Abnormal Crowd Behavior Detection Using Motion Information Images and Convolutional Neural Networks

CEM DIREKOGLU[ID]

Electrical and Electronics Engineering Department, Middle East Technical University Northern Cyprus Campus, 99738 Kalkanlı, Güzelyurt, Mersin 10, Turkey

e-mail: cemdir@metu.edu.tr

ABSTRACT We introduce a novel method for abnormal crowd event detection in surveillance videos. Particularly, our work focuses on panic and escape behavior detection that may appear because of violent events and natural disasters. First, optical flow vectors are computed to generate a motion information image (MII) for each frame, and then MIIs are used to train a convolutional neural network (CNN) for abnormal crowd event detection. The proposed MII is a new formulation that provides a visual appearance of crowd motion. The proposed MIIs make the discrimination between normal and abnormal behaviors easier. The MII is mainly based on the optical flow magnitude, and angle difference computed between the optical flow vectors in consecutive frames. A CNN is employed to learn normal and abnormal crowd behaviors using MIIs. The MII generation, and the combination with a CNN is a new approach in the context of abnormal crowd behavior detection. Experiments are performed on commonly used datasets such as UMN and PETS2009. Evaluation indicates that our method achieves the best results.

INDEX TERMS Crowd behavior analysis, anomaly detection, motion information image, convolutional neural network.

I. INTRODUCTION

Analysis of crowd behavior has become a popular research field in recent years. Crowd behavior analysis can be utilized in variety of applications, for example, automatic detection of panic and escape behavior as a result of violence, riots, natural disasters, and so forth. Generally it is challenging to find effective features for crowds, since people inside the crowd may be positioned at different locations and may move in diverse directions. As a result, higher level analysis becomes difficult.

According to [1], abnormal event detection can be classified as local and global abnormal events. Local abnormal events contain an individual who acts differently from the rest of the individuals within a crowded scene. In global abnormal events, crowd behavior inside a global scene is abnormal, such as sudden escape of people during an earthquake. This work focuses on global abnormal crowd behavior detection.

A. RELATED WORK

For global crowd behavior analysis usually holistic and object-based methods are utilized. In object-based

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Anisetti[ID].

approaches, the crowd consists of groups of individuals (objects) [2], [3] and these objects are detected and tracked in order to understand the global crowd behavior [4]. Major challenges of object-based methods are accurate object identification, tracking and action recognition in dense crowds, since occlusions affect the whole process. Alternatively, in holistic methods [5]–[7], the crowd is considered as a global unit. Thus, these approaches analyzes the whole crowd itself to extract useful features (e.g. applying optical flow to the entire frame) in order to detect the crowd behavior.

In this research, we concentrate on global abnormal crowd event detection in surveillance videos, for example, sudden escape of people in the same or diverse directions. Anomaly detection consists of two main phases: event representation and anomaly measurement. For abnormal event representation, spatial-temporal information can be used, for example social force model [7], Histogram of Optical Flow (HOF) [8], Histogram of Motion Direction (HMD) [9], spatial-temporal gradient [10], chaotic invariant [11], mixtures of dynamic textures [12], sparse representation [1] and behaviour Entropy model (BE) [13]. For anomaly measurement, most of the approaches employ a one-class learning methods to learn normal samples. As a one-class learner Hidden Markov Model [10], Gaussian Mixture Model, one-class Support

Vector Machine (SVM) [14], Replicator Neural Networks [15], Convolutional Neural Networks [16], [17] and Bayesian model [18] can be utilized. Then, during testing, if the test sample is significantly different from the normal, it is accepted as abnormal.

There are also very recent works on abnormal crowd behaviour detection based on distribution of magnitude of optical flow (DMOF) [19], context location and motion-rich spatio-temporal volumes (CL and MSV) [20], generative adversarial nets (GAN) [21], temporal convolutional neural network pattern (TCP) [22], global event influence model (GEIM) [23], and histograms of optical flow orientation and magnitude (HOFO) [24]. Reviews on crowd behaviour analysis can be found at [25], [26]. Recently, some survey papers have also appeared for deep learning based crowd behaviour analysis [27], [28]. Below, in part *B*, we particularly summarize existing optical flow based methods both for global and local crowd anomalies since our work is also based on optical flow, and then, in part *C*, we explain our method and the difference from existing works.

B. OPTICAL FLOW BASED METHODS

Here, we summarize optical flow based methods both for global and local crowd anomalies. Social force model [7] focuses on global anomaly. A grid of particles is placed over the image plane, and they are advected with the space-time average of optical flow. Then interaction force, between particles, is estimated using social force model. The interaction force is then mapped into the image region to obtain Force Flow for every pixel in every frame. The normal crowd behaviour is modelled using the Force Flow frames. Finally, bag of words approach is used to classify frames as normal and abnormal. In [8], motion feature is obtained after binning the current optical flow distribution into angular bins, yielding a one dimensional vector on flow directions for local anomaly detection. In chaotic invariant [11], the process begins with particle advection using optical flow. Then particle trajectories are clustered to obtain representative trajectories for a crowd flow. Next, the chaotic dynamics of all representative trajectories are extracted. Probabilistic model is learned from these chaotic feature set, and finally, a maximum likelihood estimation criterion is utilized to identify a global abnormal or normal behaviour. They can also predict the location. Sparse representation [1] method uses a multi-scale histogram of optical flow (MHOF) that also preserves spatial contextual information to identify local and global anomalies. They concatenate optical flow direction and energy (magnitude) information at multiple scales to generate a motion histogram. Behaviour Entropy model (BE) [13] use optical flow magnitude information in local regions to model behavior certainty, behavior entropy, scene behavior entropy in order to analyse crowd behaviour for global and local anomaly detection. Gnouma *et al.* [19] present a method based on local distribution of magnitude of optical flow (DMOF) for global anomaly detection. Patil and Biswas [20] also proposed a method for global anomaly

