Classification of Moving Crowd Based on Motion Pattern

Aneek Roy
Jadavpur University
Kolkata, India
aneek.roy5@gmail.com

Nimagna Biswas
Jadavpur University
Kolkata, India
nimagna0072@gmail.com

Sanjoy Kumar Saha
Jadavpur University
Kolkata, India
sks_ju@yahoo.co.in

Bhabatosh Chanda
Indian Statistical Institute
Kolkata, India
chanda@isical.ac.in

***Abstract*—Crowd behavior analysis is a significant task in the context of surveillance and crowd management. For a moving crowd, analyzing the motion pattern is very important. In this work, we present a simple scheme to categorize such crowds as *structured*, *semi-structured* and *unstructured* ones. The categorization is achieved based on the regularity of the motion pattern of the collection of objects (humans, in this case). In case of structured one, the movement is coherent and uniform in nature. It is expected that the crowd as a whole or individual segment of it reflects consistent orientation and speed of movement. For unstructured crowd, on the other hand, the movement is random. Hence, diversity is there in terms of orientation and speed. The semi-structured one stands in between and makes the classification problem difficult. In this work motion orientation based feature is computed to represent the motion pattern. A set of interest points detected in the initial frame are tracked over the sequence using optical flow. Thus, motion orientations are obtained. A frame is divided into blocks, and distribution of the orientation of motion of the interest points in each block is summarized in a four dimensional histogram. Block level histograms are concatenated to form the frame level descriptor. Finally, frame level descriptors are taken together to represent the sequence. In this experiment, artificial neural network (ANN) is used as classifier. Experiment is carried out on collectiveness dataset. Proposed method provides better classification accuracy in comparison to state-of-the-art techniques.**

***Index Terms*—Crowd Behaviour Analysis, Crowd Classification, Crowd Motion Analysis**

I. Introduction

In the context of video surveillance study of crowd motion has gained tremendous importance. Crowd is the collection of people where the individuals loose their distinct identity and their motion is dominated by the overall crowd flow [1], [2]. The movement of the crowd influences that of each individual; however, on the other hand, the movement of each individual contributes towards and thus helps the global motion pattern of the crowd to take its shape. Manual observation of crowd movement is quite tedious and boring from the observer point of view, and may lead to relaxed and intermittent monitoring. For an automated system to keep vigil on moving crowd, a fundamental step is to characterize the crowd as a whole. Based on the motion pattern a crowd can be broadly categorized as *structured* crowd and *unstructured* ones. In between a *semi-structured* category may also be considered.

Crowd movement is a reflection of social behaviour that follows certain principle [3]. This social aspect imposes a tendency to move in a coherent manner following the path of least resistance subjected to the constraints put by the environment and global flow. As a result an individuals tends to follow a path in the crowd with motion components that are similar to motion of the crowd itself. However, individual has his/her own characteristic behaviour and movement pattern as well as objectives. Eventuality this introduces deviation into the crowd flow and may lead to change in the crowd motion pattern. In a structured moving crowd, collection of people as a whole or the individual parts of it moves coherently with a regularity in local orientation and speed of the movement. Under normal scenario movement in the defined track imposed by environmental conditions like elevators, staircase, crosswalks gives rise to structured motion. Movement of a disciplined crowd like march past by the soldiers is an ideal example of structured behaviour. On the contrary, collection of random motions of a group of individuals results in unstructured category. The motion pattern for semi-structured crowd is neither too random nor strongly uniform. Automatic classification of crowd based on their motion pattern can help crowd management as well as determining subsequent action of disaster response team or rapid action force.

Rest of the paper is organized as follows. Section II presents a brief review of past work. Proposed methodology is elaborated in Section III. Experimental results are placed in Section IV and concluded in Section V.

II. Past Work

Computer vision community has focused in the area of crowd behaviour and automated scene analysis since last decade.Crowd analysis can be done based on local (microscopic) level feature or global level (macroscopic) features. At the global level, the global behaviour of the crowd or a group of people is tracked, without focusing on any particular individual. At microscopic level behaviour of single element is important and it can be effectively applied on low density population [4], [5]. For high density crowd, tracking an individual is either impossible or very difficult. So one will have to rely on macroscopic features.

