1-1-2014

Crowd behavior recognition using dense trajectories

Muhammad Rizwan KHOKHER
University of Wollongong, mrk840@uowmail.edu.au

Abdesselam Bouzerdoum
University of Wollongong, bouzer@uow.edu.au

Son Lam Phung
University of Wollongong, phung@uow.edu.au

Crowd behavior recognition using dense trajectories

Abstract

This article presents a new method for crowd behavior recognition, using dynamic features extracted from dense trajectories. The histogram of oriented gradient and motion boundary histogram descriptors are computed at dense points along motion trajectories, and tracked using median filtering and displacement information obtained from a dense optical flow field. Then a global representation of the scene is obtained using a bag-of-words model of the extracted features. The locality-constrained linear encoding with sum pooling and L2 plus power normalization are employed in the bag-of-words model. Finally, a support vector machine classifier is trained to recognize the crowd behavior in a short video sequence. The proposed method is tested on two benchmark datasets, and its performance is compared with those of some existing methods. Experimental results show that the proposed approach can achieve a classification rate of 93.8% on PETS2009 S3 and area under the curve score of 0.985 on UMN datasets respectively.

Keywords

behavior, dense, recognition, trajectories, crowd

Disciplines

Engineering | Science and Technology Studies

Publication Details

M. Rizwan. Khokher, A. Bouzerdoum & S. Lam. Phung, "Crowd behavior recognition using dense trajectories," in Digital Image Computing: Techniques and Applications (DICTA), 2014 International Conference on, 2014, pp. 1-7.

Crowd Behavior Recognition using Dense Trajectories

Muhammad Rizwan Khokher, Abdesselam Bouzerdoum, Son Lam Phung
School of Electrical, Computer and Telecommunication Engineering
University of Wollongong, NSW, 2522, Australia
mrk840@uowmail.edu.au, a.bouzerdoum@uow.edu.au, phung@uow.edu.au

***Abstract*—This article presents a new method for crowd behavior recognition, using dynamic features extracted from dense trajectories. The histogram of oriented gradient and motion boundary histogram descriptors are computed at dense points along motion trajectories, and tracked using median filtering and displacement information obtained from a dense optical flow field. Then a global representation of the scene is obtained using a bag-of-words model of the extracted features. The locality-constrained linear encoding with sum pooling and L2 plus power normalization are employed in the bag-of-words model. Finally, a support vector machine classifier is trained to recognize the crowd behavior in a short video sequence. The proposed method is tested on two benchmark datasets, and its performance is compared with those of some existing methods. Experimental results show that the proposed approach can achieve a classification rate of 93.8% on PETS2009 S3 and area under the curve score of 0.985 on UMN datasets respectively.**

***Keywords*—Crowd behavior recognition; dense trajectories; motion boundary histogram; bag-of-words; support vector machine**

I. Introduction

Crowd behavior recognition is one of the most challenging tasks and is an open research area. The goal is to detect and recognize different crowd behaviors using temporal information. Crowd behavior understanding faces many challenges including group-level relationships, complex interactions, emergent behaviors and self-organizing activities [1]. To tackle these challenges, many techniques have been developed [2–14] which lead to various useful applications like crowd management, automatic surveillance, public space design and virtual and intelligent environments.

There may be hundreds of objects in a crowd scene. Therefore, traditional methods like segmentation, object detection and tracking face challenges due to presence of similar appearances, small objects and inter-object occlusion. Although, some of the methods work for low density crowd scenes, they are likely to fail for high density crowd scenes. To overcome these problems, researchers are shifting focus towards developing the methods that model the overall dynamics of a crowd by considering the crowd as a single entity. In general, methods for crowd behavior analysis can be divided into two categories: object-based and holistic methods. Object-based methods deal with a crowd as a collection of objects/individuals [2–4]. In holistic methods, a crowd is considered as a single entity and no segmentation or detection of individual objects is required [5–14]. Since the proposed method falls in the latter category, in the remainder of this section different holistic methods for crowd behavior analysis are discussed.

Andrade *et al*. used optical flow and hidden Markov models (HMMs) to categorize the behavior of a crowd [5]. An unsupervised algorithm was used for spectral clustering and feature extraction, and a model for background subtraction was developed for optical flow computation. Another optical flow based method was proposed by Ali and Shah [6]. The crowd instabilities were detected based on Lagrangian particle dynamics by segmenting the crowd. The crowd movements were captured and used to generate a velocity field, and the particle movement on the velocity field helped in constructing a flow. These methods based on optical flow worked well but they faced challenges due to complex and incoherent motion patterns in dense crowd scenes.

