CROWD ATTRIBUTES RECOGNITION AND ALERTING USING SPARSE RECOGNITION

Arya Mohan.S.[1], Danya Mercline[2]

PG Student [1](M.E-Applied Electronics) , Assistant professor [2]

Department of Electronics and Communication Engineering, Narayanaguru College of Engineering,Manjalumoodu, TamilNadu

Abstract— Human behavior analysis has become a critical area of research in computer vision and artificial intelligence research community. In recent years, video surveillance systems of crowd scenes have witnessed an increased demand in different applications, such as safety, security, entertainment, and personal mental health. Although many methods have been proposed, certain limitations exist, and many unresolved issues remain open. In this work, the proposed novel-temporal sparse coding representation, based on sparse coded features with k-means singular value decomposition for robust classification of crowd behaviors has been considered. Extensive experiments have shown that, with sparsely coded features captured with vital structures of video scenes yields discriminant descriptors for classifications than conventional bag-of-visual-features. Relying on the measurable features of crowd scenes and motion characteristics, it can be used to represent different attributes of crowd behaviors. Experiments on hundreds of video scenes were carried out on publicly available datasets. Quantitative evaluation indicates that, the proposed method display superior accuracy, precision and recall in classifying human behaviors with linear support vector machine when compared with the state-of –the-art methods. The detected image will be processed by a processing unit and then an alert can be send to the authorized person through Gmail.

Keywords— Human behavior, crowd scenes, histogram of optical flow, histogram of oriented gradient, artificial intelligence, sparse coding, bag-of-visual features.

INTRODUCTION

In recent years, recognizing human activity in physical environment finds importance in different applications in areas such as intelligent video surveillance[14], and healthcare surveillance[15]. However, accurate detection and recognition of human activity remain an open problem due to the background noise, occlusions, scale and view point changes. Moreover, the interest in the field of video surveillance technology[7],[12], is increasing with the availability of low-cost sensors and processors; especially in order to understand human behavior. As a result, a large volume of video data containing excessive behavior information is available, which is not adequately analyzed by human operators. The field of intelligent crowd surveillance recently received an increase in global funding and attention, because of its usefulness in the monitoring of public areas: shopping centers, banks, airports, train stations, subway stations, sports areas, and traffic control and congestion prediction. Crowd analysis can be sub-classified based on the following area of Applications: crowd behavior, crowd segmentation, crowd tracking, crowd motion detection and crowd density estimation. A substantial analysis of the existing reviews in [7] and [6] reveals a wider perspective on the potential application of computer vision and artificial intelligence in the efficient representation of human behavior in crowded scenes.

There are different contextual terms used in the definition of human behavior including atomic actions, action sequences, and activity[3]. Atomic actions correspond to instantaneous entities upon which an action is formed. Actions correspond to a sequence of atomic actions that fulfils a task or purpose; while activities consists of a sequence of actions over space and time. The overall objective of any behavior recognition system is to detect, analyze and interpret the contextual activity. To bring about intelligent aid to human capability, the crowd analysis aims to create an efficient platform with the capability of mimicking the human intelligence by learning and classifying behaviors in normal and abnormal context. Despite many efforts, detecting appearance and dynamic information from a crowd scene remains challenging in activity analysis. Theoretically, in most activity analysis findings, objects of interests are extracted from the background, and followed by tracking the object. One can inferred that the foreground extracted is used in holistic behavior understanding. Unluckily, this mainstream representation is considered too difficult for crowd videos analysis.

Human behavior analysis in crowd offers opportunities for a wide range of applications; for instance, visual surveillance, video indexing, searching, robotics, health-care, animation, and gaming. Visual surveillance plays a vital role in the monitoring of the human activities in realistic scenes using visual sensors; such as Closed-Circuit Television (CCTV)cameras. The traditional surveillance systems essentially depend on human operates to monitor people activities, a region of interest tracking from one camera to another, or detecting of anomaly occurrence in a scene. However, the majority of the miss-detected incidences in crowd scenes are due to the manual handling of the system and poor observation by the CCTV operators. These shortcomings could be due to excessively displayed video screens, boredom due to prolonged observation, and lack of basic knowledge in classifying behaviors, and distractions by additional operational responsibilities; consequently, important events may be miss-detected in critical surveillance task[3]. A human behavior analysis system tends to solve the aforementioned problems, by providing automated high-performance framework, which assists human operators to achieve an effective management of crowds.