Ihaddadene and Djeraba [6] proposed a non-parametric method considering crowd density, velocity and direction. Finally, statistical measurements are used to estimate abnormality of a crowd. The local contextual information is ignored

978-1-7281-0297-9/19/$31.00 ©2019 IEEE

in their process. Optical-flow based motion descriptors [7]–[9] are widely used to represent complex crowd flow in the scenes. Cong et al. [10] proposed a novel feature descriptor called multi-scale histogram of optical flow (MHOF), where spatial contextual information is preserved along with motion information. Ozturk et al. [11] considered motion based features using SIFT features flow vectors. A frame is divided into number of blocks for which local dominant flow vectors are computed and finally summarized into a global dominant flow. Based on local and global flows, crowd is characterized.

Ali et al. [12] proposed a floor field based model for structured crowd analysis, where they considered that crowd movement is constrained by the external environment. The scene layout like exit gates, no go zone walls etc. were considered in defining various fields like static force field, dynamic force field, boundary force field and scene specific force field. These force fields are quite elemental and effective for defining the various external forces acting on the crowd as a whole to influence its motion behaviour. For unstructured crowd analysis, Rodriguez et al. [13] deployed the concept of correlated topic model (CTM) where correlation corresponds to low level quantized motion features and topics correspond to crowd behaviours.

Zhou et al. [14] in their work termed cumulative behavioural pattern of crowd as *collectiveness*. It indicates the degree of individual acting as a union in collective motion. Li et al. [15] followed similar concept and proposed a measure for *collectiveness* using refined topological similarity (RTS). It looks for the path similarity of pairs of tracked feature points based on their velocity correlation and spatial information. Ren Weiya [16] followed graph based method to measure the *collectiveness*. The motion coherence between two nodes of clique is considered as the measure. All such measures require tracking of large number of points and are computationally very expensive.

III. Proposed Methodology

In the context of crowd behaviour analysis there may be several problem areas like identification of anomaly or abnormality, tracking of individuals or groups, recognizing the activity. In our work, we focus on macro level categorization of moving crowd into three categories namely *structured*, *semi-structured* and *unstructured*. If a crowd as a whole maintains a regularity in motion or its segments individually coherent in terms of motion then it is considered as structured crowd. Crowd with random motion is termed as unstructured one. Whereas for a semi-structured, behaviour is neither strictly disciplined nor too random. Proposed methodology relies on motion based descriptor to classify the moving crowd into such categories.

The approaches to analyze the behaviour of moving crowd can be broadly categorized as *top-down* and *bottom-up* [17]. In bottom up approach the motion of individual is taken into account to model the crowd behaviour. On the other hand, global features of the crowd motion are used in top-down approach. Both have their own merits and demerits. Difficulty in tracking individual entity in a crowd (more so for a dense one) is the major problem in bottom-up approach [2], [17]. Moreover, it is time consuming. On the other hand, in a top-down approach localized pattern in the different segments of the crowd is ignored. Proposed method follows an intermediate approach, where a set of interest points are tracked over the sequence. The distribution of motion orientation of interest points is studied for different blocks in a video frame. It helps to capture the local pattern to an extent. A video data is divided into number of sequences with fixed duration. Each such sequence is then categorized. Major steps of the work are: *identification of interest points*, *tracking of interest points*, *computation of descriptor for the sequence*, and *classification of the sequence*. Overall block diagram is shown in Figure 1 and the steps are detailed as follows.