For dense crowd scenes, Kratz and Nishino presented an approach for anomaly detection using spatio-temporal information from the scene [7]. A set of spatio-temporal cuboids was extracted from the video sequence using 3-D multivariate Gaussian distribution. Then HMMs were used to build the connection between motion patterns. However, a cuboid based strategy is likely to result in a loss of information because different cuboids may separate coherently meaningful features.

In [8], a method was proposed for abnormal crowd behavior detection based on social force model originally introduced in [9]. The crowd was modeled using socio-psychological studies. The particle motion is used to estimate the social forces after a set of particles are overlaid onto the image. These social forces were used to classify each frame of the scene as normal or abnormal. In this method, the use of socio-psychological studies was an interesting aspect but it faced difficulties in localizing anomalies with the bag-of-words model.

In another approach, Mehran *et al*. presented a streakline technique to model the flow in crowded scenes [10]. Some important features were computed, such as potential functions and crowd flow, using Helmholtz decomposition theorem. This gave a better representation of the flow in comparison with common flow representations, by more accurately recognizing the temporal and spatial changes in the scene. The use of

potential function was also presented in [11] for abnormal crowd detection. An interaction energy potential function was designed to model the current behavior states of the subjects and their actions were represented by their velocities. The use of spatio-temporal interest points provided good modeling of the group interactions.

Cong *et al*. proposed a method for abnormal event detection in crowded scenes based on sparse reconstruction over the normal basis [12]. They measured the normal behavior of a test sample by proposing the sparse reconstruction cost (SRC) over the normal dictionary. During sparse reconstruction, a prior weight for each basis was introduced to make the proposed SRC more robust, and a novel dictionary selection method was designed to condense the over completed normal basis into a compact dictionary. Both local and global abnormal event detection were acheived by designing different types of spatio-temporal basis.

Li *et al*. designed a joint detector of spatial and temporal anomalies to detect and localize anomalous behaviors in crowded scenes [13]. A set of mixture of dynamic texture models was used to account for both appearance and dynamics for the proposed detector. Spatial and temporal saliency scores were obtained through different models and used as potentials of a conditional random field for anomaly detection.

Zhang *et al.* presented a bag of trajectory graphs (BoTG) method for dense crowd event recognition [14]. A group-level representation was designed to overcome the loss of information (i.e., variability of motion and crowd structure) in previous particle flow approaches. This method achieved a very good performance in comparison with previous methods for crowd behavior analysis.

So far, all the holistic methods presented for crowd behavior recognition face challenges due to complex motion patterns in dense crowded scenes. For a better representation, we propose to use dense trajectories [15, 16] and motion boundary histogram (MBH) [17] descriptor for the first time for crowd behavior recognition problem. The motivation behind using dense trajectories is that, based on derivatives of optical flow, the MBH computed along dense trajectories helps suppress the irrelevant motion patterns. This is different from other techniques where particle flow is used directly which is likely to affect to accuracy in the presence of complex motion patterns. Usually we get too many trajectories [15, 16], the feature representation can be further improved by incorporating only those trajectories that are on the motion boundary [18]. The feature representation using dense trajectories along with the bag-of-words (BoW) and support vector machine (SVM), build a unique model for crowd behavior recognition. For the BoW model, different coding and pooling methods are investigated to get better results.

The remainder of the paper is organized as follows. The proposed method is introduced in Section II including local feature extraction, global feature representation and classification details. In Section III, detailed experiments and analysis are presented along with the datasets used. Section IV concludes the paper.

II. Proposed Method

We propose a new method for crowd behavior recognition based on dense trajectories. In order to extract the information of dynamic nature, dense feature points are sampled from each frame using multiple spatial scales. These points are then tracked using median filtering and the information provided by displacement from a dense optical flow field. The tracked points in subsequent frames yield a trajectory. Fig. 1 shows the dense trajectories computed for different crowd scenes from the UMN dataset [19]. Different feature descriptors are computed along the trajectories obtained. These feature descriptors include, histogram of oriented gradients (HOG) and MBH. The next step is to generate a global feature representation from the extracted feature descriptors. For this purpose, a codebook is generated using k-means. Given the codebook, the feature descriptors are encoded through locality-constrained linear encoding (LLC). After feature encoding, sum pooling is used along with L_2 norm and power normalization to get a final global representation. In the last stage, a support vector machine (SVM) classifier is used to classify crowd behavior.

