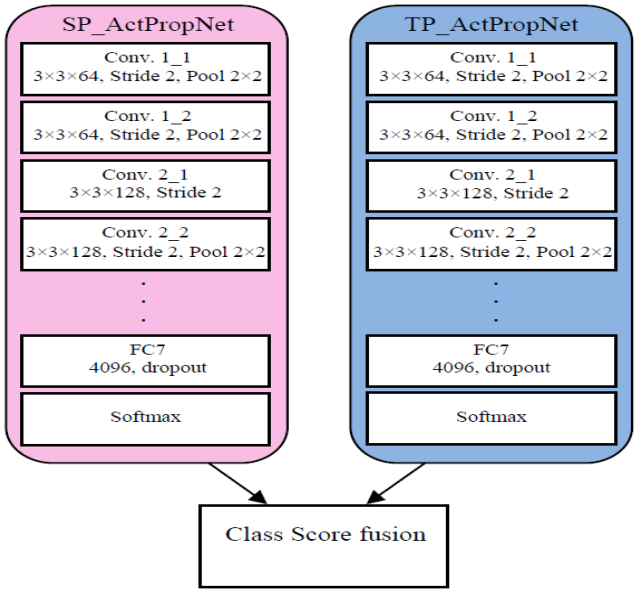

Figure 9. The training network architecture for human action recognition.

E. Implementation Details

There are two networks created in the training phase: The SP_ActPropNet and the TP_ActPropNet which are shown in the Figure 9. The appearance clues of the scene are taken as input by the first one and it is trained using region proposed spatial images. The second one uses the region proposal optical flow images produced using the proposed ARP scheme. Caffe framework is used to train both networks with back propagation. The ImageNet model is selected as the initial configuration for training both the spatial and temporal system. The rate of learning is fixed to 0.001 at the beginning and it changes 3 times throughout the training process. In addition a weight decline of 0.0005 and the momentum of 0.9 are also used. The networks are initialized well in order to train the spatial network for 15K iterations as it needs more iteration. The temporal network is also trained using a similar architecture, in which the initial rate of learning is fixed as 0.005 and it is altered five times and the iteration is extended up to 40K times. The rate of dropout for completely connected layers varies from the spatial network. The rate of dropout 0.7 and 0.9 are used for layers FC6 and FC7 respectively for the training.

3.5.1 Testing the Network

During testing, the strategy proposed in [55] is used to confirm a fair comparison. Test data is built by isolating 25 video clips of spatial images as well as optical flow fields and the functioning of SP_ActPropNet and TP_ActPropNet is tested respectively by comparing their performance.

3.5.2 Fusion of CNN Features

Discriminative action classifiers were used to create predictions for every region on spatio-temporal characteristics. The last fully associated layer of the CNNs (FC7) are used to obtain the characteristics of CNN. The characteristics of CNN from SP_ActPropNet and TP_ActPropNet are integrated. This contains 2×4096 dimensional descriptors and a linear support vector machine is trained using these descriptors to act as the last classifier.

The process of combining

spatial and motion indications and feeding into the SVM classifier is depicted in figure 3.

Experiments and Results

Table 1: Per-class breakdown and mAP on the KTH dataset. The improved scores of every label are illustrated in bold lettering.

Method	Boxing	Hand clapping	Hand waving	Jogging	Running	Walking	mAP (%)
BDI-CNN	86.11	80.56	72.22	72.22	80.56	75.00	77.78
MHI-CNN	88.89	91.67	83.33	86.11	88.89	86.11	87.50
WAI-CNN	94.44	94.44	91.67	86/11	91/67	88.89	91.20
Region Proposed method	**100**	**97.22**	**100**	**91.67**	**94.44**	**94.44**	**96.30**

