

## Article

# IDENTIFYING HUMAN BEHAVIOURS USING DEEP TRAJECTORY DESCRIPTORS

Tauseef Ali<sup>1</sup>, Eissa Jaber Alreshidi<sup>2</sup>

<sup>1</sup> Affiliation 1; [tauseeftwente@yahoo.com](mailto:tauseeftwente@yahoo.com), University of Twente, The Netherlands

<sup>2</sup> Affiliation 2; [e.alreshidi@uoh.edu.sa](mailto:e.alreshidi@uoh.edu.sa), University of Hail, Saudi Arabia

\* Correspondence: [tauseeftwente@yahoo.com](mailto:tauseeftwente@yahoo.com)

**Abstract:** Identifying human actions in complex scenes is widely considered as a challenging research problem due to the unpredictable behaviors and variation of appearances and postures. For extracting variations in motion and postures, trajectories provide meaningful way. However, simple trajectories are normally represented by vector of spatial coordinates. In order to identify human actions, we must exploit structural relationship between different trajectories. In this paper, we propose a method that divides the video into N number of segments and then for each segment we extract trajectories. We then compute trajectory descriptor for each segment which capture the structural relationship among different trajectories in the video segment. For trajectory descriptor, we project all extracted trajectories on the canvas. This will result in texture image which can store the relative motion and structural relationship among the trajectories. We then train Convolution Neural Network (CNN) to capture and learn the representation from dense trajectories. Experimental results shows that our proposed method out performs state of the art methods by 90.01% on benchmark data set.

**Keywords:** Support vector machine, motion descriptor, features, human behaviors

## 1. Introduction

Automatic recognition of human actions play the importance role and is an active topic in computer vision and pattern recognition research [1], [2], [3]. Action recognition has wide range applications in area of automated surveillance [7], [8], [9] [10], in-home elder monitoring [4], [5], [6], and public security [11], [12], [13] and customer behavior [14], [15], [16]. Technically speaking, the task of human action is to identify and recognize actions and behaviors of one or more persons from video sequences. For example, an in-home elder is doing an exercise or doing some other activity by using his/her hands, arms, legs, and other body parts. There are two ways to observe his/her action, 1) bare eyes, 2) camera. We can easily understand and classify his/her action with the bare eyes, since our minds are trained for specific categories. For example, we can easily classify the behavior of a person is walking or running. Such manual classification is tedious and hectic job and we need automatic methods that automatically can identify and classify these actions. For example, in rehabilitation process, it is important to observe and monitor human actions for a long period of time [17], [18]. It is impossible for humans to manually analyze these action for long period of time due to limited capabilities [19][20][58]. Therefore, we need some method that can automate this process [55].

The alternative solution is to develop a virtual machine [21], [22], [23] that can accurately analyze and understand humans' actions from the videos in order reduce to human labor and error. For example, in smart rehabilitation, a virtual analyst with the help of camera can automatically understand and analyze his/her behavior. There are many applications of such virtual analyst. We can prevent the patient from injuries by predicting his/her behavior. Other important applications include video surveillance, entertainment. Motivated by the importance of automatic action recognition methods, several work reported in literature [24], [25], [26], [27] with aim to automatically classify human actions in analyzing a large-scale of video data and provide understanding on the current state and future state.

Cutler et al. [30] [31] detect and identify the periodic motion in extremely low resolution videos. They compute similarity among the different frames of the video sequence and find whether an action is periodic. The limitation of this method is that they only use appearance features while we argue that similarity among the appearance features cannot distinguish the variation of posture. In order to overcome the above limitation, motion gradient information was used to classify actions. Efros et al. [32] introduced a novel motion descriptor which was based on the optical flow and motion similarity. They compute optical flow from consecutive frames and smooth out the optical flow in four separated channels. Spatial-temporal motion descriptors is then computed to identify different actions. Chaudhry et al. [36] proposed a histogram of oriented Optical flow (HOOF) and used Binet-Cauchy kernels to classify human actions. HOOF was very effective in identifying the human action since this method can alleviate the effect of noise, scale and motion variation [59][60]. From the above discussion, it is revealed that representation of motion information play an important role in identifying human actions. In the most recent literature, the fisher vector representation integrated with improved dense trajectory descriptor is most effective way of capturing motion information [56]. Originally, dense trajectory features are computed by sampling and tracking dense points from the each frame in multiple scales, while the improved trajectory feature is computed by estimating the camera motion and replaces the bag-of-features with Fisher vector. With such adaptation, they achieved state of the art performance on several action recognition datasets. However, in both these approaches, trajectories are represented as vector of spatial coordinates, hence losing structural relationship between different trajectories.