A novel spatio-temporal sparse coding representation**,** based on Sparse Coded features with K-means Singular Value Decomposition(SCKSVD) for robust classification of human behavior especially student for learning, improve accuracy and simplify model structure. The notion of space-time features used is [17] extended, which was formulated based on 3D interest points and bag-of-visual feature. The bag-of-visual feature was used for quantization and estimating the motion and appearance gradients of a video scene. However, among the limitations of the Bag-of-visual feature, it was observed that spatial relationship among the video patches was poorly captured which resulted to low recognition rate in [17] . The SCKSVD models was evaluated on the following crowd datasets; University of Central Florida (UCF), Chinese University of Hong Kong (CUHK), and University Technology PETRONAS Crowd dataset(Crowd-UTP),in comparison with baseline methods.

RELATED WORKS

In this section, the review of existing studies on crowd analytic models are presented. Existing works on crowd analysis can be categorized into holistic methods, particle/track let based methods, trajectory based and spatio-temporal models.

A. HOLISTIC METHODS

Holistic methods analyze the crowd pattern in global perspective[4],[10] and [11]. Chan and Vasconcelos in[8] analyzed crowd setup based on dynamic textures models, which represent video sequences as observations from a linear dynamical system. Mahadevan *et al.* in extend the concept of dynamic texture models to anomaly detection in crowd scenes. There are also works that utilize local level features, e.g., optical flow, to build up models for analyzing the motion trajectories in crowded scenesin X.wang *et al.*[18] and D.Kuettel *et al.*[8]. These approaches prove effective in global scene visualization. However, these approaches are computationally complex due to tedious training. In addition, as a result of applicability of some local level visual features, pedestrian’s motion detection is difficult in sparse crowd scenes.

B TRAJECTORY-BASED METHODS

Trajectory-based methods consider a crowd as a collection of pedestrians and analyze the communality between individual pedestrians. Trajectories analysis of crowd scene is presented in [9] and [19]. Choi and Savares [5] reported a hierarchical activity model that identify the relationship between individual trajectories and collective actions. Wang *et al.*[19] proposed a technique for trajectory clustering and semantic region recognition. These techniques analyze the motion patterns of individual pedestrian. However, the computed models are not deployable to different scenes. Morris and Trivedi in [13] proposed a model that captured pedestrians trajectories, action recognition and anomaly detection in crowd scenes. The goal of our work is different. We aim at recognizing and classifying crowd scenes by analyzing the crowd spatio-temporal features of the input scenes.

C. PARTICLE-BASED METHODS

Particle-based also called tracklet-based models are effective in analysis of dense crowd scenarios [12] ,[1] and [16] Saad work in [16] is based on a particle-based approach that is used in analysis of interactions between crowd members and also in measurement of flow complexity. The tracklet-based methods deal with tracklets generated by Kanade Lucas Tomasi (KLT) trackers in [20] and.Zhou *et al* [20]. in proposed a mixture model to learn motion patterns and predict individual behaviors from tracklets. Shao *et al.* in measure the collective crowd motions based on the path similarities on the collective manifold. These methods provide a trade-off between holistic approaches and trajectory-based methods. The limitation is that individual information cannot be fully covered in the scene. Our proposed method studies not only the global information but also the individual local information.