detection. First, optical flow is computed at each frame of a video. Then the video is divided into spatio-temporal volumes (STV). In each volume, mean value of the optical flow magnitudes is computed. Next, STVs with the higher mean values are used for testing the anomaly. Histogram of flow orientation information to together with mean value of the flow magnitudes in that volume is used as a feature vector for abnormal crowd behaviour detection. Generative adversarial nets (GAN) [21] use optical flow magnitude images for global and local abnormal behaviour detection. Ravanbakhsh *et al.* [22] fuse appearance and optical flow magnitude image using a convolutional neural network for global and local abnormal crowd behaviour detection. Pan *et al.* [23] performs global abnormal behaviour detection using a combination of features such as combination of scale, velocity and disorder features. In their work, velocity feature is based on optical flow magnitude. Colque *et al.* [24] also proposed an optical flow based feature descriptor for global and local anomaly detection. These features are represented by histograms of optical flow orientation and magnitude and entropy. This is a three-dimensional histogram consist of orientation, magnitude and entropy of orientation dimensions.

C. CONTRIBUTION

We present a new work for abnormal crowd event detection. The key contribution is new motion information image (MII) generation using optical flow. The proposed MIIs can represent and discriminate normal and abnormal events well, and when MIIs are input to a CNN for training and testing, it achieves very promising results in this domain. Both normal and abnormal MIIs are trained using a CNN that means we have two categories in the CNN network. According to our observation, during an abnormal event, people start to run. Especially in the motion regions, this abnormal behavior increases the angle difference between the optical flow vectors computed in the previous frame and in the current frame at each pixel location. In addition, we also observe that the optical flow magnitude increases too. We introduce a mathematical formulation to produce a MII. As a first step, optical flow angle differences are computed for each pixel location based on the current frame and the previous frame. However, some optical flow measurements are small and noisy, and their angle difference affect the observation. To overcome this problem, the angle difference is multiplied with the optical flow magnitude computed in the current frame, and form the MII. We compute a MII for each frame. Finally, a CNN is used to learn normal and abnormal crowd behaviors using MIIs. In the testing phase, the CNN classifies the input MII image.

It is important to emphasize that though there are many optical flow based algorithms have been introduced for crowd behavior understanding, the MII generation is a completely new concept that is based on the angle difference between optical flow vectors in consecutive frames, and the optical flow magnitude in the current frame. Our studies show that when MIIs are combined with CNN for classification, it outperforms the existing methods in abnormal crowd behavior

detection. Our experiments are performed on two commonly used public datasets, such as UMN [29] and PETS2009 [30]. Results illustrate that our method achieves the best results in both datasets.

In our preliminary work [14], optical flow-based features are used together with one class SVM for abnormal crowd behaviour detection. In [14], we created a one-dimensional feature vector based on a combination of optical flow magnitude and optical flow angle difference information. The proposed feature vectors are extracted for frames representing normal behaviour, and then we use a one class SVM to train these feature vectors. Finally, if a test frame is significantly deviating from the normal type, it is labelled to be abnormal. Our earlier work is significantly different from the current work since in this paper we generate a novel MII representation that provides a visual appearance of crowd motion. The MIIs are input to CNN for training and testing of two classes: Normal and Abnormal crowd behaviours. It is also important to note that Hatirnaz *et al.* [31] adopted our preliminary work [14] to develop a concept-based semantic search interface. They use semantic web technologies to improve video retrieval for abnormal crowd behaviors in a surveillance system. The novelty of this work is about using semantic web technologies for annotation of surveillance videos and developing an intelligent semantic search interface. They use the existing work in [14] for crowd behavior feature extraction.

In this paper, Section 2 introduces motion information image (MII) generation. Section 3 presents abnormal crowd event detection using CNN. Section 4 presents experiments in UMN and PETS2009 datasets, as well as, discusses parameter selection and computational complexity evaluations. Section 5 is conclusions.

II. MOTION INFORMATION IMAGE GENERATION
The proposed motion information image (MII) generation is based on optical flow. The optical flow at each frame is computed using the Lucas-Kanade algorithm [32]. In a panic situation, each person in the crowd may move in different directions or in the same direction. Therefore, the MII must be invariant to the direction of movement, as well as it must be discriminative enough so that the normal and abnormal events can be separated at every time frame. For example, in Fig. 2 (a), when we look at the first and third images, we can observe that each person is moving (e.g. scattering) mostly towards different directions. On the other hand, in the second image in Fig. 2(a), everyone is moving towards the same direction (i.e. to the right side). All of these situations are panic and escape situation despite the direction of movement of each person (whether everyone moves in different direction or same direction). Thus, the MII must be invariant to the direction of motion, and it must be discriminative so that the normal and abnormal events can be identified at every time frame.

