Journal of Multimedia Information System VOL. 4, NO. 4, December 2017(pp. 225-232): ISSN 2383-7632(Online)
http://dx.doi.org/10.9717/JMIS.2017.4.4.225

Fight Detection in Hockey Videos using Deep Network

Subham Mukherjee[1], Rajkumar Saini[2,*], Pradeep Kumar[2], Partha Pratim Roy[2], Debi Prosad Dogra[3], and Byung-Gyu Kim[4]

Abstract

Understanding actions in videos is an important task. It helps in finding the anomalies present in videos such as fights. Detection of fights becomes more crucial when it comes to sports. This paper focuses on finding fight scenes in Hockey sport videos using blur & radon transform and convolutional neural networks (CNNs). First, the local motion within the video frames has been extracted using blur information. Next, fast fourier and radon transform have been applied on the local motion. The video frames with fight scene have been identified using transfer learning with the help of pre-trained deep learning model VGG-Net. Finally, a comparison of the methodology has been performed using feed forward neural networks. Accuracies of 56.00% and 75.00% have been achieved using feed forward neural network and VGG16-Net, respectively.

Key Words: Hockey Videos, FFT, Radon Transform, Feed Forward Neural Network, Convolutional Neural Network, VGG Net.

I. INTRODUCTION

Recognition of Human Actions is a challenging task and has received significant amount of attention amongst researchers in this field [8]. Unlike analyzing such actions from still images, action detection from the temporal frames in videos provides an additional clue for recognition as a number of actions can be analyzed using motion information. Going by a human way of perceiving a fight video, we classify a fight video on the basis of certain observations such as changes in movement direction, velocity, acceleration in the consecutive video frames. In this paper, we make use of similar features for action recognition to detect fight scene in Hockey videos.

Video based action recognition has been explored with more attention by various researchers in computer vision community. A number of research work has been done in last few years. Guo et al. [4] have proposed the use of nearest neighbor and sparse linear approximation classifiers to classify covariance matrices based features for test videos. E.Bermejo et al. [1] proposed the well-known Bag-of-Words approach, along with the use of two best descriptors currently available, MoSIFT and STIP for action recognition, specifically for fight detection. Cheng et al. [2] used a hierarchical approach based on Gaussian mixture models and Hidden Markov models (HMM) to recognize gunshots, explosions and car-breaking in audios. Video and GPS based trajectory classification [10, 11] and clustering [9, 12] has also been studied that could be used for anomaly detection. Giannakopoulos et al. [3] has presented a method for violence detection in movies based on audio-visual information that uses an audio-statistics and motion-orientation and magnitude features in video together in a k-Nearest Neighbor classier to decide whether a given sequence is violent or not. Andrej et al. [5] extended the connectivity of a CNN from simply the time domain to the spatio-temporal domain in his paper. This is a remarkable improvement in the domain of video analysis extending the use of CNNs from image recognition to analysis of video sequences. Simonyan et al. [13] introduced a Two-Stream Convolutional Architecture which uses spatial data and temporal data such as flow fields from sequential frames to recognize human actions.

Despite a significant works, video based action recognition is still a challenging task because of the complex nature of scene such as illumination changes,

Manuscript received September 20, 2017; Revised October 22, 2017; Accepted December 1, 2017. (ID No. JMIS-2017-0034)
Corresponding Author (*): IIT Roorkee, India, E-mail. rajkumarsaini.rs@gmail.com
[1]Institute of Engineering and Management, Kolkata, India , subhammukherjee61196@gmail.com
[2]IIT Roorkee, India, rajkumarsaini.rs@gmail.com, pradeep.iitr7@gmail.com, proy.fcs@iitr.ac.in
[3]IIT Bhubaneswar, dpdogra@iitbbs.ac.in
[4]Sookmyung Womens University, Seoul, Republic of Korea, bg.kim@ieee.org

clutter, occlusions etc. The movement or distortions during video acquisition and The variability in human action makes it more difficult to correctly recognize human actions.