We propose a method to encode the trajectories in more effective way. Motivated by the great success of CNNs, we propose a novel trajectory descriptor based on CNN. Our proposed framework has the following steps: 1) Extract dense trajectories from multiple consecutive frames. 2) Project extracted trajectories on the canvas. This will result in a texture image. 3) Base on texture image, we utilize and train a convolutional neural network to learn compact representation of dense trajectories. From the experiments, we quantitatively show that our propose framework outperforms other state-of-the-art methods

## 2. Materials and Methods

The proposed methodology starts by dividing the video into N number of temporal segments. We then compute dense optical flow between multiple consecutive frames using method in [41]. The reason of computing dense optical flow to extract motion information. We then extract trajectories by particle advection approach. After extracting trajectories, we then project those on the canvas resulting in a texture image. We then utilize and train CNN to learn structural relationship from texture images.

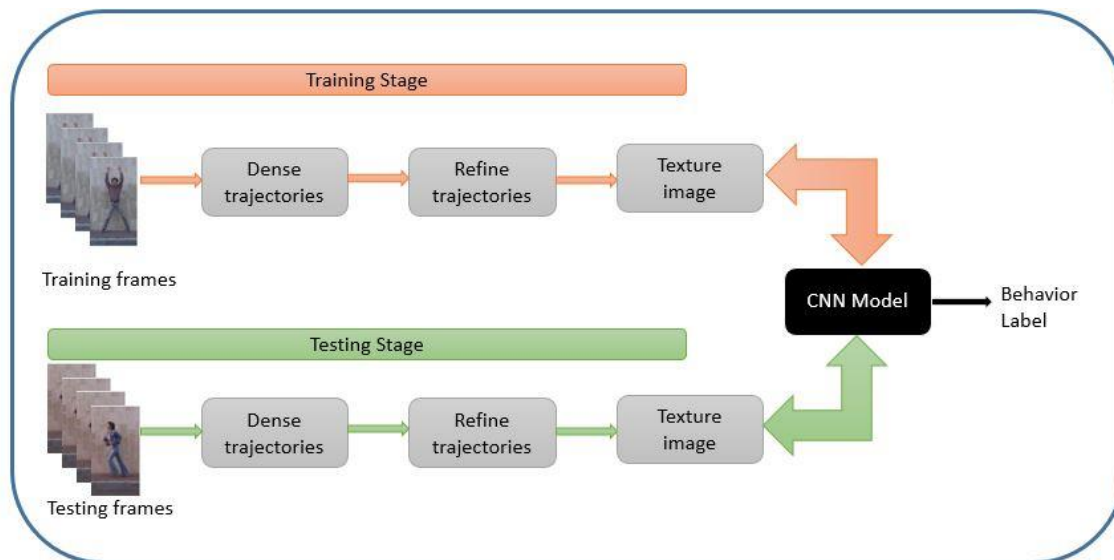


Figure 1: Pipeline of our proposed framework

## 2.1 Optical Flow Computation

The primary step to extract motion information is to compute dense optical flow between multiple consecutive frames. For computing high accurate optical flow we employ methods [42] where gray value consistency, gradient constancy and smoothness in multi-scale constraints. Consider a point  $i$  in the image at time  $t$  of a segment: its flow vector  $Z_{i,t} = (X_{i,t}, V_{i,t})$  includes the location of point and its velocity is represented by  $V_x$  and  $V_y$ . Where  $V_x$  and  $V_y$  represents the gradients in horizontal and vertical directions. After computing optical flow for each pixel, we have now motion field where higher magnitudes corresponds to the foreground and lower magnitude pixels represents the pixels belong to the background.

## 2.2 Particle Advection

After computing optical flow for whole video segment, the next step is to generate dense and long trajectories [43], [44]. We overlay grid of particles over the first optical flow field. In first optical flow field, initial location of each particle is called the source location of the particle. In order to capture better representation of motion, we keep the size of the grid as same as resolution of the frame. The size of particle is same as size of the pixel. However, this arrangement will incur high computational costs. In order to reduce the computation cost and extract better representation of motion, we slightly reduce the resolution of the grid. During the process of advection, we maintain two separate flow maps, one to maintain the horizontal coordinates and other to maintain vertical components. These map in general store the initial and following positions of the point trajectories evolved during advection process.