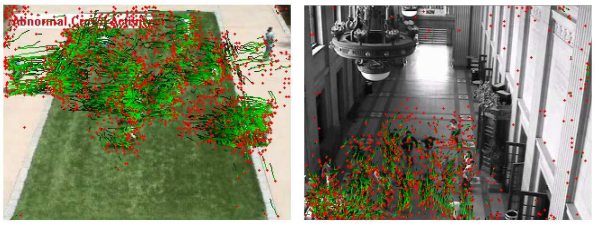

Fig. 1: Dense trajectories computed for different crowd scenes from the UMN dataset [19]. Red marks are the end points of the trajectories.

The next three subsections present in more detail the feature extraction, global feature representation and stages, respectively.

A. Feature Extraction

Previous studies have shown that feature trajectories can provide an efficient representation of a video sequence. Usually, these trajectories are computed by matching SIFT descriptors between frames. Wang *et al*. implemented this concept along with dense sampling and called it dense trajectories [15, 16]. We extract the trajectory features from videos using these dense trajectories and calculate different descriptors (i.e., HOG and MBH) along those trajectories.

In order to extract dense trajectories, multiple spatial scales are used. Fig. 2(a) shows the dense sampling process where a grid separated by W pixels is used to densely sample the feature points which are separately tracked in each spatial scale. At a frame I_t, point $P_t = (x_t, y_t)$ is

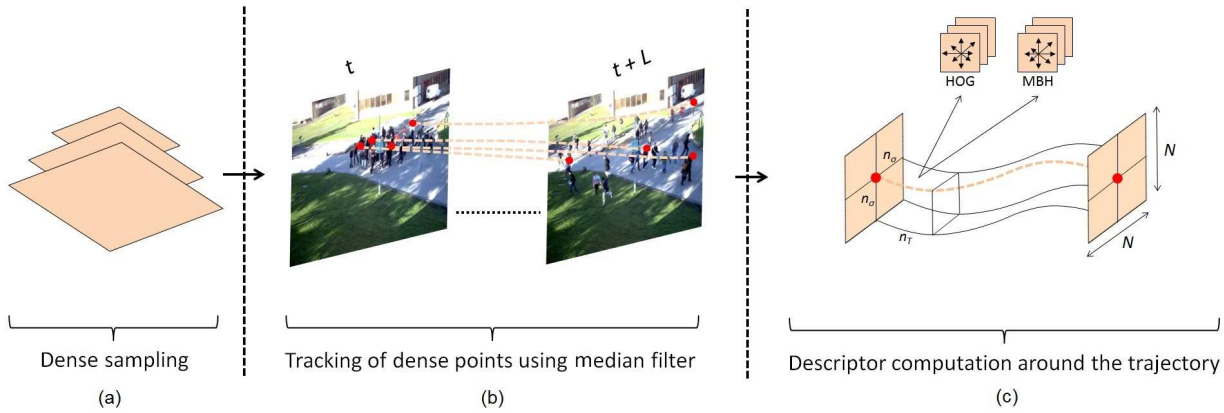

Fig. 2: Illustration of the dense trajectories method adapted from [15, 16]. (a) A grid is used to densely sample the features points for each spatial scale. (b) A median filter and dense optical flow field is used to track the points in each spatial scale. (c) Descriptors like HOG and MBH are computed along the dense trajectories within a volume of $N \times N \times L$ which is subdivided into $n_\sigma \times n_\sigma \times n_T$.

tracked in the following frame I_{t+1}, and its tracked position is smoothed by applying a median filter on a dense optical flow field $w_t = (u_t, v_t)$:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t)|_{(x_t, y_t)}, \qquad (1)$$

where M represents a median filter kernel. After the computation of dense optical flow field, points are tracked densely without additional cost. The algorithm by Farneback [21] is used to extract dense optical flow as it embeds a translational motion model between two consecutive frames. Irregular and fast motion patterns can be easily tracked because of the smoothness constraints of the dense optical flow field.