In this paper, we have done a comparative study on the application of normal feedforward, fully-connected neurons to detect human actions from acceleration and deceleration patterns in a video and using Artificial CNNs in action detection from dense optical-flow fields obtained from the temporal frames in a video. Till now most of the work done in this field are mainly audio-visual approaches where both motion and motion orientation along with statistics of audio are analyzed to estimate the output. However, analyzing human action recognition on the basis motion analysis in videos has not been done using CNN and other deep learning approaches. The first part of this paper is mainly focused on extracting the displacement patterns from consecutive frames in a given amount of time, throughout the video. Acceleration and deceleration patterns estimated on the basis of Motion Blur in the temporal frames of the videos are fed into a fully connected neural network to learn the pattern for human action classification. Next, we extend CNN to estimating human action in video data.

The neurons in Convolutional Neural Networks(CNNs) respond to features in a restricted region of space known as the receptive field. The receptive fields of different neurons partially overlap such that they tile the visual field. The response of an individual neuron to a feature within its receptive field can be approximated by a convolution operation. The Convolutional Layers in ConvNets inherit the ability to adapt themselves according to the input provided to them to produce the desired output and hence have the ability to extract the best features out of the input image. In our implementation we have used CNN because of the ability of ConvNets to efficiently fit non-linear data, the input to the convolutional layers is a stack of optical flow fields in a given amount of time.

Multilayer perceptrons of Feed-forward Neural Networks with hard-limiting (signum) activation functions can adapt to complex decision boundaries. Unlike ConvNets, neurons in these layers are fully-connected and do not have a specific receptive field. In this paper we have tested the abilities of both the networks to classify human actions, in the same dataset. Self-regulating visual vigilance systems are highly dependent upon motion patterns of moving objects to find out distinct types of activities occurring within the range of sensor. Motion patterns of the objects can be quite explanatory and often used for various assignments, such as scene semantic analysis [14], highway traffic management [6], atypical activity detection [7], etc.

II. FIGHT DETECTION USING FEED-FORWARD NN

A video is a stack of frames. A violent action in a video can be specified by a large acceleration or deceleration in movements. A large acceleration or deceleration can be analyzed by the amount of motion blur in the temporal frames of the video and by the pattern in its variation.

Motion blur is the apparent streaking of rapidly moving objects in a still image or a sequence of images such as in a video. It results when the image being recorded changes during the recording of a single exposure, either due to rapid movement or long exposure. Very fast movements usually have larger amount of motion blur in comparison to any normal movement in the video frames. Motion blur usually shifts the frequency spectrum of an image towards the lower frequency spectrum. Added to that feature motion blur also suppresses the gradient and sharpness of an image. These two features form a good estimator of acceleration or deceleration in a video. A good candidate to measure the gradient and sharpness of an image is the variance of Laplacian in the image. First we subtract two consecutive frames from each other to get a motion map, that is an image which is the summation of the Global and Local motion from two consecutive frames. Next we suppress the Global Motion in the image which is a result of camera movement and not action in the video and is hence not a good estimator of rapid action in the video. Whereas Local Motion is the result of local movements (dx) occurred in the video. Hence, this provides us with a measure of the amount of displacement in two consecutive frames. The frames per-second provides us with (dt), that is the amount of time required for displacement (dx). Hence this way we get the motion velocity in two consecutive frames (dx/dy). If this is continued for all the frames in the video we get a measure of the continuous acceleration (d(dx/dt)/dt) and deceleration in the video. Here, in our work we have used Variance of Laplacian as a measure of the displacement (dx) in each motion map (I).

After enhancing the local motion in the differential image (I) we compute the power spectrum from two consecutive frames. It can be proved that when there is sudden motion in between two consecutive frames, the power spectrum image of the second frame will depict an ellipse. The ellipse orientation is perpendicular to the direction of motion, the frequencies outside the ellipse being attenuated. In our implementation we make use of the eccentricity of the ellipse which is dependent on the acceleration.

2.1. 2D FFT(Fast Fourier Transform)

The Fourier Transform is used to decompose an image into its sine and cosine components. The output of the transformation represents the image in the Fourier or

frequency domain, while the input image is the spatial domain equivalent. In the Fourier domain image, each point represents a particular frequency contained in the spatial domain image. In our implementation we compute the power spectrum of the frames using the 2D Fast Fourier Transform.