The trajectories obtained through this process are applicable to structured crowds but for the unstructured crowds [45], [46], where people have un-predicted behaviors, such trajectories do capture the actual motion information and flows. The reason is that in unstructured crowds, the people move in arbitrary directions and in most of the case there is chance that particle belong to a person will lose its path and become the part of different motion pattern [57]. In this case, the trajectory becomes unreliable and erroneous and is not suitable to be considered. In order to tackle the above problem, we modify the above equation in the following way:

$$X_{(i,t+1)} = X_{(i,t)} + F(X_{(i,t)}) * B_i$$

Trajectories obtained using the above equation stop the trajectories drifting from one motion flow to other motion flow. Trajectories obtained by this method precisely capture the motion information in any complex situation.

After trajectories obtained using the above process, some of the trajectories will belong to the foreground while other trajectories will belong to background and noise [47][48][49][61]. In order to make the process efficient and effective, we remove trajectories belong to the background and noisy. To this end, we compute spatial extent of each trajectory by calculating the Euclidean distance between the start and last points of the trajectories. From our experiments, we observed that trajectories belong to noise and background are generally smaller in length. We use this information and set a threshold value on the length of trajectories. We remove those trajectories whose length is less than the specified threshold.

### 2.3 Generating Texture Image

After extracting trajectories, the next step is to generate texture image by projecting trajectories on the canvas. It is worthy to mention that most of the state-of-the-art techniques treat video in 3D space and extract features from spatial and temporal dimension. This method is very effective for small videos but incur high computation cost when large volume of video data is required to be processed. In order to address this problem, we propose a novel way to covert video into a two-dimension space so that can handled and processed easily.

Let Trajectory T is represented by (x, y, u, v) t, where t is the frame number, x and y are the horizontal and vertical coordinates whereas u and v represents the velocity vectors. We then simply project trajectories from the multiple frames onto canvas C according to the following formulation.

$$C_{s,t} = \sqrt{u^2 + v^2}$$

Where s is the canvas index at time t.

### 2.4 Trajectory descriptor based on Convolution Neural Network

In order to learn more compact representation of trajectories, CNN is employed to process the texture image. The main advantage of training CNN with texture image is that it learns structural relationship among trajectories. All texture images extracted from the video segments are stacked together and used as one multi-channel input for CNN. Our CNN model is shown in Figure 2. It consists of four convolution layers. Each convolution layers is followed by ReLU layer. Convolution layers are then followed by max-pooling layers with two normalization layers and three fully connected layers. The main function of the convolution layer is to perform convolution operation on the feature map of the previous layers and extract features from the local neighborhood. Then additive bias is applied and the resulting feature map is passed to ReLU activation function. Generally, the value of a unit at position (x, y) of the jth feature map in the ith layer is given by  $V_{xy}$  using the following formulation.

$$v_{ij}^{xy} = a \left( b_{ij} + \sum_{m=0}^{M-1} \sum_{w=-\Delta}^{\Delta} \sum_{h=-\Theta}^{\Theta} \omega_{ijm}^{wh} v_{(i-1)m}^{(x+w)(y+h)} \right)$$

