

# Recognition of crowd abnormal activities using fusion of handcrafted and deep features

Manasi Pathade, Madhuri Khambete

Department of Electronics and Telecommunication, MKSSS Cummins College of Engineering for Women, Pune, India

## Article Info

### Article history:

Received Jun 6, 2022

Revised Aug 5, 2022

Accepted Aug 29, 2022

### Keywords:

Deep networks

Density estimation

Merging

Optical flow

Sudden dispersion

Video surveillance

## ABSTRACT

Constant vigilance is extremely important at crowded public places where some unusual activities such as sudden dispersion or continuous gathering of people may lead to chaotic and disastrous situations. Automatic recognition of such collective activities of people is indeed an important task to ensure people safety. In this view, we propose a novel approach for automatic recognition of crowd merging and sudden dispersion events. The proposed method detects dispersion and merging using fusion of features extracted by deep networks with a novel set of optical flow based and density based handcrafted features. These proposed features are not affected by occlusion and illumination. These features complement the features extracted by deep network and their fusion improves the performance of event recognition by significant amount. The method is tested profoundly on benchmark public datasets as well as private datasets. Abnormal activity recognition data often suffers from high class imbalance. However, the proposed method could successfully recognize sudden dispersion and merging activities on such very small datasets having class imbalance. This proves the effectiveness and robustness of the proposed features. The proposed method also shows better performance than other state of the art methods based on deep networks.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Manasi Pathade

Department of Electronics and Telecommunication

MKSSS Cummins College of Engineering for Women Pune

Maharashtra, India

Email: manasi.pathade@cumminscollege.in

## 1. INTRODUCTION

In urban environment many public places are crowded most of the times. To ensure safety of the people, security forces typically monitor such places with the help of automated crowd event recognition systems equipped with closed circuit televisions (CCTV). For such places, special attention needs to be given to abnormal activities of the crowd such as sudden dispersion and merging. When people experience some threat; they get panic and suddenly start dispersing from each other. If this event is not recognized and controlled in time, it may lead to chaotic situation and such situation may claim lives of many people. On the other hand, continuous merging of more and more people/their groups at certain place may lead to congestion. If this event is not recognized and flow of people is not redirected, it may result in stampede like situation. Hence apt recognition of such events is essential to avoid future disasters.

To detect dispersion or merging, manually watching the CCTV video footages of any place is very difficult task. There is always a possibility of missing out some important clue by human operators due to fatigue caused by watching long video footages (which may run for several hours). In order to make this process less cumbersome and accurate, intelligent surveillance systems are being developed. With this

context, in this paper automated method to detect dispersion and merging events of crowd using CCTV footage is proposed.

In computer vision community, different methods are being proposed for crowd scene analysis and event recognition. Detailed review of these methods can be found in [1]–[4]. In most of the traditional methods [5]–[16], event recognition is performed by analyzing crowd motion. These methods have used different handcrafted features such as motion magnitude and direction [5], [8], [10], [14], [15], and [16], motion vector intersections [6], optical flow manifolds [9], dynamic textures [11] for characterizing the crowd behavior. Classification of events is carried out either in supervised way [5], [8], [14], [15] or unsupervised way [7], [9]–[13].

Since past few decades, researchers are widely using deep networks for various applications [17]–[39]. Al Mamun *et al.* [30] and Ahmed *et al.* [36], deep network based methods are proposed to improve the accuracy of object tracking tasks in varying illumination conditions. End to end deep architectures are proposed for anomaly detection and localization in crowded scenes [17]–[19], [28], and [29]. In these methods feature extraction as well as classification is carried out by deep networks. These networks are trained on huge data which is in the form of spatio-temporal volumes extracted from video datasets. On the other hand, methods like [20], [37] use pre-trained deep networks for extracting features directly from raw video frames. Event classification is accomplished with the help of traditional classifiers such as support vector machines (SVM). Method proposed by Elmannai and Al-Garni [35] is based on fusion of handcrafted features and features extracted by deep network. The authors have used traditional machine learning classifiers and majority voting approach for classification. Methods proposed by [21]–[24] do not extract features directly from raw video frames; instead they first extract motion information using optical flow/3D gradients of images and then use deep neural networks to extract high-level features from this primary motion information. In these methods event classification is done using supervised classifiers (like support vector machines/convolutional neural networks (CNN)) or unsupervised way such as autoencoders/decoders. Few researchers like [26] have proposed fusion of fully convolutional neural networks (FCNN) and optical flow features for detection of abnormal events. FCNN features are clustered into a set of binary codes. The variations in histogram of binary codes are compared with a statistical measure to detect abnormality.