A trajectory $(P_t, P_{t+1}, P_{t+2}, ...)$ is formed by concatenating tracked points in subsequent frames. Since trajectories can drift from their point of initialization, the length of a trajectory is restricted to L frames as shown in Fig. 2(b). The trajectory is removed from the tracking process if it exceeds the length L. The availability of a track on the dense grid in a frame is verified so that a dense coverage can be ensured. If there is no tracked point present in a $W \times W$ neighborhood then this feature point is added to tracking process after it is sampled. The points cannot be tracked in a structureless homogenous image area. The criteria from [22] is used here for tracking of sample points in structureless areas. Given the auto-correlation matrix of a sampled feature point, if the smaller eigenvalue is below a threshold, the point is excluded from the tracking process. The static trajectories and trajectories with large displacements are also removed in a pre-processing stage. If the displacement vector between two consecutive frames is larger than 70% of the overall trajectory displacement, it is considered as a large displacement.

The local motion pattern is described by the shape of a trajectory. Given a trajectory of length L, the shape is described by a vector $S = (\Delta P_t, ..., \Delta P_{t+L-1})$, where $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$ is the displacement vector. The vector S is normalized to get a trajectory descriptor given in Eq. (2),

$$S' = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} |\Delta P_j|}. \qquad (2)$$

Descriptors like HOG and MBH are computed along the dense trajectories within a space-time volume which leverages the motion information. The space-time volume has dimensions $N \times N \times L$ which is further divided into a grid of size $n_\sigma \times n_\sigma \times n_T$. This embeds the structure information as shown in Fig. 2(c).

The HOG descriptor concentrates on the static appearance information while MBH extracts the dynamic information. The orientations are quantized into β bins for HOG. L_2 norm is used for the normalization of the descriptor. The MBH descriptor computes the derivatives for vertical and horizontal components of optical flow field which encodes the relative motion between pixels [17]. The optical flow field is separated by MBH into its x and y components for which spatial derivatives are computed. The orientation information is quantized into β' bins for each component which is then normalized by L_2 norm.

Usually, there are many points to be tracked [16]. However, valid trajectories can be obtained from only few points residing on the motion boundary [18]. This concept is partly implied by the MBH descriptor. In [18], two successive frames are used to sample points which is different from [16]. Once the gradient magnitude of optical flow field is calculated, a thresholding operation is applied

and the original dense trajectory sampled points are refined through a generalized mask obtained from Otsu's algorithm [23]. The static regions with no motion foreground are deleted. For the remaining patches, the central points are refined using average location of foreground. The number of trajectories are significantly reduced in comparison with [16]. Not all the points of trajectories are forced to be on motion boundary to avoid inaccurate tracking; see [18] for further details.

B. Global Feature Representation

The bag-of-words technique has been used successfully to represent video sequences. The BoW was originally used for natural language processing. Then it has been used extensively in video and image classification tasks. Here, we use the BoW model to represent the video sequences from their feature descriptors obtained using dense trajectories.

1) Codebook Generation: To begin with, a codebook is generated using a set of descriptors. The feature space is partitioned into different regions called visual words and a codebook is generated using k-means clustering [24]. A set of local feature descriptors $\{\mathbf{x}_1, ..., \mathbf{x}_N\} \in R^D$, is partitioned into K clusters $\mathbf{L} = \{\mathbf{d}_1, ..., \mathbf{d}_K\}$, where $\mathbf{d}_K \in R^D$ is the centroid or prototype of the k-th cluster. For each feature descriptor $\mathbf{x}_i$, a corresponding binary indicator $r_{ik} \in \{0, 1\}$ is assigned which means $r_{ik} = 1$ if feature descriptor $\mathbf{x}_i$ is linked with cluster k and $r_{ij} = 0$ for $j \neq k$. An objective function can be defined as:

$$\min_{\mathbf{L}} \chi(\{r_{i,k}, d_k\}) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{i,k} \left\| \mathbf{x}_i - \mathbf{d}_k \right\|_2^2, \tag{3}$$

where $\|\cdot\|_2$ represents the L_2 norm. The objective function χ is optimized iteratively. The details are available in [24].

2) Feature Encoding: After obtaining the codebook $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_K]$ with K visual words, a code vector $\mathbf{c}_i \in \mathbf{C}$ is calculated for the local feature descriptors $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ obtained from the video sequences. The dimension of the code vector $\mathbf{c}_i$ is the same as that of codebook vectors $\mathbf{d}_k$.