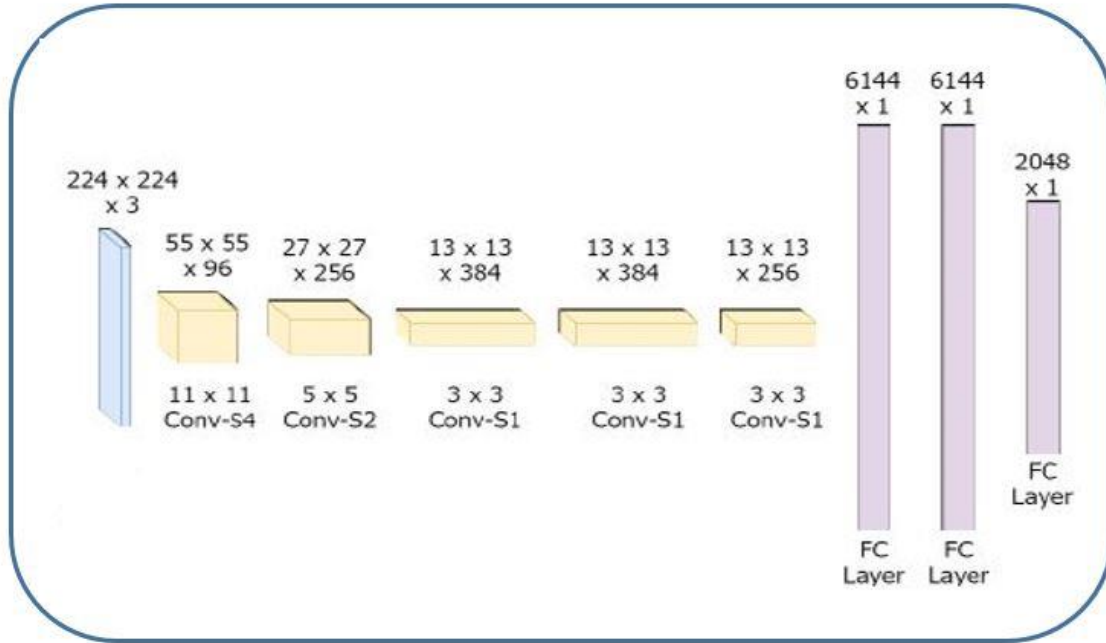


Figure 2: Convolution Neural network for human behavior recognition

Where value of  $\Delta$  is half of width and  $\Theta$  is the half of height of an input image.  $a(.)$  is the activation function,  $b_{ij}$  is the bias of feature map. During training phase, we use gradient descent method. The pooling layers pool over the local neighborhood of the feature map in the previous layer result in reducing the resolution of feature map. In this way, pooling layers enhance the invariance to distortions on the input feature map.

We first train CNN on the training set. After training phase, we re-run the CNN model on the both training and testing sets and ignore the back propagation computation. We store all the parameter values from the fourth convolution layer at every iteration as the final representation of the video segment.

### 3. Results

We use the publicly available benchmark dataset [54] for performance evaluation and comparison. The dataset contains 93 video sequences belong to 10 human actions. Sample frames of the dataset are shown in Figure 3. The actions are: bend, jump, and jack, jump forward- on-two-legs, jump-in-place-on-two-legs, run, gallop sideways, skip, walk, wave-two-hands, and wave one-hand). These actions are performed by 9 different actors. All video sequences in the dataset have the resolution of 180x144 pixels and the length of each video sequence is four seconds with average of 50 fps. The dataset also include extracted foreground, obtained by background subtraction.

In our experiment, we first compute optical flow for only the foreground objects excluding the background in order to reduce the computation time. In our experiments we use fixed configuration of CNN. The size of texture image is fixed to 165 x 165 pixels.

For training, we use 2/3 of the dataset and the rest for testing. In our experiments, in order to rigorously evaluate the performance of our proposed method, we randomly select six sequences in each action as a training set and the rest for testing