Variations in illumination, poor resolution, small size and overlapping of objects are the major issues in crowd activity recognition. Many researchers have proposed different handcrafted features but extracting accurate and reliable features is challenging. Inaccurate features may degrade the accuracy of classification. Deep neural networks have ability to automatically extract features. However, to make the deep networks extract appropriate features, they must be first trained on large training data. For crowd abnormal event recognition systems, obtaining a big dataset is a big challenge. Even though many videos of crowded scenes are available, very few of them consists of sudden dispersion and merging activities. Moreover in the available videos, number of frames corresponding to dispersion or merging is very less. This is obvious because such abnormal situations do not occur frequently. Thus, ‘class imbalance’ is inevitable which adversely affects performance of event recognition systems.

We address the above mentioned issues in the proposed method. In this method, event recognition is accomplished using fusion of deep features with a novel set of handcrafted features. The proposed handcrafted features are not affected by occlusion and illumination. Moreover, these features are derived by using human intuition/intelligence to provide additional spatial and temporal information which can complement the deep features. In this way we assimilate the advantages of modern deep network approach and handcrafted features in order to improve the performance of crowd abnormal event recognition systems especially for small datasets having very few abnormal event instances.

Our approach of abnormal event recognition is much different than the methods mentioned in previous section. In the proposed method, convolutional neural networks are employed which extract spatial information from each frame. Additionally, we propose novel handcrafted features to capture temporal variations in the attributes of crowd such as motion magnitude and density. The features extracted by CNN are fused with the handcrafted features so that both spatial and temporal features will be available for event recognition (detailed explanation of the method in Section 2.3). The salient features of the proposed method are:

- Two sets of features are proposed in this work for abnormal event recognition. One is a set of deep features while another is a set of handcrafted features.
- Deep features are extracted by convolutional neural network–GoogleNet by transfer learning approach.
- We propose a completely new set of handcrafted features which represent spatial as well as temporal variations in crowd motion patterns. The proposed features are computed without tracking the objects. Also, they are not affected by occlusion and illumination and can be used for low to high crowd density levels.

- The fusion of deep and handcrafted features integrates advantages of both the approaches and thus leads to better discriminating feature vector.

Remaining paper is arranged as: The proposed method is discussed in detail in section 2, detailed explanation of datasets used for experimentation is given in Section 3 while in Section 4 analysis of results on different datasets is presented. Conclusion and future work are discussed in Section 5.

## 2. PROPOSED METHOD

### 2.1. Background

In the proposed method, the focus is on detection of sudden dispersion and merging events of people at crowded places. As the aim of proposed work is to distinguish merging and sudden dispersion events from normal situation; it is necessary to understand the definitions of merging, sudden dispersion and normal situation (Figure 1) before the proposed method is explained in detail.

- Normal Situation: A situation is said to be normal when either there is no significant movement (refer Figure 1(a): Sequence 1.1) or when people are walking coherently (as in Figure 1(b): Sequence 1.2) or they are freely roaming (refer Figure 1(c): Sequence 1.3). Thus, in normal situation, average speed of motion and density of people generally remains constant w.r.t. time.
- Merging Event: When people continuously gather at a certain location/region, we say it is merging event (refer Figure 2, Figure 2(a): Sequence 2.1 and Figure 2(b): Sequence 2.2). In this situation, number of people in that region (i.e. density) will go on increasing w.r.t time and may even increase beyond the capacity of that place.
- Sudden Dispersion Event: When people (who were initially comfortably moving) suddenly run away from each other, we define it as sudden dispersion event (refer Figure 3). In this situation, significant increase in motion magnitude and reduction in crowd density is observed.