D. SPATIO-TEMPORAL FEATURES MODELS

Spatio-temporal models found in the literature include.C.S.J.Junior *et al.* [6] ,S.Ali *et al* [2] Moreover, model for activity recognition in a video has been reported. In this study, distance metric learning was implemented using physical activity recognition and intensity estimation. In another recent development,the authors proposed a novel multi-view feature selection method via joint local pattern-discrimination and global label-relevance analysis (mPadal). The proposed mPadal employs a new joint local-and-global approaches of human pose. Experimental findings show that mPadal outperforms other baseline methods on publicly available activity recognition dataset. Multi-view models, based on histogram of motion intensities (HMI) and histogram of oriented gradient (HOG) descriptors are used in analysis of human action in video scenes. A novel approach for human action recognition using 3D skeleton joints recovered from RGB-D cameras. Recently, a computer vision algorithm for classification of gait anomalies from kinect is proposed in for neurodegenerative diseases like Parkinson and Hemiplegia. These approaches may not be suitable for real time application because of the computational complexity involved. Another limitation of the approaches is that, they are suitable to few individual scene as such may not be applicable to crowded scenarios. Another closely related work on behavior recognition has been presented recently. They conducted a detailed survey on activity recognition using semantic space features such as pose, pose let, attributes, related object, and scene context. They extensively exploited the aforementioned features to recognize behavior in video data and still images. The literature survey revealed that the studies in the domain of activity recognition and crowd behavior classication rely heavily on BoVF algorithms in training the classification models.

BoVF algorithms A.N.Shuaibu *et al* [17] lack convergence speed and poor accuracy due to long iterative nature, and its performance is greatly affected by number of k-clusters. The local convergence of BoVF algorithms can lead to computational complexity for real-time behavior analytic application. However, sparse representation approach performs better and avoid possibility of being trapped in local minima which is commonly observed with BoVF -[17]. Sparse features algorithms has been proven to effectively work with linear classifiers for simplification of the models and performance improvement of image lesio discrimination. Reviews on more efficient methods on crowd scene classification have been reported. The proposed SCKSVD model overcomes the limitations listed in the aforementioned methods. Innovative model design, appearance and dynamic information can be effectively learned from sparsely coded features. In addition, the proposed model is capable of extracting HOG and histogram optical flow (HOF) from long-range scenes (i.e.200 frames or more) without sampling or compression. The SCKSVD addresses the common limitations of the state of-the-art models such as complexity in training and poor classification accuracy.

PROPOSED WORK

This section describes the proposed method for behavior understanding in crowd scene based on learned spatiotemporal sparse feature representation. Sparse signal representations are becoming increasingly popular and lead to state-of-the-art results in various applications such as face recognition, image demising and imprinting, and image classification. The main reason being the intrinsic sparse nature of image representations when using fixed bases such as discrete cosine transforms or wavelets. In addition, the basis vectors can be learned from the data itself and be constrained to produce a sparse representation. The entire process is presented in the flowchart.

A. DENSE SPATIO-TEMPORAL INTEREST POINTS

This section gives an overview of the feature extraction used, by reading the video input data (v), followed by extraction of interest point patches (i). We extend the concept of Space Time Interest Points (STIP) used, and detect the key interest points using spatio-temporal extension of Harris corner detector. Histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) are extracted as stated *extractHOG*(*patch*) and *extractHOF*(*patch*) respectively. The features are concatenated and train with the sparse model. In this work we carried out scale selection at multiple levels and then extracted features as compared to single scale selection. The multiple-scale selection of spatiotemporal features reduced the computational complexity and present good recognition performance as presented.