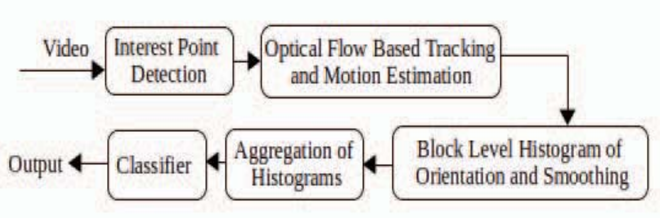
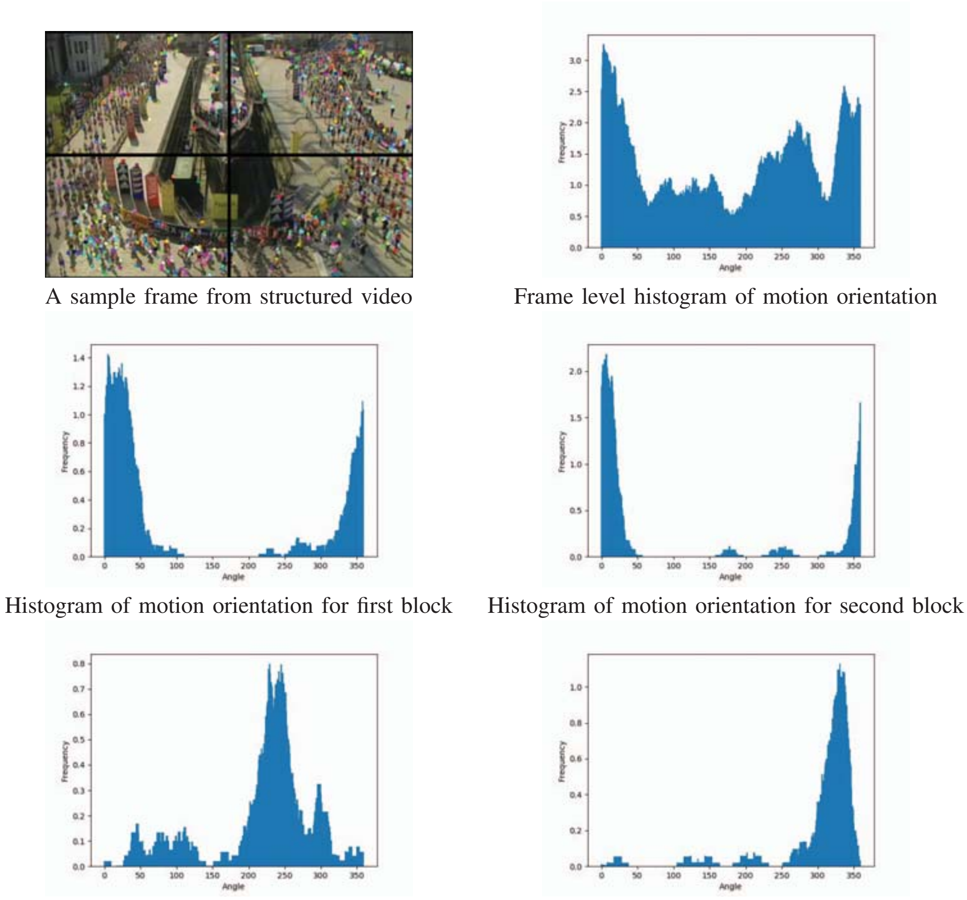

Video
Interest Point Detection
Optical Flow Based Tracking and Motion Estimation
Block Level Histogram of Orientation and Smoothing
Aggregation of Histograms
Classifier
Output

Fig. 1. Overall Block Diagram

Identification of interest points: Interest points are identified from the first frame of the sequence under study. At first corner points are detected by using Shi-Tomasi [18] algorithm. The algorithm works on the principle of Harris corner [19] detector with slight variation in the selection criteria. Spatial differential of intensity values is taken into account to detect sharp changes in pixel intensities along horizontal and vertical direction. A small patches in the image are considered. Score is assigned to the patches based on the variation between the patch and its neighbourhood. Harris corner detector and Shi-Tomasi detector differ in computing the score. Finally, the patches with score higher than a threshold are taken as the interest points. In our work, we restrict ourselves to top N strong points based on the score.

Tracking of interest points: In order to capture the motion behaviour we track only the detected interest points following optical flow algorithm [20]. Consecutive frames are considered for tracking of interest points. Tracking begins with N interest points (taken as 500 in our work) in the first frame. All the interest points of previous frame may not be available in a later point of time as some of those may exit the frame. At subsequent time instance, only mapped points are tracked. Thus, mapped Interest points in i-th frame are tracked in $(i+1)$-th frame. Based on the tracking, motion parameters are computed. To ensure sufficient data for motion characterization, it is essential to have considerable interest points available in each frame. Hence, in case the number of such points for a frame falls below a threshold at an instance, the interest points are refreshed by applying Shi-Tomasi algorithm. This also allows us to include new points entering into the sequence.