Algorithm 1 Local Motion Enhancement Procedure

For every two consecutive frames *I1,I2* **do**
  A=I1-I2
  For each pixel *p in A* **do**
    Take 3x3 matrix around *p*
    Compute the variance Mp of the matrix.
    M(p)=Mp \\ M being the variance mask. It provides the variance of each pixel w.r.t. its surrounding environment. The variance is likely to be higher for regions containing fight scene, i:e: local motion.
  EndFor
  \\ M has larger intensity in areas of local motion while very less intensity for areas with global motion.
  IM=M⊗A \\It enhances local motion and suppresses global motion.
EndFor

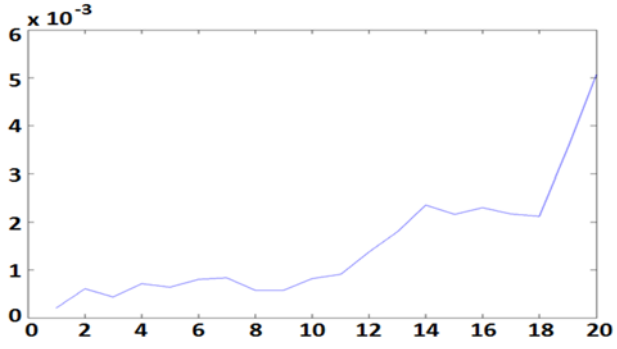

Fig. 1. Example of variance pattern of non-fight video.

2.2. Radon Transform

Radon transform computes projections of an image matrix along specified directions. In our implementation we have used projections from 0 to 180 degrees. It used to reconstruct the frequency data from 2-D FFT into a two dimensional form on which further calculations have been done. As stated above motion blur shifts the frequency spectrum of an image towards the lower frequencies. Hence applying a low pass-filter is equivalent to the motion blur in the image. When we subtract two consecutive frames I(i) and I(i-1) and suppress the global motion we extract the low frequency components from the image.

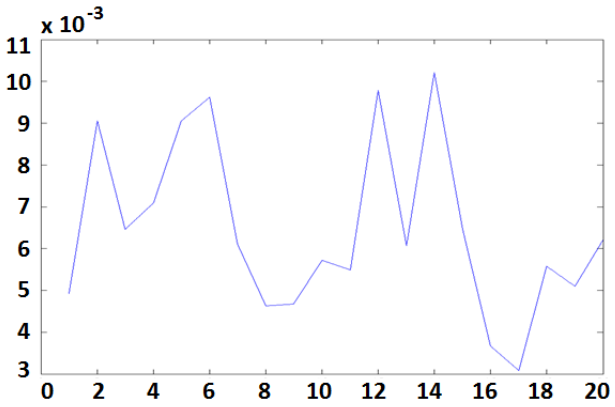

Fig. 2. Example of variance pattern of fight video.

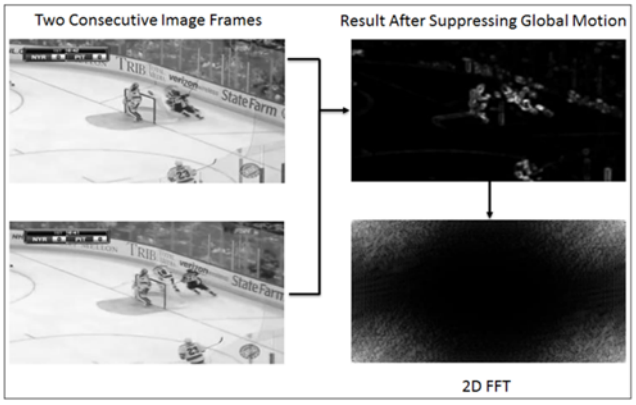

Fig. 3. Example of motion map and FFT for non-fight video.

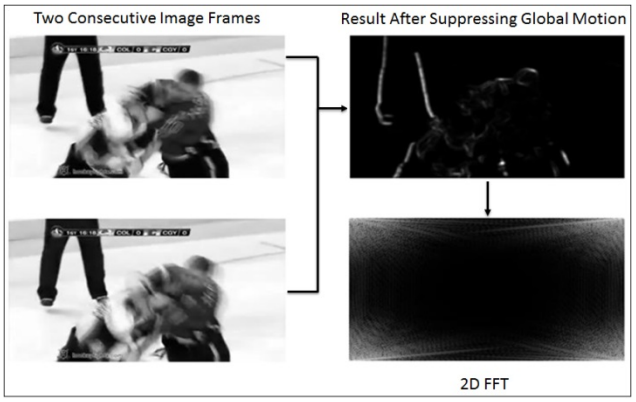

Fig. 4. Example of motion map and FFT for fight video.