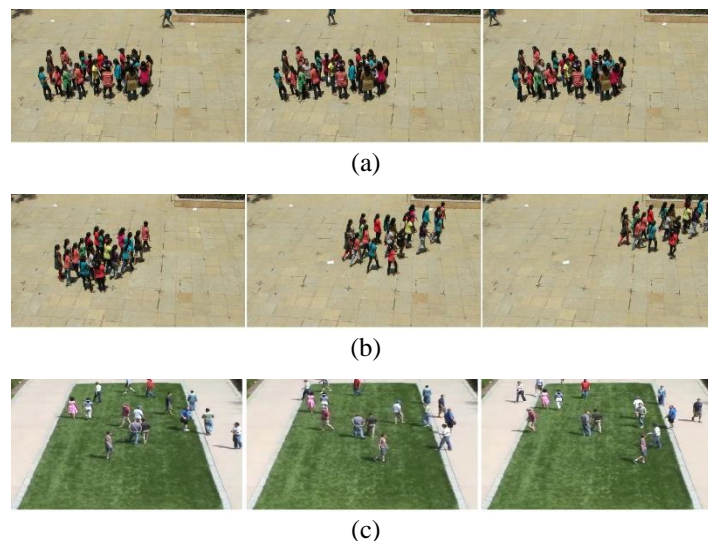


Figure 1. Examples of normal situation, (a) sequence 1.1: People standing in a group (no significant motion), (b) sequence 1.2: People walking coherently, and (c) sequence 1.3: People freely roaming

### 2.2. Block diagram

The block diagram of proposed method is shown in Figure 4. The video stream captured from CCTV is processed frame by frame. Every frame is processed simultaneously through the pipeline of deep feature extraction and handcrafted feature extraction. These features are then combined to form final feature vector which is used further for event classification. The proposed method is explained in detail in Section 2.3.

### 2.3. Deep feature extraction

Deep features are extracted from raw frames using pre-trained CNN. We have used standard 22 layer architecture of GoogLeNet [40] with 9 inception modules used for feature dimensionality reduction

without reducing the performance gain. It accepts input images of size  $224 \times 224 \times 3$ , so all the frames are resized to this size. The frame passes through multiple sets of convolutional and pooling layers and inception modules. The activation function used is rectified linear unit (ReLU)-rectified linear unit which helps to overcome the problem of vanishing gradients. We have used initial 10 layers of the network as it is i.e. their weights are not modified. However, GoogleNet is originally trained on ImageNet dataset for object recognition task. It is necessary to make it learn the features specific to crowd activity recognition task. Hence, fine tuning of 12 layers (present in 9 inception modules) is done by retraining them on crowd activity recognition datasets. The final feature map is acquired from the last average pooling layer which is a  $1024 \times 1$  dimension vector per frame. This feature map gives global/high level information of the input image.

As compared to other architectures of CNNs (such as LeNet, AlexNet, GoogleNet, and VGGNet), pre-trained GoogleNet model requires lesser memory [40]. It uses auxiliary classifiers to overcome the problem of overfitting which is typically experienced by very deep networks [40]. Moreover, it uses global average pooling layer at the end instead of fully connected layer due to which total learnable parameters are significantly reduced as well as accuracy is improved. Hence, GoogleNet is used in the proposed method.

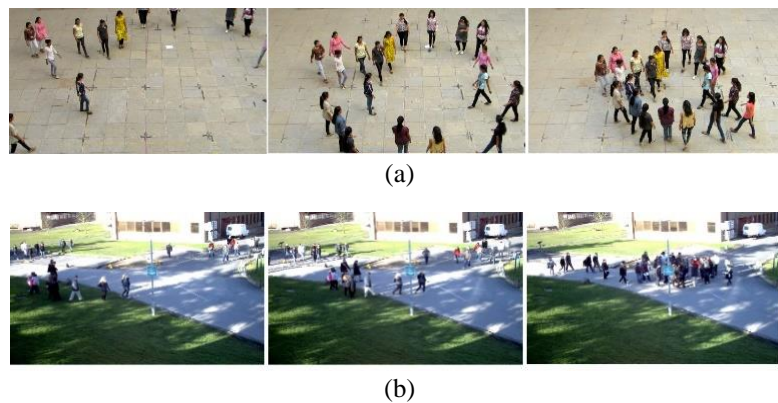


Figure 2. Examples of merging event, (a) sequence 2.1 and (b) sequence 2.2



Figure 3. Examples of dispersion situation

#### 2.4. Handcrafted features

For every incoming frame, a set of novel handcrafted features is extracted which provides spatial and temporal information about crowd. These features are extracted using macroscopic approach of crowd density and flow estimation. Detection and tracking of individual objects is completely avoided in the proposed method.