A sample frame from structured video

Frame level histogram of motion orientation

Histogram of motion orientation for first block

Histogram of motion orientation for second block

Histogram of motion orientation for third block

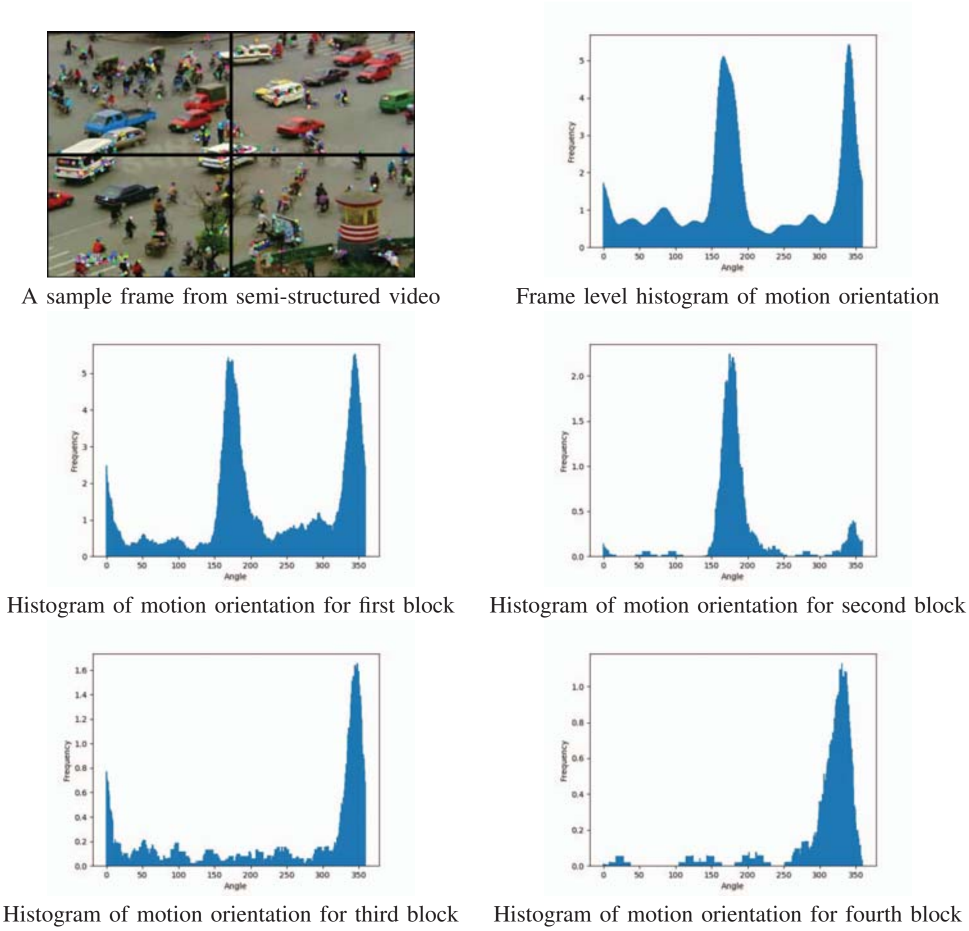
Histogram of motion orientation for fourth block

Fig. 2. Illustrates frame level and block level histograms of motion orientation for a frame in structured motion video. Original frame (at top-left) is divided into 4 equal blocks and are numbered row-wise.

In our experiment the threshold is taken as $0.75 \times N$.

Computation of descriptor for the sequence: To form the frame level descriptor, say, at the i-th frame, the tracked interest points in a pair of consecutive frames, i.e., i-th and $(i+1)$-th frames are considered. Motion direction for an interest point is obtained from the its positions in consecutive frames. Let an interest point is at (x, y) position in i-th frame and it is tracked at position $(x+u, y+v)$ in the next frame. Then the motion orientation vector for the point in i-th frame is taken to be $(u, v)^T$ and orientation angle θ is computed by an inverse trigonometric function of (u, v). Ideally there should be less variation in the orientation angle θ for a structured sequence and it increases with randomness in behaviour. A global histogram of orientation for a frame should usually reflect the pattern. But it fails to capture the local information. Even in a structured sequence, behaviour of individual segments (block) of crowd image may have uniformity, but inter segment heterogeneity can exist. Original image of a structured crowd scene as shown in Figure 2 exemplifies such a situation. Thus, a global histogram at the frame level may loose its discriminating capability. To combat such scenario, we divide the frame into K blocks. Thereafter, the block level histograms of θ are formed and concatenated to form the frame level descriptor. Block level histogram is prepared by considering only the interest points within the block. Such a modified global histogram (prior to quantization) for a frame of structured video is shown in Figure 2. The video frame depicts that moving crowd is taking a turn resulting into different motion orientations in different segments. The frame level histogram also shows significant contribution over a wide range of orientation. Thus, it deviates from expected feature of structured sequence. For visualization the frame is divided into 4 blocks, and numbered row-wise. Block level histograms shown in bottom two rows of Figure 2 reflect that the localized patterns are more uniform in nature. A careful study of the nature (or profile) of frame level and block level histograms justifies the division of a frame into blocks and computing histograms at block level. Figure 3 and 4 show the same for semi-structured and unstructured samples. Frame level histograms for structured and unstructured samples are quite distinct; whereas for semi-structured one it bears similarity with both. However, block level histograms reveal more distinctive characteristics for all the three scenarios.