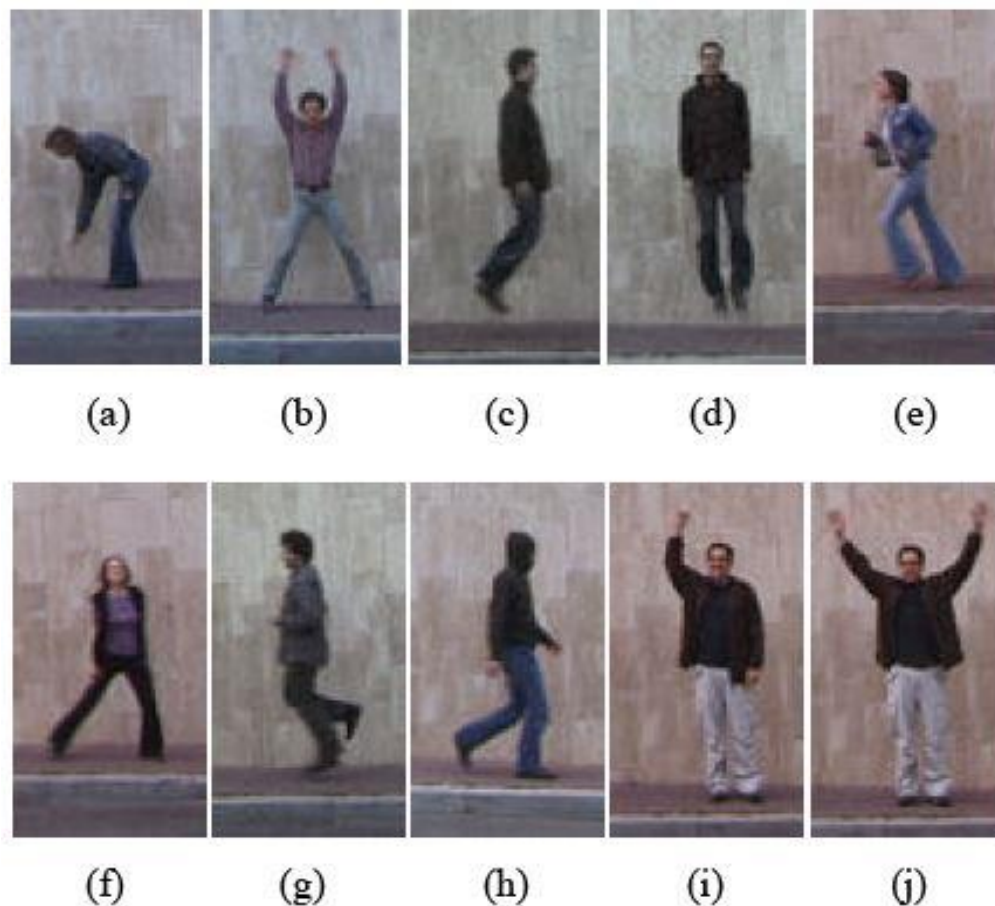


Figure 3. Sample frames from the dataset contain that 10 actions of 9 persons consists of (a) bend, (b) jack, (c) jump, (d) jump, (e) run, (f) side, (g) skip, (h) walk, (i) wave1, (j) wave2

Figure 4, shows that confusion matrix for classification, where we train the classifier on one behavior and test it on the other behaviors. The average classification rate is 89.17%. However, there are some false positives due to the misclassification of bend with pjump. It attributes to the fact similar motion vectors are generated due to the person stand still before and after bending down. In skipping, 3/4 sequences are classified as running. It was expected because their motion and posture are very likely to each other. In addition, the misclassification of pjump with jump and wave2 with wave1 cause by similar pose too.



|       | Bend | Jack | Jump | Pjump | Run  | Side | Skip | Walk | Wave1 | Wave2 |
|-------|------|------|------|-------|------|------|------|------|-------|-------|
| Bend  | 0.87 | 0.00 | 0.00 | 0.33  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00  | 0.00  |
| Jack  | 0.00 | 1.00 | 0.00 | 0.00  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00  | 0.00  |
| Jump  | 0.00 | 0.00 | 1.00 | 0.00  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00  | 0.00  |
| Pjump | 0.00 | 0.00 | 0.33 | 0.67  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00  | 0.00  |
| Run   | 0.00 | 0.00 | 0.00 | 0.00  | 1.00 | 0.00 | 0.00 | 0.00 | 0.00  | 0.00  |
| Side  | 0.00 | 0.00 | 0.00 | 0.00  | 0.00 | 1.00 | 0.00 | 0.00 | 0.00  | 0.00  |
| Skip  | 0.00 | 0.00 | 0.00 | 0.00  | 0.75 | 0.00 | 0.74 | 0.00 | 0.00  | 0.00  |
| Walk  | 0.00 | 0.00 | 0.00 | 0.00  | 0.00 | 0.00 | 0.00 | 1.00 | 0.00  | 0.00  |
| Wave1 | 0.00 | 0.00 | 0.00 | 0.00  | 0.00 | 0.00 | 0.00 | 0.00 | 1.00  | 0.00  |
| Wave2 | 0.00 | 0.00 | 0.00 | 0.00  | 0.00 | 0.00 | 0.00 | 0.00 | 0.67  | 0.33  |

Figure 4. Confusion matrix of different human behaviors

For quantitative analysis and comparisons, we also compare our method with other reference methods and the results are reported in Table 1. The conventional classification methods use the strategy of leave-one out with nearest neighbor while we used hold out method. In our method, the sequences in training set are not used in the testing process. For Leave-one-out with X samples, the model is train on all data except for one sample and test the model with the sample in each time X. We then compute the average error of X time. So every data used in testing once and in training X – 1 times. From the Table 1, it is obvious that our method produces lower error rate compare to other state-of-the-art methods.

| METHODS          | ERROR RATE   |
|------------------|--------------|
| Khan et al [24]  | <b>89.17</b> |
| Ullah et al [43] | <b>95.24</b> |
| Kong et al [3]   | <b>85.69</b> |
| Proposed         | <b>72.11</b> |