There are different encoding techniques including vector quantization, soft encoding and locality-constrained linear encoding (LLC) [25] for the classification of images. Based on the performance in our experiments, we use LLC method to encode our features obtained for the videos using the codebook. The LLC method is different from other encoding methods because it enforces locality over sparsity. A smaller coefficient is obtained for the basis vectors that are far from the feature descriptor $\mathbf{x}_i$. The following optimization problem is solved to get the coding coefficients:

$$\begin{aligned} \min_{\mathbf{C}} \sum_{i=1}^{N} \left\| \mathbf{x}_i - \mathbf{D}\mathbf{c}_i \right\|^2 + \lambda \left\| \mathbf{s}_i \odot \mathbf{c}_i \right\|^2, \\ \text{subject to } \mathbf{1}^T \mathbf{c}_i = 1, \end{aligned} \tag{4}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{s}_i$ represents the locality adaptor. The weights for each basis vector are obtained using the following locality adaptor:

$$\mathbf{s}_i = \exp\left(\frac{\text{dist}(\mathbf{x}_i, \mathbf{D})}{\sigma}\right),$$

where $\text{dist}(\mathbf{x}_i, \mathbf{D}) = [\text{dist}(\mathbf{x}_i, \mathbf{d}_1), ..., \text{dist}(\mathbf{x}_i, \mathbf{d}_K)]^T$. The speed of weight decay is adjusted by σ for the locality adaptor. For the LLC code, the shift invariant requirements are followed by the constraint $\mathbf{1}^T \mathbf{c}_i = 1$. To improve the computational efficiency, an approximation can be implemented. The second term in Eq. (4) can be ignored and k nearest basis vectors of $\mathbf{x}_i$ can be directly selected which minimizes the first term after solving a smaller linear system. This provides the coding coefficients for k basis vectors and other coefficients are set to zero.

3) Pooling and Normalization: Once we get the coding coefficients of all the local feature descriptors in a video, a holistic representation $\widetilde{\mathbf{x}}$ is obtained through a pooling process. Here we use sum pooling strategy [26]. In sum pooling, the k-th component of $\widetilde{\mathbf{x}}$ is given by:

$$\widetilde{x}_k = \sum_{i=1}^{N} \mathbf{c}_{i,k}.$$

The pooled feature $\widetilde{\mathbf{x}}$ is then normalized using a combination of L_2 and power normalization. Using L_2 normalization [27], the feature $\widetilde{\mathbf{x}}$ is divided by its L_2 norm:

$$\mathbf{x}' = \widetilde{\mathbf{x}} / \sqrt{\left(\sum_{k=1}^{K} \widetilde{x}_k^2 \right)}.$$

The power normalization method [27] applies the following function in each dimension:

$$f(\widetilde{x}_k) = \text{sign}(\widetilde{x}_k) |\widetilde{x}_k|^{\alpha},$$

where the normalization parameter α ranges from 0 to 1.

C. Classification

After getting the global feature representation of the feature descriptors obtained from dense trajectories, SVM is used to classify the video sequences [28]. Given a set of training instance-label pairs $(\mathbf{x}'_i, \ell_i)$, $i = 1, ..., l$, where $\mathbf{x}'_i$ is a global feature vector for a video sequence and label $\ell_i \in \{1, -1\}$. Classifier training is performed by solving the following optimization problem:

$$\min_{\mathbf{w}} \left\{ \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{l} \max(1 - \ell_i \mathbf{w}^T \mathbf{x}'_i, 0)^2 \right\}, \tag{5}$$

where C is the regularization loss trade-off parameter and is set after the cross validation on training dataset. Here, $\mathbf{w}$ is the normal vector to the hyper-plane separating the samples of two classes. We use one-vs-the-rest approach for multi-class classification.

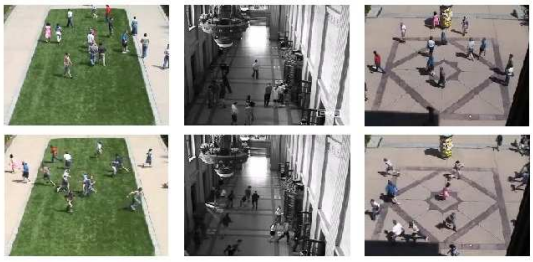

Fig. 3: Sample frames of three different indoor and outdoor scenarios from the UMN dataset [19]. Normal crowd behavior (top) and abnormal crowd behavior (bottom).

III. Results and Analysis

A. Datasets

The proposed method is tested on two benchmark datasets: UMN and PETS2009 S3. The unusual crowd activity dataset from university of Minnesota (UMN) [19], contains 11 different videos of crowd scenes in 3 different indoor and outdoor scenarios. The video scenes start from a normal crowd behavior and end with an abnormal (panic) behavior. Fig. 3 shows sample frames of the crowd scenes from the UMN dataset. The PETS2009 S3 dataset [20] contains video sequences of different crowd events including crowd walking, running, local dispersion, evacuation and subgroup formation. There are videos from four different camera views. The videos are manually segmented into 65 clips and labeled with the crowd events, walking, running, etc. The sample frames of different crowd events from PETS2009 S3 dataset are shown in Fig. 4.