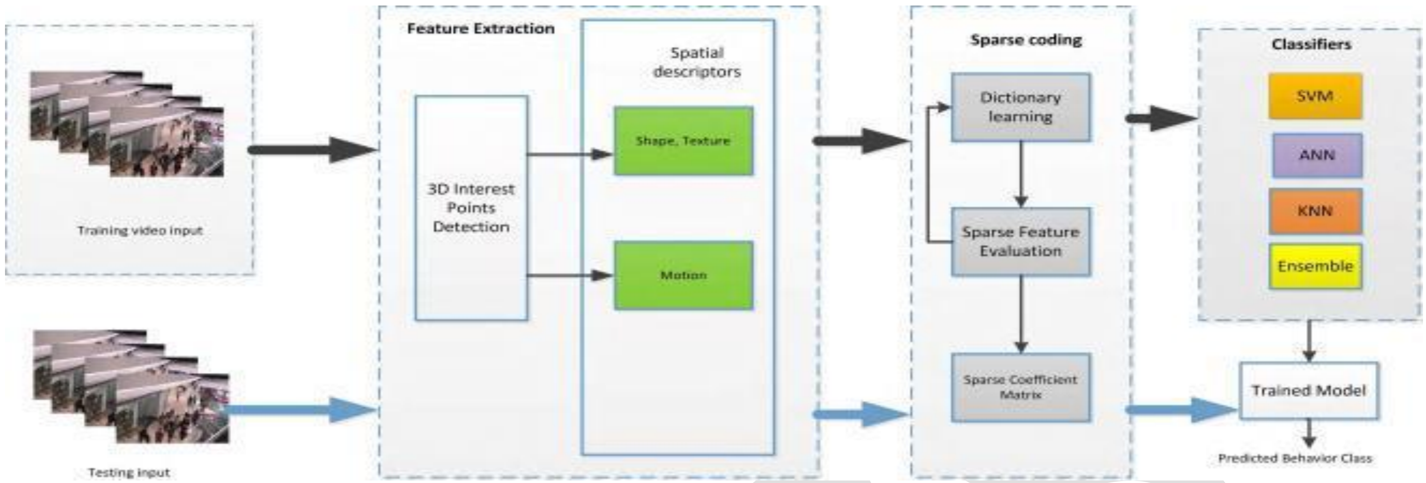

FIGURE 1. Schematic flow pattern of crowd behavior analysis.

The interest points give sign cant information on video data. They provide a compact video representation and tolerance to background clutter, occlusions and scale change. The extracted features are used in training and evaluation of specific behaviors. Appearance and motion features are obtained by computing HOF descriptor and HOG descriptor in the dense grid. It is worth mentioning that this approach does not require tracking. The motives behind the HOGHOF is that local object appearance, shape, and motion information can be effectively computed by a distribution of local intensities. This is implemented by dividing the video sequence into 3D patches 1x, 1y, and 1of spatial regions (cells). Each volume is further divided into *n* tx, *n* y, and *n*. grid of cuboids, then coarse HOG and HOF are extracted for each sub-volume segment. Then normalized histograms are concatenated into HOG, HOF and HOGHOF feature descriptor vectors and are closely alike to the well-known scale invariant feature transform descriptor. T The STIP is based on Harris 3D detector and dense detector. In the spatial domain \`*s*' an image \`*f* ' can be model in space-time domain by its corresponding linear scale representation. The convolution of *f* with Gaussian kernel of scale parameters \`_21' and \`_21 is given in Eq. (1).*L*_*x*; *y*; *t* I _2; _2_D *g*_*x*; *y*; *t* I _2; _2__ *f*.*x*; *y*; *t*(1)Where \`_' denote the convolution sign and *L* is the Gaussian operator obtained at local scale. The spacetime second moment is computed using the relation:_.:I _; _/D *g*.O*L*(:I *s*_; *s*_/) __O*L*.:I _; _/where O*L* is the space-time derivative. Interest points are detected based on local maxima *H* corresponding to positive values *H* > 0 as given in:*H* D *det*/ *ktrace*3*T*._/.*H* D _1C _2 *k*._C _2/2.

Algorithm for behaviour recognition from input video.

```
Read video data.
n ← number of videos in the dataset.
for (v:= 1 to n,v++) do
    [I]<- detect interest Point(v)
    for each i€[I] do
        patch ← extract Patch(i)
        [hog] ← extract HOG(Patch)
        [hof] ← extract HOF(Patch)
        Desct(v,i) ← (hog+hof)
    end for
end for
for ← 1,2,_ _ _ ,p do
    $D_k$ ← [Desct]y
    $v_i$ ← TrainedSparseDictionary
    $v_i$ ← histogram($v_i$)
    $v_i$ ← normalized($v_i$)
end for
```

```
18: model ⟵ TrainClassifier(V)
19: Evaluate the trained Model
```