#### 4. Conclusion

In this paper, we proposed a deep model approach for recognizing human behaviors. We proposed a novel approach to exploit motion information by projecting trajectories on the canvas. In this way, we generate texture image that captures the structural relationship among different trajectories. We then learn deep trajectory descriptor based on CNN. We demonstrated the capability of our approach in capturing the dynamics of different classes by extracting these features. The main advantage of the proposed method is its simplicity and robustness

## References

1. Wang, Limin, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. "Temporal segment networks: Towards good practices for deep action recognition." In European Conference on Computer Vision, pp. 20-36. Springer, Cham, 2016.
2. Feichtenhofer, Christoph, Axel Pinz, and Richard P. Wildes. "Spatiotemporal multiplier networks for video action recognition." In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7445-7454. IEEE, 2017.
3. Kong, Yu, Shangqian Gao, Bin Sun, and Yun Fu. "Action Prediction From Videos via Memorizing Hard-to-Predict Samples." In AAAI. 2018.
4. Ma, Shugao, Leonid Sigal, and Stan Sclaroff. "Learning activity progression in lstms for activity detection and early detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1942-1950. 2016.
5. Hu, Weiming, Dan Xie, Zhouyu Fu, Wenrong Zeng, and Steve Maybank. "Semantic-based surveillance video retrieval." IEEE Transactions on image processing 16, no. 4 (2007): 1168-1181.
6. Ben-Arie, Jezekiel, Zhiqian Wang, Purvin Pandit, and Shyamsundar Rajaram. "Human activity recognition using multidimensional indexing." IEEE Transactions on Pattern Analysis & Machine Intelligence 8 (2002): 1091-1104..
7. Saqib, Muhammad, Sultan Daud Khan, and Michael Blumenstein. "Texture-based feature mining for crowd density estimation: A study." In Image and Vision Computing New Zealand (IVCNZ), 2016 International Conference on, pp. 1-6. IEEE, 2016.
8. Cutler, Ross, and Larry S. Davis. "Robust real-time periodic motion detection, analysis, and applications." IEEE Transactions on Pattern Analysis and Machine Intelligence 22, no. 8 (2000): 781-796.
9. Efros, Alexei A., Alexander C. Berg, Greg Mori, and Jitendra Malik. "Recognizing action at a distance." In null, p. 726. IEEE, 2003.
10. Fathi, Alireza, and Greg Mori. "Action recognition by learning mid-level motion features." In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1-8. IEEE, 2008.
11. Chaudhry, Rizwan, Avinash Ravichandran, Gregory Hager, and René Vidal. "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions." In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 1932-1939. IEEE, 2009.
12. Ullah, H., Altamimi, A. B., Uzair, M., & Ullah, M. (2018). Anomalous entities detection and localization in pedestrian flows. *Neurocomputing*, 290, 74-86.
13. Ullah, H., Ullah, M., & Uzair, M. (2018). A hybrid social influence model for pedestrian motion segmentation. *Neural Computing and Applications*, 1-17.
14. Ahmad, F., Khan, A., Islam, I. U., Uzair, M., & Ullah, H. (2017). Illumination normalization using independent component analysis and filtering. *The Imaging Science Journal*, 65(5), 308-313.
15. Ullah, H., Uzair, M., Ullah, M., Khan, A., Ahmad, A., & Khan, W. (2017). Density independent hydrodynamics model for crowd coherency detection. *Neurocomputing*, 242, 28-39.
16. Khan, Sultan Daud, Muhammad Tayyab, Muhammad Khurram Amin, Akram Nour, Anas Basalamah, Saleh Basalamah, and Sohaib Ahmad Khan. "Towards a Crowd Analytic Framework For Crowd Management in Majid-al-Haram." arXiv preprint arXiv:1709.05952 (2017).
17. Saqib, Muhammad, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein. "Extracting descriptive motion information from crowd scenes." In 2017 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1-6. IEEE, 2017..
18. Ullah, M., Ullah, H., Conci, N., & De Natale, F. G. (2016, September). Crowd behavior identification. In *Image Processing (ICIP), 2016 IEEE International Conference on*(pp. 1195-1199). IEEE.
19. Khan, S. "Automatic Detection and Computer Vision Analysis of Flow Dynamics and Social Groups in Pedestrian Crowds." (2016).
20. Arif, Muhammad, Sultan Daud, and Saleh Basalamah. "Counting of people in the extremely dense crowd using genetic algorithm and blobs counting." *IAES International Journal of Artificial Intelligence* 2, no. 2 (2013): 51..