We propose following five features for event recognition: average distance from centroid of crowd, rate of change of average distance from centroid, rate of change of average density, average magnitude of motion and rate of change of magnitude of motion. Out of these five features, average distance from centroid of crowd, rate of change of average distance from centroid and rate of change of average crowd density are spatial features. These features are computed from foreground pixels and they represent distance between people and density of people. Remaining two are motion features which represent average speed of people. The process of 'handcrafted feature extraction' is described in this section.

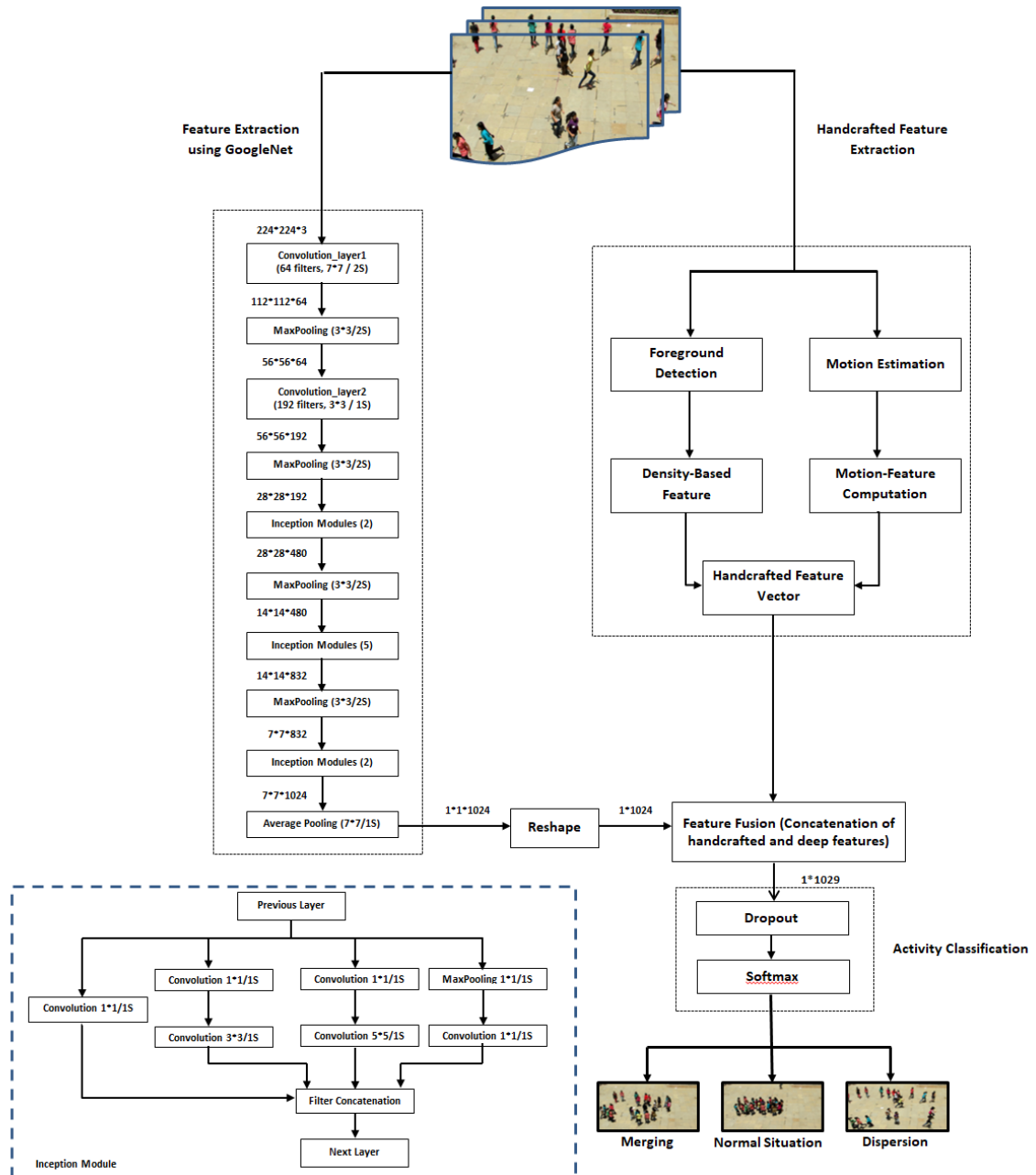


Figure 4. Block diagram of proposed method

**2.4.1. Spatial feature extraction**

Spatial features are extracted in two steps: foreground detection followed by feature computation;