Shape cues (BDI) lead to captured silhouette features from the spatial domain and can effectively identify the posture of an actor. They can provide over 95% mAP at the posture level of the sub-action descriptor. Motion history cues (MHI), even simple and fast temporal features, are of crucial importance for recognizing sub-actions of the locomotion level: stationary, walking, and running. However, deciding the memory capacity of MHI is a highly action-dependent issue. As determined from a large set of experiments, correctly recognizing one action takes approximately 2s with the ICVL dataset. The combination of shape and historical movement cues (WAI), when the weighted average was used with them, can provide further improvement in performance for the gesture level of the sub-activity descriptor. Shape and historical movement cues are complementary for gesture-level sub-action recognition. From an ablation study, it was noted that the historical movement cue provide more information when compared to shape cue if the two were used individually. However, after combining them, the shape cue contributed much more than the motion cue for gesture- level sub-action recognition. And also, we have implemented the two-stream network suggested in [46] using UCF101 dataset. We have attained nearly matching performance (82.1% for spatial CNN with region proposal algorithm and 83% for temporal CNN when L=1)

Table2. Comparison of performance with the up-to-date deep networks on UCF 101 dataset.

Training Setting	Accuracy of spatial network on UCF 101	Accuracy of temporal Network on UCF 101	Final accuracy after fusion on UCF 101
ActPropNet	82.1%	87.65(L=1)	92.11(L=1)
Two-stream CNN[1]	75.7%	78.3%(L=1) 83%(L=10)	N/A(L=1)
Single Frame[2]	76%	81.2%	86.04%
LRCN-fc6[2]	77.12%	79.95%	84.95%
Two – stream + LSTM[3]	75.1%	N/A	91%
Deep net [4]	68.4%		

Table3. Comparison of performance of the up-to-date work with our scheme, before and after the manual removal of background using UCF101 dataset

Model of training	Accuracy using UCF 101 dataset	Accuracy on a subset of 200 physically eliminated background images
Implementation of spatial CNN[1] with Region proposal algorithm	82.1%	69%
SP-ActPropNet with Region proposal algorithm	78.2%	71.98%

Then, another experimentation was accompanied to measure the efficiency by using the trained spatial network in the control case of manual background removal.

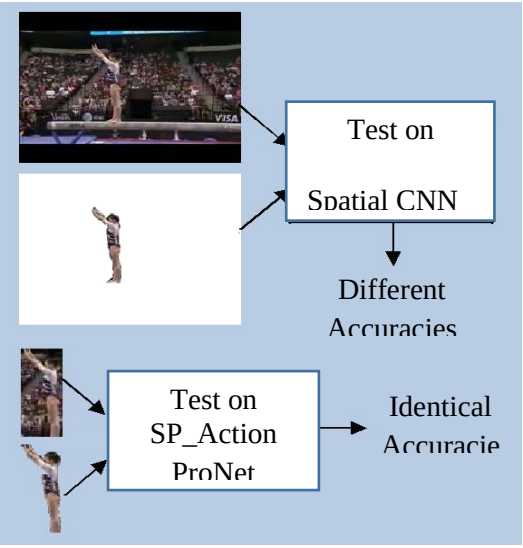

Figure 10. Depicts the replacement of specific sample background images from UCF101 dataset with a colourless background.

This experiment is conduct using generated foreground clues in a small subclass of sample dataset i.e., 200 images.

An experiment was conducted on trained spatial network and the performance was approximately fell down by 8%.Since the adequate robotics datasets are not available deployable, beneficial action recognition system could be developed. The network can be on huge publicly available datasets rather than training it from the scratch in order to enhance the architecture to produce the preferred resultant task. Hence, we require a consistent pre-trained model. Test on Spatial CNN Different Accuracies Figure 10. Manual background removal scenario.

Action regions were selected from the training and test data (refer section 3.1) then the training procedure on these regions' proposed images was performed rather than taking full images for both spatial and temporal networks. Even though the precision of spatial system was somewhat reduced by 2.6% when compared with [55], the accuracy of temporal network was increased greatly up to 6.8%. But, it is proved that the lower accuracy for spatial stream utilizes the background cues for the classification purposes. Once combining the learned spatial and temporal CNN characteristics through SVM, we attained the up-to-date matching accuracy of (92.11%). Table 2 illustrates the results and compares with the five up-to-date methods. Further, the same controlled experiment was carried out for manually removing the background in order to test SP_ActPropNet on the suggested image space that was mined from the recently produced sample set. We have perceived an interesting fact that the precision of SP_ActPropNet continued to be a constant, which proves that our method's consistency irrespective of the background and context. Table 3 shows the outcome of the experiment conducted for background removal with full input images for trained spatial network of [55] and region proposed input images for SP_ActPropNet.