21. Ullah, H., Ullah, M., Afridi, H., Conci, N., & De Natale, F. G. (2015, September). Traffic accident detection through a hydrodynamic lens. In Image Processing (ICIP), 2015 IEEE International Conference on (pp. 2470-2474). IEEE.
22. Ullah, H. (2015). Crowd Motion Analysis: Segmentation, Anomaly Detection, and Behavior Classification (Doctoral dissertation, University of Trento).
23. Khan, Sultan D., Stefania Bandini, Saleh Basalamah, and Giuseppe Vizzari. "Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows." *Neurocomputing* 177 (2016): 543-563..
24. Shimura, Kenichiro, Sultan Daud Khan, Stefania Bandini, and Katsuhiro Nishinari. "Simulation and Evaluation of Spiral Movement of Pedestrians: Towards the Tawaf Simulator." *Journal of Cellular Automata* 11, no. 4 (2016).
25. Khan, Sultan Daud, Giuseppe Vizzari, and Stefania Bandini. "A Computer Vision Tool Set for Innovative Elder Pedestrians Aware Crowd Management Support Systems." In *AI\* AAL@ AI\* IA*, pp. 75-91. 2016.
26. Saqib, Muhammad, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein. "A study on detecting drones using deep convolutional neural networks." In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-5. IEEE, 2017..
27. Khan, Sultan Daud, Giuseppe Vizzari, Stefania Bandini, and Saleh Basalamah. "Detection of social groups in pedestrian crowds using computer vision." In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 249-260. Springer, Cham, 2015..
28. Khan, Sultan Daud, Fabio Porta, Giuseppe Vizzari, and Stefania Bandini. "Estimating Speeds of Pedestrians in Real-World Using Computer Vision." In *International Conference on Cellular Automata*, pp. 526-535. Springer, Cham, 2014..
29. Khan, Sultan D., Luca Crociani, and Giuseppe Vizzari. "Integrated Analysis and Synthesis of Pedestrian Dynamics: First Results in a Real World Case Study." *From Objects to Agents* (2013)..
30. Khan, Sultan D., Luca Crociani, and Giuseppe Vizzari. "PEDESTRIAN AND CROWD STUDIES: TOWARDS THE INTEGRATION OF AUTOMATED ANALYSIS AND SYNTHESIS."
31. Ullah, H., Ullah, M., & Conci, N. (2014, September). Dominant motion analysis in regular and irregular crowd scenes. In *International Workshop on Human Behavior Understanding*(pp. 62-72). Springer, Cham.
32. Saqib, Muhammad, Sultan Daud Khan, and Michael Blumenstein. "Detecting dominant motion patterns in crowds of pedestrians." In *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, vol. 10225, p. 102251L. International Society for Optics and Photonics, 2017..
33. Ullah, H., Ullah, M., & Conci, N. (2014, March). Real-time anomaly detection in dense crowded scenes. In *Video Surveillance and Transportation Imaging Applications 2014*(Vol. 9026, p. 902608). International Society for Optics and Photonics.
34. Ullah, H., Tenuti, L., & Conci, N. (2013, March). Gaussian mixtures for anomaly detection in crowded scenes. In *Video Surveillance and Transportation Imaging Applications* (Vol. 8663, p. 866303). International Society for Optics and Photonics.
35. Rota, P., Ullah, H., Conci, N., Sebe, N., & De Natale, F. G. (2013, September). Particles cross-influence for entity grouping. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European* (pp. 1-5). IEEE.
36. Ullah, H., & Conci, N. (2013, September). Structured learning for crowd motion segmentation. In *Image Processing (ICIP), 2013 20th IEEE International Conference on* (pp. 824-828). IEEE.
37. Ullah, H., & Conci, N. (2012, November). Crowd motion segmentation and anomaly detection via multi-label optimization. In *ICPR workshop on Pattern Recognition and Crowd Analysis*.
38. Khan, W., & Ullah, H. (2010). Authentication and Secure Communication in GSM, GPRS, and UMTS Using Asymmetric Cryptography. *International Journal of Computer Science Issues (IJCSI)*, 7(3), 10.
39. Ullah, H., Ullah, M., Uzair, M., & Rehman, F. (2010). Comparative study: The evaluation of shadow detection methods. *International Journal Of Video & Image Processing And Network Security (IJVIPNS)*, 10(2), 1-7.
40. Khan, W., & Ullah, H. (2010). Scientific Reasoning: A Solution to the Problem of Induction. *International Journal of Basic & Applied Sciences*, 10(3), 58-62.
41. Uzair, M., Khan, W., Ullah, H., & Rehman, F. U. (2009, December). Background modeling using corner features: An effective approach. In *Multitopic Conference, 2009. INMIC 2009. IEEE 13th International* (pp. 1-5). IEEE.