Next, a 2-D FFT is used view the power spectrum of these low frequency components. Then Radon Transform is used to reconstruct the low frequency components from the power spectrum acquired from FFT and its vertical maximum projection vector (vp) is obtained and normalized to maximum value 1. The kurtosis (K) and the mean (M) of this vector is therefore taken as an estimation of the acceleration. An example of the Radon Transform of two consecutive frames in a fight video is shown below: Since kurtosis alone cannot be used as a measure, since it is

obtained from a normalized vector (i.e. it is dimensionless), the average power per pixel P of image C is also computed and taken as an additional feature. Without it, any two frames could lead to high kurtosis even without significant motion.

III. FIGHT DETECTION USING CNN

CNN in machine learning is a type of feed-forward architecture in which the connectivity pattern amongst the neurons is similar to the human visual cortex. CNNs make use of the fact that the input data consists of images and hence constrain their architecture in a more advantageous pattern. The neurons of these neural networks are arranged over 3 dimensions: width, height and depth. These networks usually consist of three layers (i) Convolutional Layer: This layer consists of multiple learnable filters that adopt their weights accordingly to extract a learnable feature from the input data. In our implementation we have used a 3-D convolutional Layer to process a stack of frames along spatial dimensions. The parameters used here are the 2-D dense optical flow fields in the temporal frames of the video and the time elapsed which is available from the number of frames per second in the video. (ii) Pooling Layer: This layer performs down-sampling operations along spatial dimensions to reduce the complexity of the data. The combination of the Convolutional Layers and the Pooling Layers give rise to Locally-Connected Layers. (iii) Fully Connected Layer: This layer is similar to the feed-forward neural network architecture. All the neurons here are connected each other. The output of the Convolutional Layer and the Pooling Layer is being learned by the Fully Connected Layer to give the desired output.

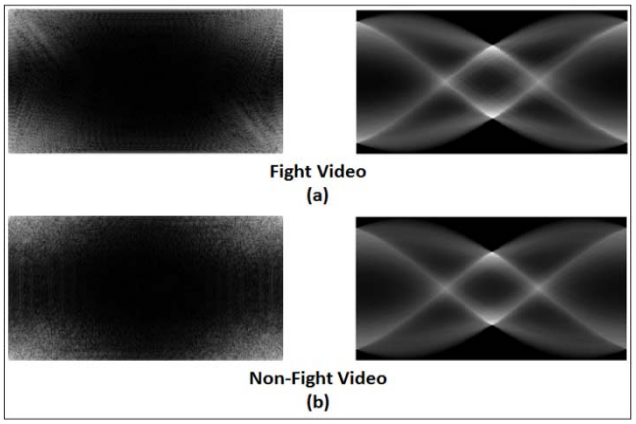

Fig. 5. Example of FFT and radon transformation for fight and non-fight video.

3.1. Dense Optical Flow

Optical flow is a pattern in which visible objects are apparently moving. It is a measure of the relative motion of objects or edge or surfaces between the observer and the scene. In our implementation we use FarneBacks algorithm to compute dense flow fields of two temporal frames. The flow fields consist of a (dx) component and a (dy) component (i.e. the magnitude component and the phase component of displacement). These components can give a clear idea about the acceleration and the deceleration in the consecutive frames throughout the video. Unlike normal convnet models, the input to our model is formed by stacking optical flow displacement fields between several consecutive frames. Such an input clearly describes the motion between video frames, which makes the recognition easier.

Algorithm 2 Feature Extraction Procedure

1 **For** every two consecutive video frames *I1,I2* **do**

2 $D(x,y,t)=I1(x,y,t)-I2(x,y,t-1)$

3 $A(x,y,t)=D(x,y,t)$ \\ with enhanced Local Motion as mentioned in Algorithm 1.

4 $F(v,w,t)=FT(A(x,y,t))$ \\ *FT* denotes 2D fourier transform.