- Foreground detection: In this step, moving objects are separated from background. In our application, we need to get a constant model of background so that correct foreground can be detected even if objects remain stationary for a longer time. Considering this, we have decided to use background subtraction approach. This is the simplest and the fastest technique which does not need any background model to be predefined. The frame without any moving objects is considered as the background frame. Following equations are used to obtain foreground (1) and (2).

$$imdiff = |I_{current} - I_{background}| \tag{1}$$

$$I_{foreground} = \begin{cases} 1, & imdiff \geq \text{threshold} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Threshold is computed using Otsu's algorithm [41]. In order to reduce the false positive pixels, median filtering is applied on thresholded image. The foreground image for sample frame is shown in Figure 5.



Figure 5. Foreground detection of sample frame of our new dataset

- Spatial feature computation: Using foreground pixels extracted in the previous step, various spatial features are computed. These features describe the distribution of moving objects in the scene, their density and distance between them. Following features are computed:

1. Average distance of foreground points from centroid of crowd: This feature describes the overall spread of people w.r.t. the centroid. Centroid of foreground region describes the center of the mass i.e. mean location around which the moving objects are distributed. It is observed that in case of merging situation, average distance of people from centroid keeps reducing; in case of dispersion situation it increases while in normal situation it does not change significantly. Centroid of foreground region is calculated as (3).

$$C(x, y) = \frac{\sum_{i=1}^n P_i(x, y)}{n} \quad (3)$$

where,  $n$  = total number of foreground pixels and  $P(x, y)$  is the spatial location of each foreground pixel  $P$ . Average distance from centroid is calculated using (4) and (5):

$$d(P_i) = \sqrt{(C(x) - P(x))^2 + (C(y) - P(y))^2} \quad (4)$$

$$d_{avg} = \frac{\sum_{i=1}^n d(P_i)}{n} \quad (5)$$

2. Rate of change of average distance: Variation in average distance from centroid is shown in Figure 6 for PETS dataset. We see that when people are merging, the average distance is reducing; when people are just standing in a group (i.e. normal situation), the average distance is constant and at its minimum level; while in case of dispersion situation, the distance from centroid keeps on increasing. From Fig. 6, it can be understood that monitoring the change in the average distance w.r.t. time is important to detect the type of event (i.e. merging or dispersion). Hence, we decided to consider rate of change of average distance w.r.t. time (computed over a window of 25 frames) as one of the distinctive features.
3. Rate of change of crowd density: This feature is estimated in three steps:
  - a. Estimation of absolute density around centroid: This feature estimates number of people around the centroid within certain distance. First, number of foreground points situated in a circular region with radius 'r' around the centroid 'C' is counted. This is defined as 'absolute density' (refer (6)).

$$\text{Absolute density} = \sum_{i=1}^n f_i \in \{P_i (d_i \leq r)\} \quad (6)$$

In order to make the method scene-independent, 'r' is not kept constant; but it is assigned a value of 'd<sub>avg</sub>' (computed using (5)). Depending on the distribution of people in the scene, 'd<sub>avg</sub>' changes (as demonstrated in Figure 6). In case of dispersion situation, average distance keeps increasing while in case of merging it reduces.

- b. Estimation of average density around centroid: For the prediction of crowdedness around centroid, average density is estimated using (7).

$$\text{Average density} = \frac{\text{Absolute density}}{d_{avg}} \quad (7)$$

- c. Estimation of Rate of change average density: Variation in average density is shown in Figure 7 for PETS dataset. It can be seen that, density in the region around the centroid keeps on increasing under merging situation. Under normal situation, the density does not change much, however when people start dispersing from each other, density goes on reducing. Thus, we consider 'rate of change of average density computed over a time period of 25 frames' as one of the features for event classification.