B. SPARSE MODELING REPRESENTATION

The aim of sparse coding is to locate an effective method of images pattern illustration by fusion of multiple features selected from a dictionary. Given a sparse dictionary matrix *D* D *d* that contains *K* atoms as column vectors *d*, the sparse coding problem of extracted HOG-HOF video descriptor *T* (HOG,HOF,HOGHOF) can be stated sanding the sparsest vector *y* such that *T* is approximately as *Dy*. A signal *T* can be represented by the linear combination of atoms;1; *d*2; : : : *dT nn*X1*yDn*_12_.1C_/D *Dy* (5) Where *n* is the total number of sparse dictionaries, *y* is the sparse code representation (mostly zeros entries) of T over D. The procedure to solve *y* is called sparse decomposition. The system of linear Eq. (5) is ill-posed and has no unique solution; however,, if *y* is sparse or approximately sparse, it can be uniquely determined by solving the following Optimization problem:*y* D*minyk*y*k0subject to*k*T Dy*k Where k_k represents the *lp*2 *norm* operator and _ is the reconstruction error of the signal *T* using the dictionary *D* and the sparseness code vector *y*. The combinatorial problem associated with Eq. (6) is NP-hard [38], so it is impossible to solve this problem by analyzing all possible sparse subsets. There are two types of methods to solve this problem. The first is the greedy algorithms [50], [51]; the second method relaxes the highly discontinuous *l norm*, replacing it by a continuous or even smooth approximation. Typically, when *l*00 *norm* is used, the problem becomes D*miny*k*T Dy*k2 *subject to*k*y*k0 where _ indicates the sparsity level. This convex minimization problem can be cast as a least squares problem with a penalty.

C. VIDEO CLASSIFICATIONS WITH SPARSE CODED FEATURES

As stated, sparse coding process follows immediately after the extraction of 3D spatio-temporal interest points and HOG, HOF, HOGHOF descriptors. The sparse representation is to learn a dictionary that captures vital and discriminative features of the video scenes. SCKSVD is employed for dictionary learning process.

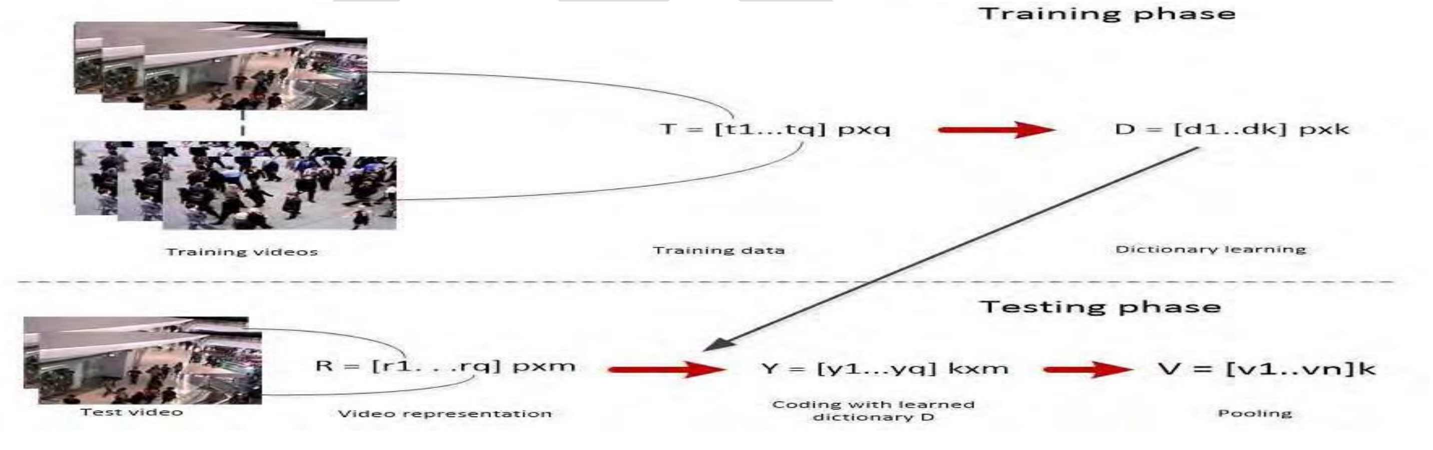

FIGURE 2. Pipeline for feature extraction and sparse coding technique.