IV. CONCLUSION

The present study proposes a consistent method for action detection and recognition of real world surveillance video datasets. At first, real time action detection and recognition were done using BDI, MHI and WAI. Extensive experiment validated that a sub activity descriptor brings entire information of human activities (82.11%) and altogether disposes of all misclassifications using more activities on some independent stages. Second, CNNs were used on the basis of appearance and movement clues. In a series of investigations conducted, common CNN models were used based on learned features from the foreground and background clues. The study, by using various experiments provides a solution on appropriate system irrespective of the related background information. Also, an "action region proposal" scheme is used, that dynamically shift the attention to the regions where the possibility is more for the actions to happen. Throughout these experiments, the temporal network beaten the best in class methods utilizing an optical flow domain and the performance of spatio-temporally merged activity recognition method synchronised or overtook the state-of-the-art, with an accuracy of 92.11%.

REFERENCES

1. E. Ricci et al., "Uncovering Interactions and Interactors: Joint Estimation of Head, Body Orientation and Formations From Surveillance Videos," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4660–4668.
2. B. Wang et al., "Abnormal crowd behavior detection using size-adapted spatio- temporal features," Int. J.Control. Autom. Syst., vol. 9, no. 5, pp. 905–912, Oct.2011.
3. M. Yang et al., "Detecting human actions in surveillance videos," in TREC video retrieval evaluation workshop,2009.
4. S. Kim et al., "Intelligent visual surveillance — A survey," Int. J. Control. Autom. Syst., vol. 8, no. 5, pp. 926–939, Oct.2010.
5. C.Schüldt et al., "Recognizing human actions: Alocal SVM approach, "in Proceedings - International Conference on Pattern Recognition, 2004, vol. 3, pp. 32–36.
6. L. Gorelick et al., "Actions as space-time shapes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 12, pp. 2247–2253,2007.
7. S. Oh et al., "A large-scale benchmark dataset for event recognition in surveillance video, "in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 3153–3160.
8. F. Smeaton, et al., "Evaluation campaigns and TRECV id," in Proceedings of the 8th ACM international workshop on Multimedia information retrieval-MIR'06,2006,p.321.
9. Zhou et al., "Object Detectors Emerge in Deep Scenes" CNNs, arXiv:1412.6856 [cs.CV].
10. M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real-time range sensor," 2010 IEEE Compute. Soc. Conf. Compute. Vis. Pattern Recognition. - Work., pp. 1–8, 2010.
11. C. Plagemann, et al., "Real-time identification and localization of Body parts from depth images in 2010 IEEE International Conference on Robotics and Automation, 2010,pp. 3108–3113.
12. R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 781–796, 2000.
13. B. Zhang et al., "Real-time Action Recognition with Enhanced Motion Vector CNNs," Apr. 2016.
14. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, 2005, vol.I, pp. 886–893.
15. C. Yu and T. Joachims, "Learning structural SVMs with latent variables," in … International Conference on Machine Learning, 2009, pp. 1–8.
16. Dollar et al., "Behavior recognition via sparse spatio -temporal features". In VS-PETS, 2005. Laptev. On space-time interest points. IJCV, 64(2/3):107-123, 2005.
17. Wang et al.," Latent hierarchical model of temporal structure for complex activity classification". TIP, 23(2), 2014.
18. Gkioxari et al., "Contextual action recognition with r* cnn." arXiv:1505.01197, 2015.
19. J. Aggarwal and M. Ryoo, "Human activity analysis: A review," ACM Compute. Surv., vol. 43, no.3, p. 16:1- 6:43, 2011.
20. R. Poppe, "A survey on vision-based human action recognition," Image Vis. Compute., vol. 28, no.6, pp. 976–990, 2010
21. B. Sapp, D. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011, pp.1281–1288
22. Cherian et al., "Mixing body-part sequences for human pose estimation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, pp. 2361–2368.
23. J. Shotton et al., "Real-time human pose recognition in parts from single depth images," Commun. ACM vol. 56, no. 1, pp. 116–124, 2013.
24. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 3, pp. 257–267, 2001.
25. J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," Proc. IEEE Compute. Soc. Conf. Compute. Vis. Pattern Recognition., vol. 23, no. 402, pp. 928–934, 1997.
26. C. Bin Jin et al., "Real-time human action recognition using CNN over temporal images for static Video Surveillance cameras," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 9315, pp. 330–339.
27. S. Ali et al., "Chaotic invariants for human action recognition," in Proceedings of the IEEE International Conference on Computer Vision, 2007.

28. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008.
29. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.
30. Lowe, 2004] DG Lowe. Distinctive image features from scale-invariant key points. In IJCV, 2004.
31. N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In Proceedings of the European Conference on Computer vision (ECCV), 2006.
32. Krizhevsky et al., " Image Net classification with deep convolution neural networks". In Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), pages 1106–1114, 2012.
33. Baccouche et al., "Sequential deep learning for human action recognition. "HBU Springer, pages 29–39, 2011.
34. Ji et al., "3D Convolutional Neural Networks for Human Action Recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, January 2013.
35. Wang et al., Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
36. Du et al., "Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015
37. Karpathy et al., " Large-scale video classification with convolutional neural networks. In Proceedings Of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014
38. Taylor et al., "Convolutional learning of spatio-temporal features. In Proceedings of the European Conference on Computer vision, 2010.
39. Ji et al.,"3D Convolutional Neural Networks for Human Action Recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, January 2013.
40. Du et al., "Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
41. Karpathy et al., " Large-scale video classification with convolutional neural networks". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014
42. Simonyan and Zisserman, " Two-stream convolutional networks for action recognition in videos". In proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), pages 1-8, 2014.
43. C. Yu and T. Joachims, "Learning structural SVMs with latent variables," in International Conference on Machine Learning, 2009, pp. 1–8.
44. Yu and Yuan, " Fast action proposals for human action detection and search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
45. Lawrence et al., "Edge boxes: Locating object proposals from edges." In Proceedings of the European Conference on Computer vision (ECCV), 2014.
46. Rezazadegan et al.," Evaluation of object detection Proposal under condition variations. In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015.
47. Brox et al., " High accuracy optical flow estimation based on a theory for warping. In Proceedings of The European Conference on Computer vision (ECCV), pages 25–36, 2004.
48. Gkioxari and Malik, "Finding action tubes". In Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition (CVPR), 2015.
49. S. Li, "Human re-identification using soft biometrics in video surveillance," 2015.Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. 2004, vol. 2, no. 2, pp. 28–31.
50. B. Babenko and S. Belongie, "Visual tracking with online Multiple Instance Learning," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 983–990.
51. Soomro et al., "UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, abs/1212.0402, 2012
52. Simonyan and Zisserman, "Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014
53. Wang et al., " Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
54. Gkioxari and Malik, " Finding action tubes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
55. .Jia et al., "Convolutional architecture for fast feature embedding". In ACM MM, 2014.

AUTHORS PROFILE

N. Kumaran Completed his B.E (Computer Science and Engineering) from CIT, Coimbatore, Tamil Nadu, India and M. Tech (Information Technology) from Sathyabama University, Chennai, in the year 1998 and 2007 respectively. Currently pursing Ph.D (Video Analysis) in NIT, Tiruchirappalli.-15,India. His areas of interest include video processing and Computer Networks. He has got around 18 years of teaching experience in various Institutions

Dr. U. Srinivasulu Reddy received his Ph.D. from National Institute of Technology, Tiruchirappalli.. M.Phil. & MCA from Bharathidasan University, Tiruchirappalli.. He is currently working as Assistant Professor in the Department of Computer Applications, National Institute of Technology, and Tiruchirappalli.. He is a life time member of the Computer Society of India (CSI). His research interests include Big Data Analytics, Machine Learning and Bioinformatics

Dr. S. Saravana Kumar is a professor in Shanmuganathan College of Engineering, Pudukkottai, India. He received Ph.D., from National Institute of Technology, Tiruchirappalli. He received M.E Computer Science and Engineering from Anna University, Chennai. His research interest includes image and video processing and object detection.