In reality, we need not discriminate between minor changes



A sample frame from semi-structured video

Frame level histogram of motion orientation

Histogram of motion orientation for first block

Histogram of motion orientation for second block

Histogram of motion orientation for third block

Histogram of motion orientation for fourth block

Fig. 3. Illustrates frame level and block level histograms of motion orientation for a frame in semi-structured motion video. Original frame (at top-left) is divided into 4 equal blocks and are numbered row-wise.

in orientation. To address the issue, Gaussian function is applied on the histogram elements to distribute the impact of angle θ over the range $\theta - 90$ to $\theta + 90$ and summing up the distributed contributions modified histogram is obtained. The orientation range (*i.e.*, angle) is further quantized into four bins denoting the four quadrants. The modified histogram is then summarized into four dimensional histogram. The steps for computing the frame level descriptor are summarized as follows.

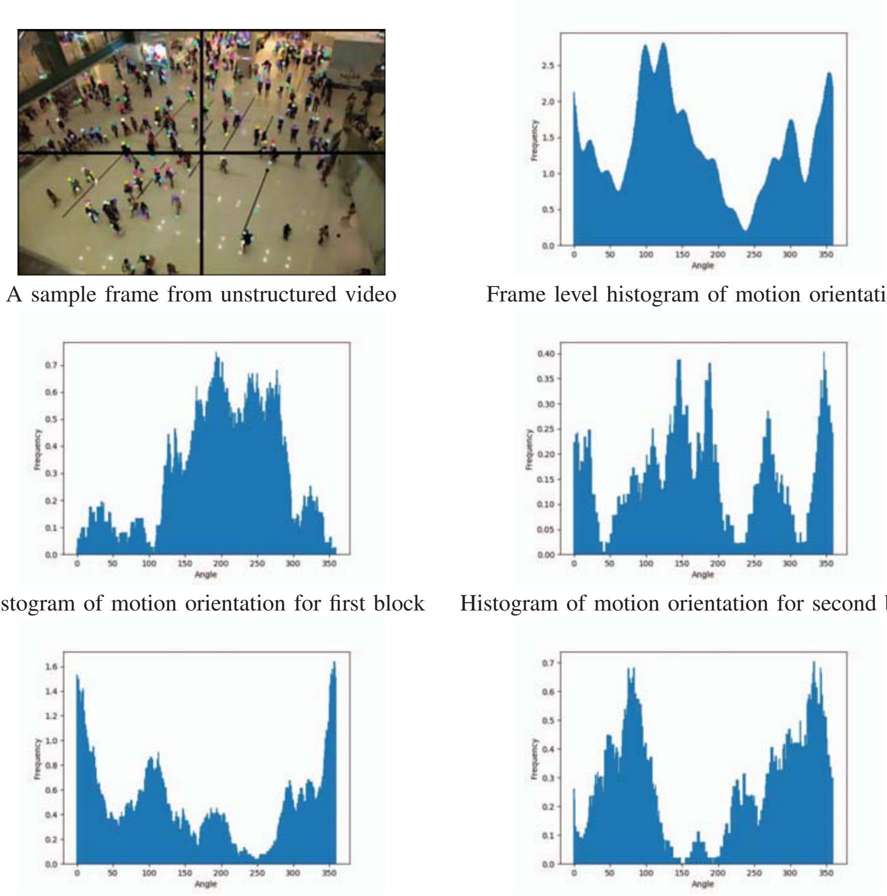
- Obtain the orientation (angle) of motion for interest points based on the tracking outcome of successive frames.
- Divide the frame into K blocks and prepare angle histogram at the block level.
- Apply moving Gaussian averaging on the histogram elements and recompute the histogram.
- Quantize the angle range according to quadrants and form four dimensional block level histogram.
- Concatenate block level histograms to form frame level descriptors.