5 Apply Radon transform, $R(d,\theta)=Radon(f(v,w,t))$

6 Normalize, $P(\theta)=max(R(d,\theta))$

7 $K=Kurtosis(R(d,\theta))$

8 $M=Mean(R(d,\theta))$

9 **EndFor**

10 **Return** *Histogram(P, nbins), Histogram(K,nbins), Histogram(M,nbins).*

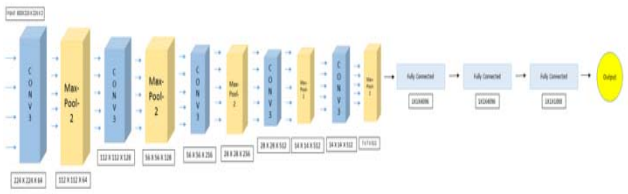

Fig. 6. VGG-16 Network.

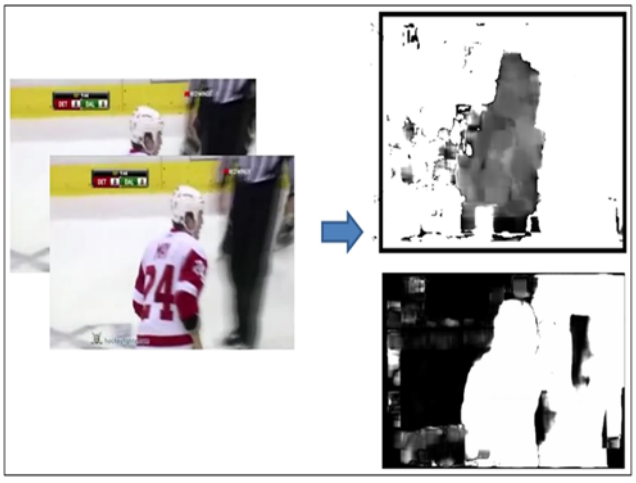

Fig. 7. Example of flow field extracted from fight video.

IV. EXPERIMENTAL RESULTS

In this section, we present the outcomes obtained by applying our method on National Hockey League (NHL) dataset.

4.1. Datasets and Ground Truths

The dataset consists of 1000 video clips of NHL games. Each video clip consists of 50 frames. The resolution of each frame is 720 × 576. The video frames are labeled into two classes i.e. fight and non-fight. Each video clip is further divided into two consecutive parts consisting of 25 frames each to increase the number of training samples. 70% of the total data is used to train the networks, 10% for validation and the left 20% is used for testing the trained models. The training and validation clips are randomly picked from the dataset. The data is randomly shuffled after each epoch for robust training and testing.

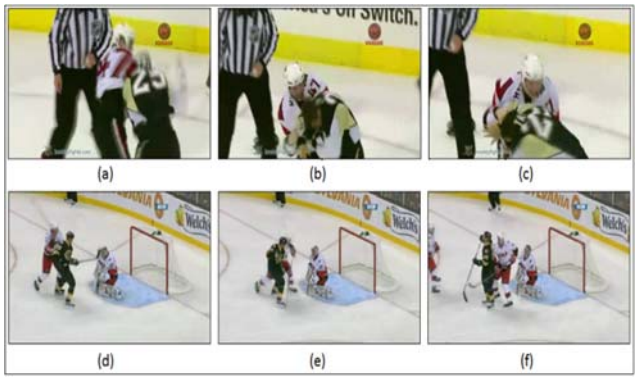

Fig. 8: Example of video frames from dataset. (a), (b), (c) are 3 consecutive frame of fight video.

4.2 Fight Detection Results

The results of fight detection in Hockey videos have been discussed in this section.

4.2.1. Without CNN

Here we discuss the classification results using a Feed-Forward Neural Network. The network is trained with the features extracted from each video as mentioned in algorithm 2. All the data are first normalized and zero centered before training and the network weights are then initialized using Xavier initializer. Next the network is trained using RMSprop with a learning rate of 1 exp -2 and a decay rate of 1 exp -6. The batch size is fixed and equal to 16. A dropout of 20% is introduced after every dense layer while training to prevent the model from overfitting. Figure 9 shows the accuracy using feed-forward NN by varying number of epochs. It can be observed from figure that accuracy does not improve sufficiently after 200 epochs and the accuracy of 56.00% has been recorded.