In an unusual situation, people panic and scatter around. In such a situation, we observe that, especially in motion areas, the angle difference between the optical flow vectors
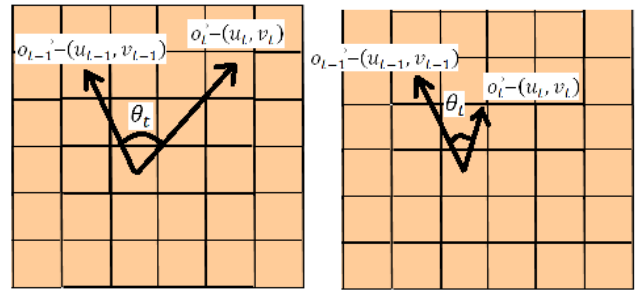
FIGURE 1. Optical flow angle difference. (a) Observed behaviour in abnormal situation, (b) and in normal situation.

in consecutive frames increases at each pixel location. The angle difference between two vectors, at each pixel location, is calculated as follows:

$$\theta_t(x, y)$$
$$= \arccos\left(\frac{(u_{t-1}(x, y).u_t(x, y) + v_{t-1}(x, y).v_t(x, y))}{\left(\sqrt{u_{t-1}^2(x, y) + v_{t-1}^2(x, y)} \cdot \sqrt{u_t^2(x, y) + v_t^2(x, y)}\right)}\right)$$
(1)

where $\overrightarrow{o_{t-1}}(x, y) = (u_{t-1}(x, y), v_{t-1}(x, y))$ and $\overrightarrow{o_t}(x, y) = (u_t(x, y), v_t(x, y))$ are optical flow vectors, respectively, in the previous frame $(t-1)$ and in the current frame (t) at each pixel location (x, y). θ_t is the angle difference at the current frame. The optical flow angle difference between these two vectors is also shown in Fig. 1 (a) and (b). To our observation, the angle difference appears to be higher, as shown in Fig. 1 (a), when there is an abnormal behaviour (i.e. Escape or panic situation), and the angle difference is smaller as in Fig. 1 (b) when the behavior is normal. However, there are also some optical flow measurements appear on the image not because of object motion but because of noise or lighting change in still areas (no motion areas). In still areas, under ideal conditions, optical flow measurements should be zero (magnitude is zero, and angle difference is zero). However in practical applications, on real world images, optical flow measurements usually appear to have small optical flow magnitude in still areas because of noise or lighting change. The angle difference between the vectors in consecutive frames may be higher in still areas. We don't want these noisy measurements to affect our observation since MIIs are based on angle difference of vectors in consecutive frames. To overcome this problem, the angle difference is multiplied with the optical flow magnitude computed in the current frame as illustrated below,

$$I_t(x, y) = \sqrt{u_t^2(x, y) + v_t^2(x, y)}.\theta_t(x, y) \qquad (2)$$

where $\sqrt{u_t^2(x, y) + v_t^2(x, y)}$ is the optical flow magnitude in the current frame (t) at each pixel location (x, y). $\theta_t(x, y)$ is the angle difference calculated in Equation 1. I_t represents the motion information image (MII) for the current frame (t). If magnitude and angle difference values are high,
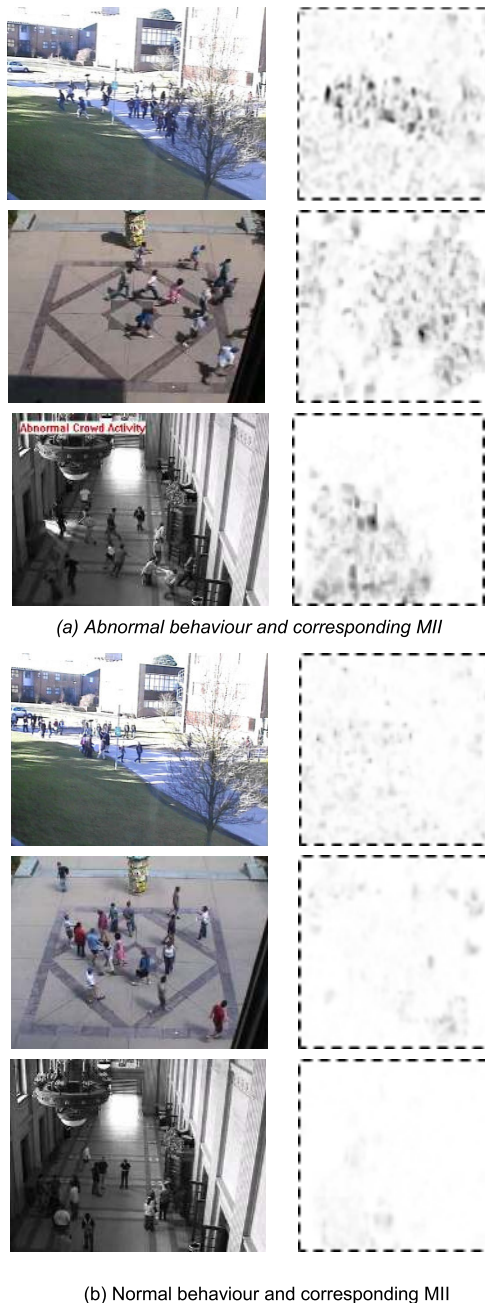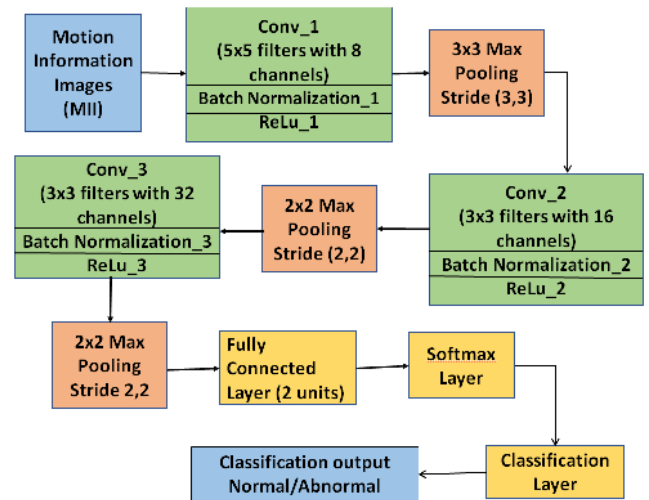
(a) Abnormal behaviour and corresponding MII