Frame level histograms for all the frames are flattened into a sequence, and then these are concatenated to generate the sequence descriptor of dimension $4 \times K \times F$ where K and F represent number of blocks in a frame and number of frames in the sequence respectively.

Classification of the sequence: Sequence descriptors are fed to neural network based classifier. Motivated by the findings of Hinton et al. [21], we have considered multiple hidden layers (four in our case). Number of nodes in input and output layers are same as the dimension of input vector and number of classes respectively. Intermediate layers were designed to have half the number of nodes of the previous layer. For the intermediate layers, the rectified linear unit function (Relu) is used as activation function. However, in the output layer Softmax [22] function is used, which helps in determining the probability of the input to belong to a particular class. Back propagation [23] technique is used to train the model considering binary cross-entropy [24] as the loss function. For optimization, rmsprop [25] is used.

IV. Experimental Results

In this work, Collectiveness dataset [14] is used to carry out the experiment. The said dataset contains 413 crowd video sequences and these already have ground truth information.



A sample frame from unstructured video

Frame level histogram of motion orientation

Histogram of motion orientation for first block

Histogram of motion orientation for second block

Histogram of motion orientation for third block

Histogram of motion orientation for fourth block

Fig. 4. Illustrates frame level and block level histograms of motion orientation for a frame in unstructured motion video. Original frame (at top-left) is divided into 4 equal blocks and are numbered row-wise.

TABLE I
CONFUSION MATRIX (FIGURES IN %)

Detected / Actual	Unstruct.	Semi-struct.	Struct.
Unstruct.	80.23	12.79	06.98
Semi-struct.	11.43	79.05	09.52
Struct.	02.56	05.13	92.31
Overall			86.01

That means the sequences have already been categorized as *structured*, *semi-structured* and *unstructured* based on the opinion of ten raters. Details are given in [14]. There are 92 unstructured, 110 semi-structured and 211 structured crowd motion sequences. Each sequence consists of 100 frames.

In order to compute the descriptor, Frame is divided into K blocks. A very low value of K is almost as good as global histogram. On the other hand, very high value represents local behaviour. Hence we have chosen a moderate value and it is taken as 32. We have classified the sequences into three categories following the proposed methodology. Randomly chosen 90% data (i.e., sequences) of each category is used to train the classifier and remaining 10% is used for testing. This is done 10 times and average performance is computed. Table I reports the average classification outcome in the form of confusion matrix. Note that structured motion can be better recognized compared to unstructured or semi-structured. This is because tracking is more accurate in case of structured motion than the other two categories. This is a rational argument as a systematic algorithm for tracking may fail to locate a position that is randomly moved to. The explanation is strengthened if we look at the mis-classification between unstructured and semi-structured motions. It is observed that confusion mostly arises with semi-structured class. This is expected as semi-structured motion contains some parts performing structured motion; while other parts unstructured. A loss of balance between these two, may push the sequence to be identified as unstructured or structured. However, accuracy for the individual classes as well as the overall accuracy are quite satisfactory. Instead of block level histograms, we tried to work with global histogram. But the performance was much poor with 65% as the overall accuracy. To quantize the orientation, instead of 4 bins, we considered higher number of bins (4 and 8). It resulted into marginal improvement in accuracy (around 1%) but computational cost increased significantly.

We have also compared the performance of our proposed method with that of Zhou et al. [14] and Ren Weiya [16]. In both the works, classification accuracy has been reported by taking two classes at a time, not three together. Performance of the methods decreases as difficulty level in distinguishing between the classes increases. We have also measured the performance of our proposed work following the same strategy. Comparative results are shown in Table II, which indicates that overall performance of the proposed methodology is better or at least comparable. Both the works [14], [16] are built upon the interaction between interest points and their surroundings. Thus it captures localized information and spurious interest points may affect the performance particularly for close categories. Proposed descriptors are neither too local nor too global and hence has an edge. Moreover, proposed methodology is quite simple in comparison to those works and can work in real time.