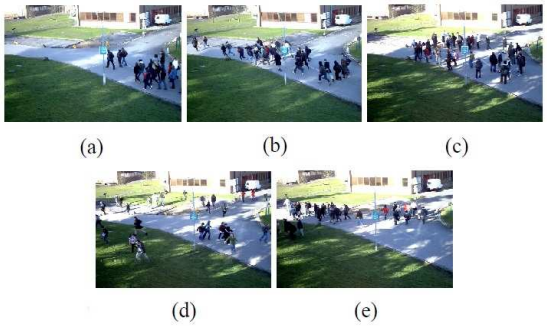

Fig. 4: Sample frames of different crowd events from the PETS2009 S3 dataset [20]. (a) Walking, (b) running, (c) local dispersion, (d) evacuation and (e) group formation.

B. Parameter Selection

During feature extraction using dense trajectories, the same parameter values are set as in [15, 16, 18]. The dense sampling step size is set to $W = 5$ pixels which is dense enough to yeild good results. In total, 8 spatial scales are used which increase by factor $1/\sqrt{2}$. The median filter kernel M is 3×3 pixels for the tracking of points effects. The length of the trajectories is set to $L = 15$ frames to prevent trajectories from drifting from their point of initialization. A trajectory length $L = 15$ frames gives a 30 dimensional trajectory descriptor S'. In order to exclude a point from the tracking process for homogenous structureless image areas, the threshold τ on the eigenvalues for each frame I is set as follows:

$$\tau = 0.001 \times \max_{i \in I} \min(\lambda_i^1, \lambda_i^2),$$

where $(\lambda_i^1, \lambda_i^2)$ are the eigenvalues of the point i.

In order to compute different descriptors (i.e., HOG and MBH) along dense trajectories, the parameter values for volume $N \times N \times L$ and spatio-temporal grid $n_\sigma \times n_\sigma \times n_\tau$ are set to $N = 32$, $n_\sigma = 2$ and $n_\tau = 3$. The final dimension for HOG descriptor is 96 using $\beta = 8$ bin quantization for orientations. The MBH descriptor separates the optical flow field into its x and y components. A histogram with $\beta' = 8$ bins is computed for each component which provides two intermediate descriptors MBHx and MBHy. The dimension of both MBHx and MBHy descriptors is 96. The final MBH descriptor is obtained by merely concatenating of MBHx and MBHy.

For the BoW, the codebook is generated using k-means for a dictionary size of 100 for the UMN dataset [19] and 200 for the PETS2009 S3 dataset [20]. To reduce the complexity during codebook learning, a subset of features is randomly selected from each training video sequence, containing 25 and 50 descriptors for the UMN and PETS2009 S3 datasets accordingly. For feature encoding using LLC the default parameters are used as in [25] (i.e., $\lambda = 1 \times 10^{-4}$). For power normalization the parameter α is set to 0.5.

A one-vs-the-rest linear SVM is used for training and testing of video sequences given their global feature vectors. The value of parameter C in Eq. (5) is set after the cross validation on training dataset. We use LIBLINEAR toolbox [29] to implement the linear SVM.

C. Abnormal Behavior Recognition

To test the performance of the proposed method for abnormal crowd behavior recognition, experiments are conducted on the UMN dataset. The videos from the dataset are segmented into clips of 15 frames each and labeled as normal or abnormal behavior event. In order to extract different features, we calculate descriptors including spatial (HOG descriptor computed along the dense trajectories), temporal (MBH descriptor computed along the dense trajectories) and spatial+temporal (combination of HOG and MBH via concatenation of visual words).

In order to get a better global feature representation from feature descriptors, different feature encoding methods were tested including vector quantization (VQ) [30], soft-assignment encoding (SA) [31] and LLC. For pooling process, sum and max pooling [32] methods were tested.

Table I presents the area under the ROC curve (AUC) scores of abnormal crowd behavior recognition on the UMN dataset, using VQ, SA and LLC encoding with sum and max pooling. As per the results in Table I, LLC with sum pooling is a better choice here.

TABLE I: AUC scores on UMN dataset using different encoding (VQ, SA and LLC) and pooling (sum and max) methods.