(b) Normal behaviour and corresponding MII

FIGURE 2. Some example frames representing abnormal and normal behaviours and their corresponding MIIs.

the multiplication output will be high as well (this is a case for motion regions). If magnitude is small and the angle difference is high, the multiplication output will be smaller (this may be a case for still regions). If magnitude is small and the angle difference is small, the multiplication output will be even smaller. Therefore, multiplying magnitude with the angle difference generates differences on MIIs. This process generates a significant difference between the abnormal and normal Motion Information Images (MIIs). In Section IV (E), we demonstrate that this multiplication improves the performance considerably.

Fig. 2 (a) shows some example frames representing abnormal behaviour, and their corresponding MII. On the other

FIGURE 3. The CNN structure.

hand, Fig. 2 (b) illustrates some example frames representing normal behaviour and their corresponding MII. All of the MIIs are resized to have dimensions 75×75 that will be input to a CNN. In addition, for better illustration, the MIIs are inverted in Fig. 2. It can be observed that MIIs produced by abnormal behaviours are significantly different than the MIIs produced by normal behaviours.

III. CNN TRAINING AND CLASSIFICATION

We use a simple 2D CNN structure, and train the CNN network with MIIs to achieve abnormal crowd behavior detection. In the CNN network, there are two classes: Normal and Abnormal behavior.

A. CNN ARCHITECTURE AND TRAINING

MIIs are experimented with simple CNN architectures with varying number of convolutional layers and channels, various filter sizes, and pooling layers to achieve the best accuracy in both datasets. We observe that the CNN architecture that uses MIIs for abnormal crowd behavior detection can be the one with three convolutional layers as shown in Fig. 3. It is also important to note that we also tested the MIIs with popular deep networks in image recognition, where the results are presented in Section IV-C.

In the simple CNN structure, shown in Fig. 3, MII inputs are resized to have dimensions 75×75. The first convolution layer uses 5×5 filters with 8 channels. After that we perform a batch normalization, a rectified linear unit (ReLu) activation, and a max pooling (with 3×3) operation. The second convolutional layer uses 3×3 filters with 16 feature maps followed by a batch normalization, a ReLu activation, and 3×3 max pooling. The last convolutional layer utilizes 3×3 filters with 32 feature maps followed by a batch normalization, a ReLu, and 2×2 max pooling. Then, we form a fully connected layer with two nodes since we have two classes, and finally employ the softmax layer for predictions. The input MII image is recognized as normal or abnormal using the classification layer. During the training a stochastic gradient

TABLE 1. Analysis of the CNN network.

Layer	Name	Activation Resolutions	Learnables	Total Learnables
1	MII input	75x75x1	-	0
2	Conv_1 (8 5x5x1 convolutions)	71x71x8	Weights: 5x5x1x8 Bias: 1x1x8	208
3	BatchNorm_1 (with 8 Channels)	71x71x8	Offset:1x1x8 Scale:1x1x8	16
4	ReLu_1	71x71x8	-	0
5	Maxpool_1 (Block 3x3, Stride 3x3)	23x23x8	-	0
6	Conv_2 (16 3x3x8 convolutions)	21x21x16	Weights: 3x3x8x16 Bias: 1x1x16	1168
7	BatchNorm_2 (with 16 Channels)	21x21x16	Offset:1x1x16 Scale:1x1x16	32
8	ReLu_2	21x21x16	-	0
9	Maxpool_2 (Block 2x2, Stride 2x2)	10x10x16	-	0
10	Conv_3 (32 3x3x16 convolutions)	8x8x32	Weights: 3x3x16x32 Bias: 1x1x32	4640
11	BatchNorm_3 (with 32 Channels)	8x8x32	Offset:1x1x32 Scale:1x1x32	64
12	ReLu_3	8x8x32	-	0
13	Maxpool_3 (Block 2x2, Stride 2x2)	4x4x32	-	0
14	Fully-Connect (2 units)	1x1x2	Weights: 2x512 Bias: 2x1	1026
15	Sofmax	1x1x2	-	0
16	Class Output (Normal or Abnormal)	-	-	0
Total Learnable Parameters in the Network =				7154

descent with momentum method is used as a solver. The learning rate is 0.01, mini-batch size is 50, and the maximum number of epoches is 10. These parameter values are determined experimentally to achieve the best performances with the proposed MIIs. The same CNN structure and parameter values are utilized both in UMN and PETS2009 datasets. TABLE 1 also shows the details of the network such as activation map resolutions at each layer, total learnable parameters at each layer, and in the whole network.