The dictionary is learn from low-level features obtained from the training video data. Literatures have shown that using sparse representation method to video descriptors can reveal high-level of discriminated features. In this work, we resize the input videos to size 120 _ 160 resolution. The following low-level feature descriptors are extracted from the video scene:

HOG: Histogram of Oriented Gradient (HOG) are extracted over a regular grid on each frame of the video. In our evaluation, each video sequence is split into 3 _ 3 _ 2 grid cells with parameters *n* D 3, *n y* D 3, and *n* D 2 and 4-bin, a descriptor dimension 72 is extracted. The choice of the grid scale is motivated by the reduced computational complexity.

HOF: Histogram of Optical Flow (HOF) are extracted over regular grid as computed for HOG, with *n x* D 3, *n* D 3, and *n* D 2 and 5-bin. For every frame, we extract a feature vector of dimension 90.*t*

HOG-HOF: This is the overall concatenation of the aforementioned extracted features. For every frame sequence, we have 72 C 90 D 162 extracted feature vector.

The vector y comprises the representation coefficients of the feature descriptor signal *T* w.r.t. the dictionary *D*. The main problem encountered in practical applications is the choice of dictionary size *D*. One can use a pre-determined and fixed dictionary size as commonly used in wavelets, curve lets, and steerable _liters transforms. It is imperative to use varying dictionary size for a given set of training data in order to justify the number of atoms that give best performance measures. For each extracted feature type, they are arranged in the column of a matrix *T* D [*t*], where each *ti*1*y*; : : : *t*is a feature vector and q is the total number of the local patches in the training and testing videos. A dictionary *D* D [*d*] is learnt with K-SVD algorithms [33], where *k* is the size of the visual dictionary. A test video can be coded with the learnt dictionary *D* by creating a feature matrix *R* D [*r*11; : : : *d*; ::*rqk*] with a set of low-level features from the video descriptors.

EXPERIMENTS AND RESULTS

In this section describe the evaluations of sparse coding technique with STIPs 3D feature descriptors, and presents the results for combined descriptors:HOG-SCKSVD, HOFSCKSVD, and HOGHOF-SCKSVD for video classifications. In particular, it show the discrimination between crowd walking, crowd crossing and crowd merging, intervened escalator, smooth escalator etc. using sparse coding and bag-of-visual-features representation.

The dataset is split based on Leave-One-Out (LOO) cross validation scheme. It is apply in order to optimize the sparse coefficients matrix. The analysis is repeated over all the training set in each case by leaving one sample out for testing. The training group is used to build SVM classifier and the testing group is used to calculate the performance of the classifier. The performance evaluation is carried out in terms of classification accuracy, sensitivity and specificity accuracy, recall and precision. They are regarded as more appropriate than other statistical metrics in performance evaluation on common platform.

Performance evaluation is computed using confusion matrix. Confusion matrix comprises of actual and predicted classifications with the following components: True Positive(*TP*), False Positive (*FP*), False Negative (*FN*), and True Negative (*TN*) as presented in contingency. Accuracy is given by the total number of correct prediction. The True Positive Rate (*TPR*) also called Sensitivity or the Recall is the proportion of actual positive classes that were correctly identified. The Precision also called Positive Predictive Value (*PPV*) is the proportion of predicted positive cases that were correct. The True Negative Rate (*TNR*) also known as Specificity is referred to the proportion of negatives cases that were classified correctly.

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + FP + TN + FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}}$$

The performance analysis of the sparse coding is presented for STIPs feature descriptors. It carry out different experiments, and in each case LOO cross validation with linear SVM is used. The dataset is randomly divided into (*n* - 1) clusters, and the visual dictionaries are learn with (*n* - 1) clusters in each case by testing on one set. The analysis is repeated for total number of videos *n* times using both training and testing videos and the mean classification result is recorded by taking the average. As stated in pedestrian sample frames , then eight classes of pedestrian attributes, compute the classification result for each behaviour classes.