UMN dataset	**VQ**		**SA**		**LLC**	
	Max	**Sum**	**Max**	**Sum**	**Max**	**Sum**
Spatial	0.91	0.925	0.82	0.75	0.91	0.925
Temporal	0.925	0.94	0.785	0.725	0.925	0.94
Spatial+Temporal	0.96	0.975	0.91	0.88	0.96	**0.985**

Fig. 5 shows the ROC curve of the abnormal behavior recognition on the UMN dataset for the proposed method using spatial+temporal descriptor.

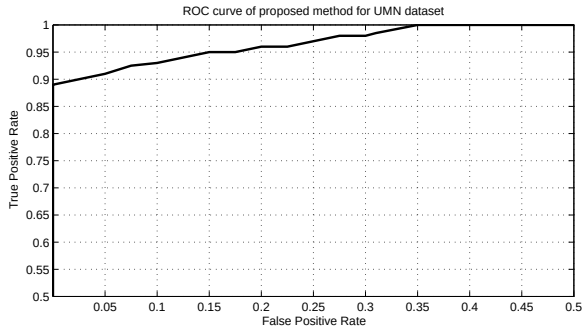

Fig. 5: ROC curve of abnormal behavior recognition for the proposed method on the UMN dataset.

The performance of the proposed method is compared with state-of-the-art methods including BoTG [14], interaction energy potentials (IEP) [11], sparse reconstruction cost (SRC) [12], social force (SF) [8], streakline potential (SP) [10] and optical flow (OF). Table II presents the comparison of the proposed method with the previous methods through the AUC scores. From Table II, we can see that the proposed method outperforms most of the previous methods and has comparable performance with BoTG and IEP methods.

TABLE II: Comparison of the proposed method with other high-level methods through AUC scores on UMN dataset.

Method	**OF**	**SP** [10]	**SF** [8]	**SRC** [12]	**IEP** [11]	**BoTG** [14]	**Proposed Method**
AUC	0.86	0.9	0.96	0.98	0.985	0.99	0.985

D. Crowd Event Recognition

For crowd event recognition, five different crowd event classes from PETS2009 S3 dataset are classified using the proposed method. From each class, 60% of the clips are randomly selected for training and the rest are used for testing. Table III shows the average classification scores in percentage using VQ, SA and LLC encoding methods with sum and max pooling techniques. Here, the LLC method with sum pooling technique performs best.

TABLE III: Classification of crowd scenes from PETS2009 S3 dataset using different encoding (VQ, SA and LLC) and pooling (sum and max) techniques.

PETS2009 S3 dataset	**VQ**		**SA**		**LLC**	
	Max	**Sum**	**Max**	**Sum**	**Max**	**Sum**
Spatial	81.5	83.1	81.5	76.9	83.1	84.6
Temporal	83.1	86.1	83.1	73.8	87.7	89.2
Spatial+Temporal	92.3	89.2	89.2	83.1	92.3	**93.8**

Fig. 6 shows the confusion matrix of the proposed method for five different classes classified using spatial+temporal descriptor. We can see that running and evacuation classes are confused with each other due to the similarity of their dynamics.

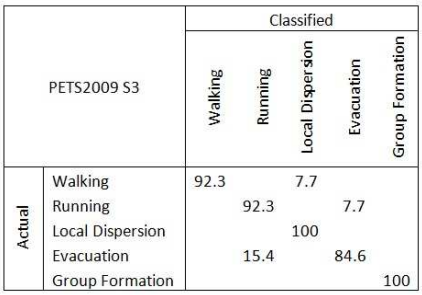

Fig. 6: Confusion matrix of crowd event recognition on PETS2009 S3 dataset using the proposed method.

Table IV presents the overall classification scores for crowd event recognition for the proposed method and the state-of-the-art BoTG method. The proposed method outperforms BoTG by 2.6%.

TABLE IV: Comparison of the proposed method with state-of-the-art on PETS2009 S3 dataset.

Method	**BoTG** [12]	**Proposed Method**
Classification Accuracy	91.2	**93.8**

BoTG works slightly better than the proposed method for the UMN dataset. The reason being, BoTG uses orientation distribution as a group attribute while generating trajectory graph. This static feature gives BoTG a better representation. But the same feature is likely to degrade the performance if there is a variation in view point and frame scale (PETS2009 S3 dataset contains four different view points). That is why BoTG does not outperform the proposed method on PETS2009 S3 dataset.