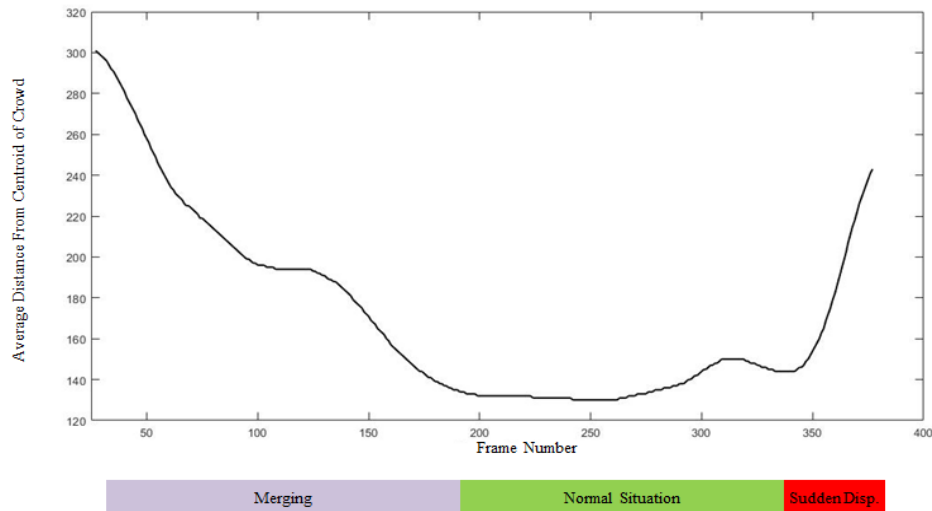


Figure 6. Variation in average distance from centroid w.r.t time

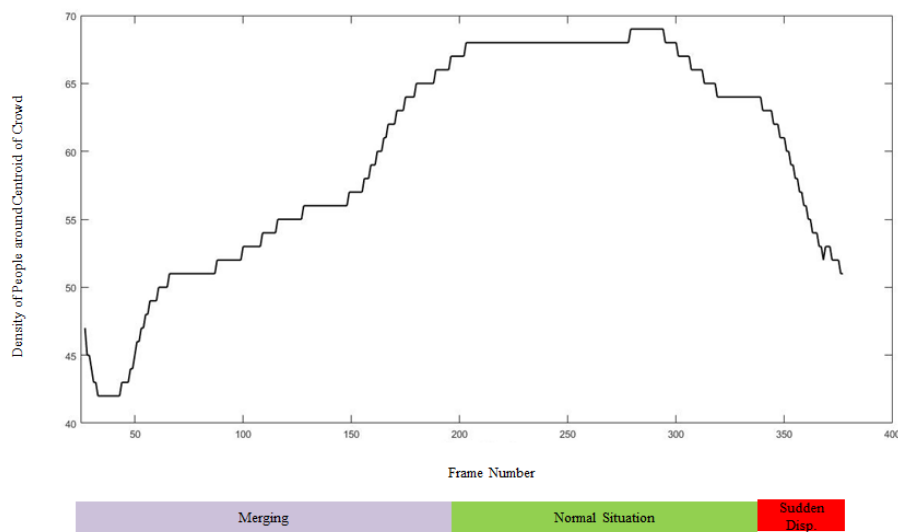


Figure 7. Variation in average density from centroid w.r.t time

#### 2.4.2. Motion feature extraction

Under normal situations, people move leisurely at their comfortable speed. When they experience any threat, they suddenly run away from each other. Thus the type of the flow is completely different under normal situation and sudden dispersion. Hence in addition to density based spatial features, motion based temporal features are also extracted. There are two steps in this process: global motion estimation and feature computation.

- Global motion estimation: It is accomplished using Horn and Schunk Optical Flow algorithm [42], [43]. The flow vectors at foreground pixels are only considered for further feature extraction step.

- Computation of motion based features: After acquiring the flow vectors, following two features are computed:
1. Average magnitude of motion is computed using (8).

$$V_{avg} = \frac{\sum_{i=1}^n V_i}{n} \tag{8}$$

The velocity magnitude  $V_i$  of each flow vector is calculated using (9).

$$V_i = \sqrt{V_{xi}^2 + V_{yi}^2} \tag{9}$$

$V_{xi}$  and  $V_{yi}$  are the vertical and horizontal components of flow vectors.