CONCLUSION

This paper proposed an approach that combines spatio-temporal feature descriptors of crowd scenes with sparse coding model. The sparse signal representation techniques was used to detect robust features of crowd scenes and produce discriminant descriptors for classification between videos containing different behaviour classes. The proposed technique achieves higher performance than other feature extraction schemes such as bag-of-visual features approaches and deep learning approaches. Experimental results was tested on three publicly available datasets (CUHK, UCF and Crowd-UTP). The statistical paired test comparing the average accuracies of both the hand-designed and deep learning-based approaches shows that there was no difference in the result obtained from the techniques on CUHK, UCF and Crowd-UTP datasets. It obtained excellent and comparable classification results on all the datasets. In particular, using a dictionary of size 1000 and a linear SVM with sparse coded features, the method achieves better results in terms of accuracy, precision and recall for all the behaviour classes. The model proposed in this study will be of useful help in the area of crowd surveillance and monitoring. The proposed model is built based on eight (8) attributes of crowd videos, which restrict prediction of behaviours outside the scope of the datasets. It intend to extend to 3D adaptive features learning approach with large attributes dataset to accommodate broader prediction of behaviour classes.

REFERENCES:

[1] S. Ali and M. Shah, ``Floor fields for tracking in high density crowd scenes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 1-14.

[2] S. Ali and M. Shah, ``A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2007, pp. 1-6.

[3] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, ``Effective active skeleton representation for low latency human action recognition," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 141-154,Feb. 2016.

[4] A. Chan and N. Vasconcelos, ``Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909-926, May 2008.

[5] W. Choi and S. Savarese, ``A uni_ed framework for multi-target tracking and collective activity recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 1-6.

[6] J. C. S. J. Junior, S. R. Musse, and C. R. Jung, ``Crowd analysis using computer vision techniques," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 66-77, Sep. 2010.

[7] V. J. Kok, M. K. Lim, and C. S. Chan, ``Crowd behavior analysis: A review where physics meets biology," *Neurocomputing*, vol. 177, pp. 342-362, Sep. 2015.

[8] D. Kuettel, M. Breitenstein, L. van Gool, and V. Ferrari, ``What's going on? discovering spatio-temporal dependencies in dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2010, pp. 1-6.

[9] R. Li, R. Chellappa, and S. K. Zhou, ``Recognizing interactive group activities using temporal interaction matrices and their Riemannian statistics," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 305-328, 2013.

[10] W. Liu, R. Lau, and D. Manocha, ``Robust individual and holistic features for crowd scene classi_cation," *J. Pattern Recognit.*, vol. 58, pp. 110-120, Apr. 2016.

[11] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, ``Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2010, pp. 1975-1981.

[12] R. Mehran, A. Oyama, and M. Shah, ``Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2009, pp. 935-942.

[13] B. Morris and M. Trivedi, ``Learning and classi_cation of trajectories in dynamic scenes: A general framework for live video analysis," in *Proc. IEEE Adv. Video Signal-Based Surveill.*, Apr. 2008, pp. 154-161.

[14] Y. Nam, S. Rho, and J. H. Park, ``Intelligent video surveillance system:3-tier context-aware surveillance system with metadata," *Multimedia ToolsAppl.*, vol. 57, no. 2, pp. 315-334, 2012.

[15] B. Pogorelc, and M. Gams, ``Detecting gait-related health problems of the elderly using multidimensional dynamic time warping approach with semantic attributes," *Multimedia Tools Appl.*, vol. 66, no. 1, pp. 95-114, 2013.

[16] A. Saad, ``Measuring _ow complexity in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 1097-1104.

[17] A. N. Shuaibu, A. S. Malik, and I. Faye, ``Behavior representation in visual crowd scenes using space-time features," in *Proc. 6th Int. Conf. Intell. Adv. Syst. (ICIAS)*, Kuala Lumpur, Malaysia, Sep. 2016, pp. 1-6.

[18] X. Wang, K. Ma, G. Ng, and W. Grimson, ``Trajectory analysis and semantic region modelling using nonparametric hierarchical Bayesian models," *Int. J. Comput. Vis.*, vol. 95, no. 3, pp. 287-312, 2011.

[19] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, ``Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2009, p. 12411.

[20] B. Zhou, X. Wang, and X. Tang, ``Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 1-8.