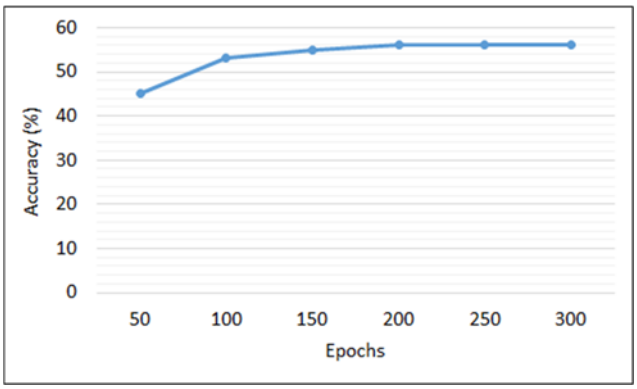

Fig. 9. Accuracy using feed forward NN by varying number of epochs.

4.2.2. With CNN

In this part of the discussion, we observe the classification results using a deep CNN. The CNN architecture used for classification is the VGG16 framework. The network is initialized and trained in a similar manner as discussed in Section 4.2. Figure 10 shows the accuracy using VGG16-Net by varying number of epochs. After 550 epochs the performance does not change and the accuracy of 75.00% has been recorded using VGG16-Net. Figure 11 shows the comparison of simple feed-forward NN and VGG16-Net. The VGG16-Net performs better as compared to simple feed-forward NN.

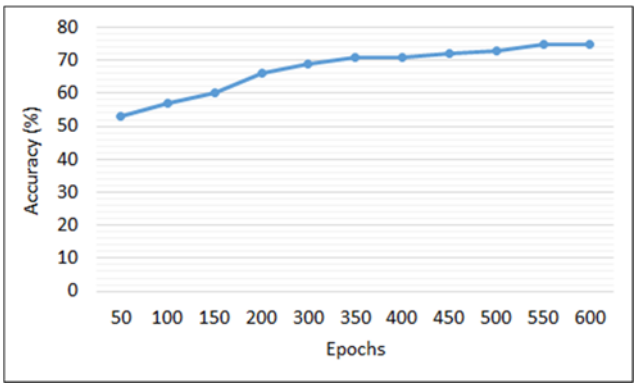

Fig. 10. Accuracy using VGG Net by varying number of epochs.

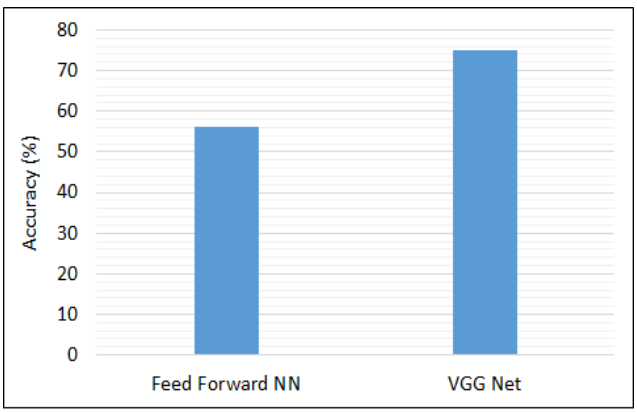

Fig. 11. Performance comparison of Feed Forward NN and VGG Net.

V. CONCLUSION

Violence detection in videos can be useful in order to prevent children from such content that can put ill effect on their minds. Effective systems can be built and embedded into video players/browsers assuming the viewers are children so that unsuitable content is restricted from being viewed. In this paper, detection of fights in Hockey videos has been done using blur and radon transform with the help of feed-forward NN. The performance has also been fine-tuned using pre-trained VGG16 deep CNN. The performance of simple feed-forward NN is recorded as high as 56.00%, whereas after fine-tuning using VGG16-Net the performance increases to 75.00%. In future, more complex videos could be analyzed to design more robust system.

REFERENCES

[1] E. B. Nievas, O. D. Suarez, G. B. García and R. Sukthankar, "Violence detection in video using computer vision techniques," *In International conference on Computer analysis of images and patterns*, pp. 332-339, Heidelberg, 2011.

[2] W. H. Cheng, W. T. Chu and J. L. Wu, "Semantic context detection based on hierarchical audio models," *In Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 109-115, 2003.

[3] T. Giannakopoulos, A. Makris, D. I. Kosmopoulos, S. J. Perantonis and S. Theodoridis, "Audio-Visual Fusion for Detecting Violent Scenes in Videos," *In SETN*, pp. 91-100, 2010.