42. Ullah, M., Ullah, H., & Alseadoon, I. M. HUMAN ACTION RECOGNITION IN VIDEOS USING STABLE FEATURES.
43. Khan, W., Ullah, H., & Hussain, R. (2013). Energy-Efficient Mutual Authentication Protocol for Handheld Devices Based on Public Key Cryptography. *International Journal of Computer Theory and Engineering*, 5(5), 754.
44. Khan, A. (2015). Coordinated Unmanned Aerial Vehicles for Surveillance of Targets (Doctoral dissertation, Queen Mary University of London).
45. Khan, A., Rinner, B., & Cavallaro, A. (2015, September). Multiscale observation of multiple moving targets using micro aerial vehicles. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on* (pp. 4642-4649). IEEE.
46. Ahmad, K., Mekhalfi, M. L., Conci, N., Melgani, F., & Natale, F. D. (2018). Ensemble of Deep Models for Event Recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2), 51.
47. Ahmad, K., Conci, N., & De Natale, F. G. (2018). A saliency-based approach to event recognition. *Signal Processing: Image Communication*, 60, 42-51.
48. Arif, Muhammad, Sultan Daud, and Saleh Basalamah. "People counting in extremely dense crowd using blob size optimization." *Life Science Journal* 9, no. 3 (2012): 1663-1673.
49. Saqib, Muhammad, S. D. Khan, and S. M. Basalamah. "Vehicle Speed Estimation using Wireless Sensor Network." In *INFOCOMP 2011 First International Conference on Advanced Communications and Computation, IARIA*. 2011.
50. Khan, Sultan Daud. "Estimating Speeds and Directions of Pedestrians in Real-Time Videos: A solution to Road-Safety Problem." In *CEUR Workshop Proceedings*, p. 1122. 2014.
51. Khan, Sultan Daud, and Hyunchul Shin. "Effective memory access optimization by memory delay modeling, memory allocation, and buffer allocation." In *SoC Design Conference (ISOCC), 2009 International*, pp. 153-156. IEEE, 2009.
52. Khan, Sultan Daud, Giuseppe Vizzari, and Stefania Bandini. "Facing Needs and Requirements of Crowd Modelling: Towards a Dedicated Computer Vision Toolset." In *Traffic and Granular Flow'15*, pp. 377-384. Springer, Cham, 2016.
53. Saqib, Muhammad, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein. "Person Head Detection in Multiple Scales Using Deep Convolutional Neural Networks." In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7. IEEE, 2018.
54. Khan, S., Giuseppe Vizzari, Stefania Bandini, and Saleh Basalamah. "Detecting dominant motion flows and people counting in high density crowds." *Journal of WSCG* 22, no. 1 (2014): 21-30.
55. Khan, Sultan Daud, and Habib Ullah. "A survey of advances in vision-based vehicle re-identification." *Computer Vision and Image Understanding* (2019).
56. Ullah, Habib, Muhammad Uzair, Arif Mahmood, Mohib Ullah, Sultan Daud Khan, and Faouzi Alaya Cheikh. "Internal Emotion Classification Using EEG Signal with Sparse Discriminative Ensemble." *IEEE Access* (2019).
57. Saqib, Muhammad, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein. "Crowd Counting in Low-Resolution Crowded Scenes Using Region-Based Deep Convolutional Neural Networks." *IEEE Access* 7 (2019): 35317-35329.
58. Ullah, M., Cheikh, F. A., & Imran, A. S. (2016, August). Hog based real-time multi-target tracking in bayesian framework. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 416-422). IEEE.
59. Ullah, M., Mohammed, A., & Alaya Cheikh, F. (2018). PedNet: A Spatio-Temporal Deep Convolutional Neural Network for Pedestrian Segmentation. *Journal of Imaging*, 4(9), 107.
60. Ullah, M., & Alaya Cheikh, F. (2018). A Directed Sparse Graphical Model for Multi-Target Tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1816-1823).
61. Ullah, M., Mohammed, A. K., Cheikh, F. A., & Wang, Z. (2017, September). A hierarchical feature model for multi-target tracking. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 2612-2616). IEEE.