B. CLASSIFICATION USING THE MIIs AND CNN

The MIIs of test frames are obtained as it is explained in Section II. In the UMN dataset, the test frame is recognized using the 28-by-28 neighbourhood frames that means the window size is 57 (including the test frame). Each of the frames in the window is labelled with the CNN classifier, and then the most frequent class represents the behaviour (normal or abnormal) of the test frame. On the other hand, in the PETS 2009 dataset, the test frame is recognized using the 21-by-21

neighbourhood frames that means that the window size is 43 (including the test frame). Each frame in the window is labelled with CNN classifier, and then the most frequent class represents the behaviour (normal or abnormal) of the test frame. The window size for the UMN and PETS datasets are determined experimentally that will be discussed in the evaluations section.

IV. EVALUATION AND RESULTS

Experiments are performed on commonly used, and publicly available two different datasets in this domain: UMN and PETS2009 datasets. The proposed work is also compared to the existing works in this domain (global anomaly detection) such as Optical Flow Features (OFF) [14], the method based on Bayesian model (BM) [18], sparse reconstruction cost (SRC) [1], chaotic invariants (CI) [11], the social force model (SF) [7], the force field model (FF) [33], behaviour Entropy model (BE) [13], distribution of magnitude of optical flow (DMOF) [19], context location and motion-rich spatio-temporal volumes (CL and MSV) [20], generative adversarial nets GAN [21], temporal CNN Pattern (TCP) [22], global event influence model (GEIM) [23], and histograms of optical flow orientation and magnitude (HOFO) [24]. We measure the accuracy of the methods, which is the percentage of correctly classified frames in comparison to the ground truth. The same accuracy measurement has been employed by the methods above. It is also important to note that, recently, Sultani et al. [34] also proposed a framework for anomaly detection. However, they need a large dataset for training since they use the complex C3D [35] network to learn spatiotemporal features with 3D convolutions. They constructed a large datasets (1900 videos) consisting of surveillance videos for abnormal events. However, the anomalies, in their dataset, involve behaviors by individuals (one or a few people performing abnormal actions such as one person abusing an animal, one person breaking the class of a shop, two persons stealing something from a car). Their dataset is not about crowd behavior. Although the method looks effective in their dataset, they did not experiment their work on UMN [29] and PETS2009 [30] for crowd behavior analysis mainly because they need large datasets for training.

A. EVALUATION ON UMN DATASET

The UMN dataset [29] consist of 11 videos, and each video contains normal and abnormal crowd behaviors. There are three different scenes in this dataset (two outdoor scenes and one indoor scene). Scene 1 is an outdoor scene that consist of two videos (e.g. Video 1 and Video 2). For testing the Video 1, we use the MIIs of Video 2 to train the CNN. Similarly, for testing the Video 2, the MIIs of Video 1 are used for training the CNN.

Scene 2 consists of six videos that are captured in an indoor environment. While testing a particular video in scene 2, we leave the testing video out, and use the MIIs of the rest of the videos in scene 2 to train the CNN.

TABLE 2. Accuracy comparison of methods in the UMN dataset.

	MII+CNN (Our Work)	OFF	BM	FF	CI	SF	SRC	DMOF	GAN	TCP	GEIM	BE
Scene 1	98.55	99.10	99.03	88.69	90.62	84.41	90.52	98.84	-	-	99.18	-
Scene 2	98.91	94.85	95.36	80.00	85.06	82.35	78.48	97.72	-	-	98.03	-
Scene 3	99.77	97.76	96.63	77.92	91.58	90.83	92.70	98.7	-	-	98.19	-
Overall Accuracy	99.08	96.46	96.40	81.04	87.91	85.09	84.70	98.42	99	98.8	98.47	99

FIGURE 4. Crowd escape behaviour detection results for Scene #1 in the UMN dataset. (a) Ground truth, (b) detection by the proposed work, (c) detection by OFF, (d) detection by BM, (e) detection by FF, (f) detection by CI, (g) detection by SF, (h) detection by SRC, (i) detection by DMOF, and (k) detection by GEIM.

FIGURE 5. Crowd escape behaviour detection results for Scene #2 in the UMN dataset. (a) Ground truth, (b) detection by the proposed work, (c) detection by OFF, (d) detection by BM, (e) detection by FF, (f) detection by CI, (g) detection by SF, (h) detection by SRC, and (i) detection by GEIM.

Scene 3 has three videos that are captured in an outdoor environment. While testing a particular video in scene 3, we leave the testing video out, and use the MIIs of the rest of the videos in scene 3 for training purpose.

TABLE 2 illustrates the accuracy of twelve methods for three different scenes. In overall, the proposed method (MII + CNN) outperforms the existing methods in UMN dataset with accuracy 99.08 %. The overall accuracy of proposed method is better than the recently published works such as GAN (99 %), BE (99 %), TCP (98.8 %), GEIM (98.47 %) and DMOF (98.42%). GAN, TCP and BE methods did not provide their performances for individual scenes. Our earlier work, OFF, achieves 97.32 %. We improve the performance of the earlier work, as well as perform better than the older works in this dataset such as the accuracy of BM (96.40%), FF (81.04%), CI (87.91%), SF (85.09%) and SRC (84.70%) stay below the proposed method.