IV. Conclusion

This paper presents a new method for crowd behavior recognition based on dense trajectories. Local features of

dynamic nature are extracted through HOG and MBH descriptors, calculated along motion trajectories. Different encoding and pooling techniques for global feature representation are tested on two datasets. The LLC encoding along with the sum pooling provides the highest classification rates, using linear support vector machine classifier.

The proposed method was compared in terms of classification accuracy with several state-of-the-art methods; the proposed method outperforms most of them.

REFERENCES

[1] D. Helbing, P. Molnar, I. J. Farkas and K. Bolay, "Self-organizing pedestrian movement," *Environment and Planning B: Planning and Design*, vol. 28, no. 3, pp. 361–383, 2001.

[2] A. Basharat, A. Gritai and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[3] S. Pellegrini, A. Ess, K. Schindler and L. V. Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," *In Proc. IEEE International Conference on Computer Vision*, pp. 261–268, 2009.

[4] Z. Cheng, L. Qin, Q. Huang, S. Jiang and Q. Tian, "Group activity recognition by Gaussian processes estimation," *In Proc. IEEE International Conference on Pattern Recognition*, pp. 3228–3231, 2010.

[5] E. L. Andrade, S. Blunsden and R. B. Fisher, "Modelling Crowd Scenes for Event Detection," *In Proc. IEEE International Conference on Pattern Recognition*, pp. 175–178, 2006.

[6] S. Ali and M. Shah, "A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, 2007.

[7] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1446–1453, 2009.

[8] R. Mehran, A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 935–942, 2009.

[9] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, pp. 4282–4286, 1995.

[10] R. Mehran, B. Moore and M. Shah, "A streakline representation of flow in crowded scenes," *In Proc. European Conference on Computer Vision*, pp. 439–452, 2010.

[11] X. Cui, Q. Liu, M. Gao and D. N. Metaxas, "Abnormal detection using interaction energy potentials," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3161–3167, 2011.

[12] 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.

[13] W. Li, V. Mahadevan and N. Vasconcelos, "Anomaly Detection and Localization in Crowded Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, 2014.

[14] Y. Zhang, L. Qin, H. Yao, P. Xu and Q. Huang, "Beyond particle flow: Bag of Trajectory Graphs for dense crowd event recognition," *In Proc. IEEE International Conference on Image Processing*, pp. 3572–3576, 2013.

[15] H. Wang, A. Klaser, C. Schmid and C. -L. Liu, "Action recognition by dense trajectories," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3169–3176, 2011.

[16] H. Wang, A. Klaser, C. Schmid and C. -L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, pp. 60–79, 2013.

[17] N. Dalal, B. Triggs and C. Schmid, "Human detection using oriented histograms of flow and appearance," *In Proc. European Conference on Computer Vision*, pp. 428–441, 2006.

[18] X. Peng, Y. Qiao, Q. Peng and X. Qi, "Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition," *In Proc. British Machine Vision Conference*, pp. 1–11, 2013.

[19] "Unusual crowd activity dataset," *University of Minnesota*. available at http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi

[20] "PETS2009 dataset," *International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009. available at http://www.cvg.rdg.ac.uk/PETS2009/a.html#s3

[21] G. Farneback, "Two-frame motion estimation based on polynomial expansion," *In Proc. Scandinavian Conference on Image Analysis*, pp. 363–370, 2003.

[22] J. Shi and C. Tomasi, "Good features to track," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994.

[23] N. Otsu, "A threshold selection method from gray-level histogram," *In IEEE Trans. on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[24] C. M. Bishop, "Pattern Recognition and Machine Learning," *Springer*, 2006.

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang and Y. Gong, "Locality-constrained linear coding for image classification," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367, 2010.

[26] S. Lazebnik, C. Schmid and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, 2006.

[27] F. Perronnin, J. Sanchez and T. Mensink, "Improving the fisher kernel for large-scale image classification," *In Proc. British Machine Vision Conference*, pp. 143–156, 2010.

[28] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[29] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang and C. J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008. Code available at http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[30] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of keypoints," *In Proc. European Conference on Computer Vision*, pp. 1–22, 2004.

[31] J. V. Gemert, J. M. Geusebroek, C. J. Veenman and A. W. M. Smeulders, "Kernel codebooks for scene categorization," *In Proc. European Conference on Computer Vision*, pp. 696–709, 2008.

[32] J. Yang, K. Yu, Y. Gong and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1794–1801, 2009.