[4] K. Guo, P. Ishwar and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Transactions on Image Processing*. Vol. 22, No. 6, pp. 2479-2494, 2013.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725-1732, 2014

[6] J. Melo, A. Naftel, A. Bernardino and J. Santos-Victor, "Detection and classification of highway lanes using vehicle motion trajectories," *IEEE Transactions on intelligent transportation systems*, Vol. 7, No. 2, pp.188-200, 2006.

[7] C. Piciarelli and G. L. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1835-1842, 2006.

[8] R. Poppe, "A survey on vision-based human action recognition". *Image and vision computing*, Vol. 28, No. 6, pp. 976-990, 2010.

[9] R. Saini, A. Ahmed, D. P. Dogra and P. P. Roy, "Classification of object trajectories represented by high-level features using unsupervised learning," *In Proceedings of International Conference on Computer Vision and Image Processing*, pp. 273-284, Singapore, 2017.

[10] R. Saini, A. Ahmed, D. P. Dogra and P. P. Roy, "Surveillance scene segmentation based on trajectory classification using supervised learning," *In Proceedings of International Conference on Computer Vision and Image Processing*, pp. 261-271, Singapore, 2017.

[11] R. Saini, P. Kumar, S. Dutta, P. P. Roy and U. Pal, "Local behavior analysis for trajectory classification using graph embedding," *In 4th Asian Conference on Pattern Recognition,* 2017. (accepted)

[12] R. Saini, P. Kumar, P. P. Roy and D. P. Dogra, "An efficient approach for trajectory classification using FCM and SVM". *In IEEE Region 10 Symposium (TENSYMP),* pp. 1-4, 2017.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *In Advances in neural information processing systems*, pp. 568-576, 2014.

[14] X. Wang, K. Tieu and E. Grimson, "Learning semantic scene models by trajectory analysis," *Computer Vision–ECCV*, pp.110-23, 2006.

Journal of Multimedia Information System VOL. 4, NO. 4, December 2017(pp. 225-232): ISSN 2383-7632(Online)
http://dx.doi.org/10.9717/JMIS.2017.4.4.225

Authors

Subham Mukherjee is pursuing B.Tech. in Department of Electronics and communication Engineering at Institute of Engineering and Management,Kolkata, India. His research interest includes Machine Learning, Pattern Recognition and Computer Vision.

Pradeep Kumar is pursuing Ph D in Department of Computer Science and Engineering at IIT Roorkee, India. His research interest includes Human Computer Interaction (HCI) and Brain Computer Interface (BCI).

Rajkumar Saini is pursuing his PhD in Department of Computer Science at IIT Roorkee, India. His research interest includes Pattern Recognition, Machine Learning.

Dr. Partha Pratim Roy received his Ph.D. degree in computer science in 2010 from Universitat Autònoma de Barcelona, (Spain). He worked as postdoctoral research fellow in the Computer Science Laboratory (LI, RFAI group), France and in Synchromedia Lab, Canada. Presently, Dr. Roy is working as Assistant Professor at Indian Institute of Technology (IIT), Roorkee. His main research area is Pattern Recognition.

Dr. Debi Prosad Dogra (M'06) obtained Ph.D. degree from IIT Kharagpur, India in 2012, M.Tech degree from IIT Kanpur, India in 2003, and B.Tech from Haldia Institute of Technology, India in 2001. Presently, he is an Assistant Professor in the School of Electrical Sciences, IIT Bhubaneswar, India. He was with Samsung Research India (2011-2013). He worked with ETRI, South Korea (2006-2007). He worked as a faculty at HIT (2003-2006). He has published more than 50 research papers in international journals and conferences. His research interests include visual surveillance, augmented reality and human computer interface. He has obtained three US patents.

Byung-Gyu Kim has received his BS degree from Pusan National University, Korea, in 1996 and an MS degree from Korea Advanced Institute of Science and Technology (KAIST) in 1998. In 2004, he received a PhD degree in the Department of Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST). In March 2004, he joined in the real-time multimedia research team at the Electronics and Telecommunications Research Institute (ETRI), Korea where he was a senior researcher. In ETRI, he developed so many real-time video signal processing algorithms and patents and received the Best Paper Award in 2007.