To summarize, our work provides the best results in UMN dataset. In addition to these results, visual accuracy comparisons of methods for the UMN dataset are illustrated in Fig. 4 and Fig. 5. These illustrations show that our work performs very accurate segmentation in comparison to other

works. The proposed method can detect the start and the end times of the abnormal event very well.

B. EVALUATION ON PETS DATASET
In PETS2009 dataset [30], there is one scenario about abnormal crowd behaviour. This scenario was captured from four different cameras location (four different viewpoints), resulting in 4 videos. In this scenario, people come to centre from different directions, wait there for a while and suddenly they start to run around in random directions. Although the same action is performed, there are significant differences because of different viewpoints. For example, the distance between camera and the crowd appears to be different for each viewpoint, the lighting conditions, distribution and location of people and objects appear to be different as well. Therefore, we evaluate the accuracy for these four different viewpoints. Each video consist of 374 frames. For testing view #1 video of the scenario, the MIIs of the other views are used for training the CNN. For testing view #2 of the scenario, the MIIs of the other views are used for training the CNN. Similarly, we test view #3 and view #4.

TABLE 3 illustrates the accuracy of seven methods for this scenario. In overall, our work (MII + CNN) outperforms

TABLE 3. Accuracy comparison of the methods in the PETS2009 dataset.

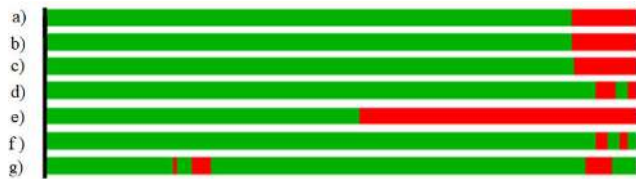|	MII+CNN (Our Work)	OFF	BM	FF	CI	SF	DMOF
View 1	99.12	98.66	96.01	94.50	94.95	91.22	99.00
View 2	99.73	99.20	94.15	63.83	92.02	89.36	99.70
View 3	99.12	99.47	95.21	95.48	94.15	94.68	99.34
View 4	95.45	89.57	91.49	96.81	89.36	64.63	87.72
Overall Accuracy	98.39	96.72	94.22	87.66	92.62	84.97	96.44

FIGURE 6. Crowd escape behaviour detection results for view #2 video of PETS2009 dataset. (a) Ground truth, (b) detection by the proposed work, (c) detection by OFF, (d) detection by BM, (e) detection by FF, (f) detection by CI and (g) detection by SF.

the other methods with accuracy 98.39 %. Other methods OFF (96.72%) BM (94.22%), FF (87.66%), CI (92.62%), SF (84.97%) and DMOF (96.44%) achieves worse than the proposed method. Only in view #4, our work ranked slightly behind the FF method. This is mainly because of the resolution problem in view #4. In view #4, there is a low resolution problem, and this is why almost all of the methods have a lower performance there in comparison to other views.

A visual accuracy comparison is also illustrated in Fig. 6 for the PETS2009 dataset. Our work performs very accurate anomaly detection in comparison to existing methods.

C. EVALUATION OF MIIs WITH OTHER NETWORKS

We also evaluate the proposed MIIs with popular deep networks for Anomaly detection such as with ResNet-50 [36], GoogleNet-V3 [37], DenseNet-250 [38], and the CNN network proposed by Oquab *et al.* [39] that is the improved version of AlexNet proposed by Krizhevsky *et al.* [40]. We performed transfer learning to tune the network parameters for two possible classes: Normal and Abnormal actions. Overall accuracy results both in UMN and PETS datasets are presented below in TABLE 4. We also included results with the simple CNN presented in Section III. During the transfer learning, particularly we adjusted the input image

TABLE 4. The proposed MIIs with different networks.

Overall Accuracy	MII + Simple CNN (7154 weights)	MII + ResNet-50 (~25 Million weights)	MII + GoogleNet-V3 (~23 Million weights)	MII + DenseNet-250 (~15 Million weights)	MII + Oquab et al. (~61 Million weights)
UMN Dataset	99.08	98.99	99.04	98.49	98.22
PETs Dataset	98.39	97.04	98.25	98.31	97.12

size according to a pre-trained network, and replace the final layers to have only two classes. Results show that all networks achieve similar results with the MIIs.
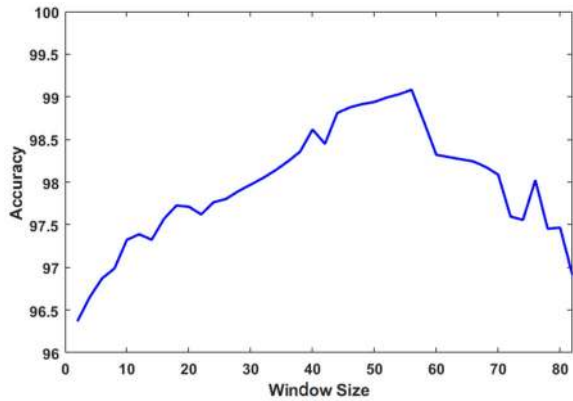
D. THE IMPACT OF WINDOW SIZE