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY (IN [0, 1])

	Unstruct. vs Semi-struct.	Semi-struct. vs Struct.	Unstruct. vs Struct.
Proposed Methodology	0.92	0.89	0.95
Methodology of Zhou et al. [14]	0.79	0.84	0.95
Methodology of Ren Weiya [16]	0.86	0.86	0.99

V. CONCLUSION

In this work a simple methodology has been proposed to categorize video sequences of moving crowd. Based on the motion pattern such crowds are classified either as structured or semi-structured or unstructured ones. Interest points detected in the first frame of the sequence are tracked over the sequence using optical flow. Thus, it requires tracking of only a subset of points in the frame. Based on the motion orientation of such tracked points descriptor is computed. By concatenating the block level histograms of motion orientation frame level feature has been computed. Thus it can well capture the localized motion patterns present in the segments of crowd. Frame level features are concatenated to represent the sequence. Finally, a neural network with multiple hidden layers has been used to classify the sequences. Experimental result on a benchmark dataset indicates that the proposed methodology performs satisfactorily and comparison with one state-of-the-art technique shows the superiority of the proposed work. As the descriptor is defined based tracking of points structured motion can be better recognized that unstructured or semi-structured motion. It should be noted that semi-structured motion creates the maximum confusion.

REFERENCES

[1] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, “Crowded scene analysis: A survey,” *IEEE transactions on circuits and systems for video technology*, vol. 25, no. 3, pp. 367–386, 2015.

[2] S. Lamba and N. Nain, “Crowd monitoring and classification: a survey,” in *Advances in Computer and Computational Sciences*. Springer, 2017, pp. 21–31.

[3] G. K. Still, “Crowd dynamics,” Ph.D. dissertation, University of Warwick, 2000.

[4] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, “A system for learning statistical motion patterns,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 9, pp. 1450–1464, 2006.

[5] B. T. Morris and M. M. Trivedi, “A survey of vision-based trajectory learning and analysis for surveillance,” *IEEE transactions on circuits and systems for video technology*, vol. 18, no. 8, pp. 1114–1127, 2008.

[6] N. Ihaddadene and C. Djeraba, “Real-time crowd motion analysis,” in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[7] S. Ali and M. Shah, “A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis,” in *Computer Vision and Pattern Recognition (CVPR), 2007 IEEE Conference on*. IEEE, 2007, pp. 1–6.

[8] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*. IEEE, 2009, pp. 935–942.

[9] X. Wang, X. Yang, X. He, Q. Teng, and M. Gao, “A high accuracy flow segmentation method in crowded scenes based on streakline,” *Optik International Journal for Light and Electron Optics*, vol. 125, no. 3, pp. 924–929, 2014.

[10] Y. Cong, J. Yuan, and J. Liu, “Abnormal event detection in crowded scenes using sparse representation,” *Pattern Recognition*, vol. 46, no. 7, pp. 1851–1864, 2013.

[11] O. Ozturk, T. Yamasaki, and K. Aizawa, “Detecting dominant motion flows in unstructured/structured crowd scenes,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3533–3536.

[12] S. Ali and M. Shah, “Floor fields for tracking in high density crowd scenes,” in *European conference on computer vision*. Springer, 2008, pp. 1–14.

[13] M. Rodriguez, S. Ali, and T. Kanade, “Tracking in unstructured crowded scenes,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1389–1396.

[14] B. Zhou, X. Tang, and X. Wang, “Measuring crowd collectiveness,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3049–3056.

[15] X. Li, M. Chen, and Q. Wang, “Measuring collectiveness via refined topological similarity,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 2, p. 34, 2016.

[16] W. Ren, “Crowd collectiveness measure via graph-based node clique learning,” *arXiv preprint arXiv:1612.06170*, 2016.

[17] J. C. S. J. Junior, S. R. Musse, and C. R. Jung, “Crowd analysis using computer vision techniques,” *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 66–77, 2010.

[18] J. Shi and C. Tomasi, “Good features to track,” in *Computer Vision and Pattern Recognition, 1994. Proceedings (CVPR), 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[19] C. Harris and M. Stephens, “A combined corner and edge detector.” in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.

[20] B. D. Lucas, T. Kanade *et al.*, “An iterative image registration technique with an application to stereo vision,” 1981.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[22] R. A. Dunne and N. A. Campbell, “On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function,” in *Proc. 8th Aust. Conf. on the Neural Networks, Melbourne*, vol. 181, 1997, p. 185.

[23] Y. LeCun, D. Touresky, G. Hinton, and T. Sejnowski, “A theoretical framework for back-propagation,” in *Proceedings of the 1988 connectionist models summer school*. CMU, Pittsburgh, Pa: Morgan Kaufmann, 1988, pp. 21–28.

[24] J. Shore and R. Johnson, “Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy,” *IEEE Transactions on information theory*, vol. 26, no. 1, pp. 26–37, 1980.

[25] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.