2. The rate of change of magnitude of motion is then computed over a window of 25 frames. The variation in average velocity magnitude is shown in Figure 8 for PETS dataset. Magnitude is seen increasing in the initial part where people are entering in the field of view, but after some time it starts reducing as people gather at some place. Under normal condition, when there is no significant motion, the magnitude is almost constant and is minimum; on the other hand when sudden dispersion occurs, the magnitude starts increasing rapidly. The rate of increase in magnitude is clearly visible in the last part of the curve. Similar variations in the features are observed on UMN dataset and our datasets. The other attribute of motion, i.e. the direction does not show any such discriminative nature (it shows similar distribution under merging and dispersion situations) and hence it is not considered in the proposed method.

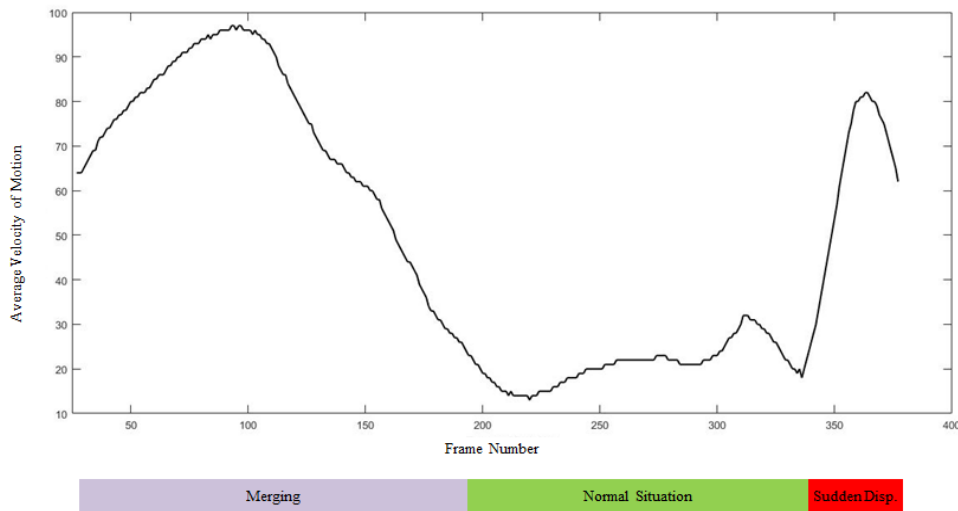


Figure 8. Variation in average magnitude of velocity w.r.t time

### 2.5. Fusion of deep features with handcrafted features

The handcrafted features are fused with the deep features. The five dimensional handcrafted feature vector is concatenated with 1024 dimensional feature vector obtained at the output of the last average pooling layer of pre-trained GoogleNet. Thus the final feature vector is of dimension 1029\*1. This feature vector is then used for recognizing the event happening in the scene.

### 2.6. Event classification

The final feature vector is applied to dropout layer before actual classification. This is required to avoid overfitting problem which is likely to occur when training datasets are small. During training phase, dropout layer temporarily removes some nodes in the network based on a probabilistic threshold. This process prevents the nodes to memorize the training data which in turn helps to prevent overfitting. The final classification is done by the softmax layer of GoogleNet. GoogleNet is originally trained on ImageNet dataset for object recognition task. Hence in the first phase, softmax layer is trained using the training set of feature vector  $S = \{s_1, s_2, \dots, s_N\}$  along with the corresponding labels  $L = \{l_1, l_2, \dots, l_N\}$  so that it learns features specific to the crowd activity recognition.. Here,  $N$  is the total number of training samples,  $s_i \in Z^d$ , is a feature vector of dimension  $d$  ( $d = 1029$ ) and  $l_i \in \{\text{normal, merging, sudden dispersion}\}$ . In the second phase



predictions are done on unseen test samples  $t_i \in Z^d$ . Fine tuning of hyper-parameters is also carried out in order to reduce classification loss. The hyper-parameters (such as learning rate, batch size and number of epochs) are selected empirically. We select minibatch size as 10, learning rate as 0.0001 (which is kept constant for all iterations) and number of epochs as 6.

### 3. RESEARCH METHOD

The proposed method is implemented on windows platform with Intel core 3 processor, 8 GB RAM and MATLAB 2019b toolbox. The method is evaluated on public as well as private datasets captured by us. Detailed information about these datasets is given below.