In both UMN and PETS2009 datasets, the impact of changing window size is evaluated. Fig. 7 (a) and (b) show the overall performance of proposed method with changing window size in UMN and PETS2009 datasets, respectively. It is seen that the best window size for our approach (MII + CNN) in the UMN dataset is 57. In PETS2009 dataset, the optimal window size for MII + CNN is 43.
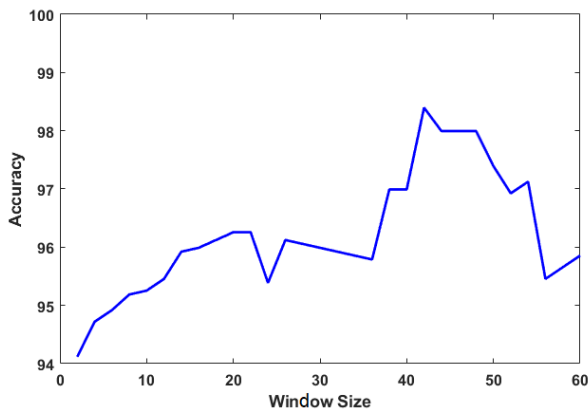
E. THE INFLUENCE OF ANGLE DIFFERENCE AND OPTICAL FLOW MAGNITUDE

We experiment the effect of angle difference and optical flow magnitude on detection accuracy and compare with the combination. In particular, we expect to observe higher detection accuracy for combination of optical flow angle difference and optical flow angle magnitude (i.e. multiplication of them as shown in Equation 2), in comparison to only using angle difference or only using magnitude. Results for the UMN dataset are presented in TABLE 5. Only angle difference performs 91.30%, only optical flow magnitude performs 94.34%, and the combination performs 99.08%. These results show that the combination increases the accuracy remarkably.

Results for PETS2009 dataset are presented in TABLE 4. Only angle difference performs 93.25%, only optical flow magnitude performs 85.49% and the combination performs 98.39%. Similar to the results in the UMN dataset, results in the PETS2009 dataset also confirm that the combination improves the accuracy significantly.

(a)

(b)

FIGURE 7. The effect of window size on accuracy. (a) UMN dataset. (b) PETS2009 dataset.

TABLE 5. The influence of angle difference and optical flow magnitude on accuracy (%) in the UMN dataset.

	Only Angle Difference (Optimal Window Size 45)	Only Magnitude (Optimal Window Size 53)	Combined (Optimal Window Size 57)
Scene 1	90.49	94.42	98.55
Scene 2	88.44	94.01	98.91
Scene 3	97.38	94.90	99.77
Overall Accuracy	91.30	94.34	99.08

TABLE 6. The influence of angle difference and optical flow magnitude on accuracy (%) in the PETS dataset.

	Only Angle Difference (Optimal Window Size 25)	Only Magnitude (Optimal Window Size 55)	Combined (Optimal Window Size 43)
View 1	98.39	89.84	99.12
View 2	89.84	82.35	99.73
View 3	94.92	89.84	99.12
View 4	89.84	79.95	95.45
Overall Accuracy	93.25	85.49	98.39

F. COMPUTATION TIME

The computational time for each phase of our method in both UMN and PETS2009 datasets are shown in TABLE 7 and TABLE 8, respectively. Results are obtained using Matlab

TABLE 7. Computation time on UMN dataset.

Performance (Average of 5 runs)	MII Formation	CNN Training	Testing
Scene 1 (2 videos)	1810.38 sec (2 videos, 1451 frames)	11.50 sec (1 video)	3.07 sec (1 video)
Scene 2 (6 videos)	6262.88 sec (6 videos, (4140 frames)	33.73 sec (5 videos)	3.42 sec (1 video)
Scene 3 (3 videos)	2657.12 sec (3 videos, 2137 frames)	17.13 sec (2 videos)	3.12 sec (1 video)

TABLE 8. Computation time on PETS2009 dataset.

Performance (Average of 5 runs)	MII Formation	CNN Training	Testing
View 1	2421.6 sec (4 videos/views, 1496 frames)	14.39 sec (3 videos: view #2, #3 and #4)	2.13 sec (1 video: view #1)
View 2	2421.6 sec (4 videos, 1496 frames)	13.32 sec (3 videos: view #1, #3 and #4)	2.09 sec (1 video: view #2)
View 3	2421.6 sec (4 videos, 1496 frames)	14.73 sec (3 videos: view #1, #2 and #4)	2.10 sec (1 video: view #3)

2018 on a Windows 7 Operating System with Intel Core i7-6700, 2.60GHz and 16GB RAM. Results show that MII formation needs considerable amount of time comparing to other stages. In addition, we observe that once the CNN is trained, the testing stage is very fast.

V. CONCLUSIONS

We presented an approach for abnormal crowd behaviour detection. The proposed approach is based on a new Motion Information Image (MII) model that is formulated using optical flow. The MII depends on the angle difference calculated between the optical flow vectors in consecutive frames. There are also some optical flow measurements that are small, and their angle difference may affect the observation. To overcome this problem, the angle difference is multiplied with the optical flow magnitude in the current frame to generate the MIIs. A convolutional neural network (CNN) is used to learn normal and abnormal events, and when a test sample is input to the CNN, it is assigned to one of the two classes (Normal or Abnormal). Evaluations are conducted on publicly available UMN and PETS2009 datasets. Results indicate that the proposed work is very effective.

REFERENCES

[1] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognit.*, vol. 46, no. 7, pp. 1851–1864, Jul. 2013.

[2] J. S. Marques, P. M. Jorge, A. J. Abrantes, and J. M. Lemos, "Tracking groups of pedestrians in video sequences," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2003, p. 101.

[3] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified crowd segmentation," in *Proc. Eur. Conf. Comput. Vis.*, vol. 5305, 2008, pp. 691–704.

[4] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 594–601.

[5] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2006, pp. 175–178.

[6] 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.*, Jun. 2007, pp. 1–6.

[7] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.

[8] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[9] H. M. Dee and A. Caplier, "Crowd behaviour analysis using histograms of motion direction," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 1545–1548.

[10] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1446–1453.

[11] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2054–2060.

[12] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[13] W.-Y. Ren, G.-H. Li, J. Chen, and H.-Z. Liang, "Abnormal crowd behavior detection using behavior entropy model," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, Jul. 2012, pp. 212–221.

[14] C. Direkoglu, M. Sah, and N. E. O'Connor, "Abnormal crowd behaviour detection using novel optical flow-based features," Presented at the IEEE Int. Conf. Adv. Video Signal based Surveill. (AVSS), Aug. 2017.

[15] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," Presented at the Int. Conf. Data Warehousing Knowledge Discovery, Sep. 2002.

[16] Z. Fang, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, and S. Chen, "Abnormal event detection in crowded scenes based on deep learning," *Multimedia Tools Appl.*, vol. 75, no. 22, pp. 14617–14639, Nov. 2016.

[17] Y. Feng, Y. Yuan, and X. Lu, "Deep representation for abnormal event detection in crowded scenes," in *Proc. ACM Multimedia Conf. MM*, 2016, pp. 591–595.

[18] S. Wu, H.-S. Wong, and Z. Yu, "A Bayesian model for crowd escape behavior detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 85–98, Jan. 2014.

[19] M. Gnouma, R. Ejbali, and M. Zaied, "Abnormal events' detection in crowded scenes," *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 24843–24864, 2018.

[20] N. Patil and P. K. Biswas, "Global abnormal events detection in crowded scenes using context location and motion-rich spatio-temporal volumes," *IET Image Process.*, vol. 12, no. 4, pp. 596–604, Apr. 2018.

[21] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," Presented at the IEEE Int. Conf. Image Process. (ICIP), Sep. 2017

[22] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-Play CNN for crowd motion analysis: An application in abnormal event detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1689–1698.

[23] L. Pan, H. Zhou, Y. Liu, and M. Wang, "Global event influence model: Integrating crowd motion and social psychology for global anomaly detection in dense crowds," *J. Electron. Imag.*, vol. 28, no. 2, p. 1, Apr. 2019.

[24] R. V. H. M. Colque, C. A. C. Junior, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude to detect anomalous events in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Dec. 2017.

[25] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[26] 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, Feb. 2016.

[27] G. Tripathi, K. Singh, and D. K. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: A survey," *Vis. Comput.*, vol. 35, no. 5, pp. 753–776, May 2019, doi: 10.1007/s00371-018-1499-5.

[28] X. Wang and C. C. Loy, "Deep learning for scene independent crowd analysis," in *Group and Crowd Behavior for Computer Vision*. Cambridge, MA, USA: Academic, Jan. 2017, ch. 10, pp. 209–252.

[29] *University of Minnesota*. Accessed: Apr. 30, 2020. [Online]. Available: http://mha.cs.umn.edu/movies/crowdactivityall.

[30] *University of Reading, PETS 2009 Dataset S3 Rapid Dispersion*. Accessed: Apr. 30, 2020. [Online]. Available: http://www.cvg.rdg.ac.uk/PETS2009/a.html#s2l1

[31] E. Hatirnaz, M. Sah, and C. Direkoglu, "A novel framework and concept-based semantic search Interface for abnormal crowd behaviour analysis in surveillance videos," *Multimed Tools Appl.*, pp. 1–39, Feb. 2020, doi: 10.1007/s11042-020-08659-2.

[32] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[33] D.-Y. Chen and P.-C. Huang, "Motion-based unusual event detection in human crowds," *J. Vis. Commun. Image Represent.*, vol. 22, no. 2, pp. 178–186, Feb. 2011.

[34] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[39] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classiffcation with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

CEM DIREKOGLU received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Eastern Mediterranean University and the Ph.D. degree in computer vision from the University of Southampton. He is currently an Assistant Professor with the Electrical and Electronics Engineering Program, Middle East Technical University Northern Cyprus Campus (METU-NCC). Before joining to METU-NCC, he was a Postdoctoral Researcher with the Insight Centre for Data Analytics, School of Electronic Engineering, Dublin City University (DCU). Before DCU, he was also a Postdoctoral Researcher with the Graphics, Vision and Visualization Group, School of Computer Science and Statistics, Trinity College Dublin. He was a member of the Information: Signals, Images and Systems Research Group in the School of Electronics and Computer Science. He has expertise in image processing and computer vision. He has been involved in diverse computer vision-based research projects and has many publications in international journals and conferences, including the high impact ones as the lead author, such as in IJCV, PR, MVA, PRL, ECCV, BMVC, and ICIP. He is a reviewer of prestigious journals in his field, such as *Computer Vision and Image Understanding*, *Pattern Recognition*, *Image and Vision Computing*, and the IEEE Transactions on Circuits and Systems for Video Technology.

• • •