#### 3.1. Publically available datasets

For crowded scene analysis, various datasets are available such as University of California San Diego (UCSD) Pedestrian Dataset (PED1 and PED2), Avenue, University of Minnesota (UMN) dataset, Performance Evaluation of Tracing and Surveillance (PETS), and Chienese University of Hong Kong (CUHK) dataset. Out of these, the merging and sudden dispersion situations are found only in UMN and PETS datasets. Hence we considered PETS and UMN datasets for evaluating our method.

PETS 2009 (S3) dataset [44] consists of four videos of outdoor scene each recorded from four different view-points. Each scene is of 1 min duration approximately. There are total 377 frames in each video. Frame rate is 7fps and resolution is  $576 * 768$ . UMN dataset consists of three scenes captured in different environments. Out of these, two are outdoor scenes and one is indoor scene. The total duration of the video is 4 min 17 sec. The frame rate is 23 fps and resolution is  $240 * 320$ . There is wide variation in illumination conditions and distance of objects from the camera in these scenes.

#### 3.2. New dataset

New dataset is created by acquiring CCTV footages at Cummins College campus. It consists of two video clips which demonstrate both merging and dispersion situations. These are outdoor scenes captured from same viewpoint but on different days and under different illumination conditions. First video clip consists of 1100 frames and the second clip consists of 1133 frames. The videos are captured at frame rate of 25 fps with resolution of  $1080 * 1920$ . Again, every frame is observed manually and then event in the frame is labeled as 'merging event' or 'sudden dispersion event'. Table 1 explains the summary of each dataset in brief. It can be observed from table I that in all datasets number of frames depicting sudden dispersion and merging event are very less as compared to normal frames (i.e. datasets are unbalanced).

Table 1. Information of different datasets used for performance evaluation

Dataset	# Total Frames	# Sudden Dispersion Frames	# Merging Frames	# Normal Frames
PETS 14to33	377	42	190	145
UMN Scene 1	1525	383	--	1142
UMN Scene 2	3728	1158	--	2570
UMN Scene 3	2070	239	--	1831
New dataset Scene 1	1100	54	263	783
New dataset Scene 2	1133	119	211	803

## 4. RESULTS AND DISCUSSION

### 4.1. Evaluation parameters

The performance of our method is quantified in terms of various quantitative parameters such as accuracy, precision, recall and F-measure. These parameters are calculated (10) to (13).

$$\% \text{ Accuracy} = \frac{TP+TN}{\text{Total Frames}} \times 100 \quad (10)$$

$$\% \text{ Precision} = \frac{TP}{TP+FP} \times 100 \quad (11)$$

$$\% \text{ Recall} = \frac{TP}{TP+FN} \times 100 \quad (12)$$

$$\% \text{ F measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \times 100 \quad (13)$$

TP: True Positive frames; TN: True Negative frames; FP: False Positive frames; FN: False Negative frames

#### 4.2. Experimental results

In this section, we present the results of the experimentation carried out to test the effectiveness of the proposed method. It is important to note that, to test efficacy of the method even with small datasets having large data imbalance, we tested our method without data augmentation or class balancing technique. The classifier is trained on few frames of each dataset and tested on the remaining frames of the same dataset. The ratio of training-testing frames is kept as 60:40.

We compare the performance of abnormal event recognition using only deep neural networks with the proposed feature fusion method (refer Table 2) on different datasets. The performance parameters are calculated for each class separately. We can see that, F-measure and accuracy values are significantly improved for recognizing dispersion and merging activities using the proposed feature fusion approach. This proves the effectiveness of proposed handcrafted features in recognizing the events on the datasets with high class imbalance.

Table 2. Results of abnormal event recognition using only CNN features and Proposed Method: CNN and handcrafted feature fusion

Dataset	Method	Precision			Recall			F-measure			Accuracy
		Dispersion	Normal	Merge	Dispersion	Normal	Merge	Dispersion	Normal	Merge	
PETS	only CNN	100	93.5	100	76.5	100	100	86.68	96.64	100	97.2
	Proposed Method	100	95.1	100	82.4	100	100	90.35	97.48	100	97.9
MVI 7	only CNN	100	99.3	92.9	84	100	96.3	91.30	99.64	94.56	98.43
	Proposed Method	100	98.7	100	100	100	92.6	100	99.34	96.16	99
MVI 11	only CNN	97.9	100	91.7	100	96.8	100	98.93	98.37	95.67	97.9
	Proposed Method	100	100	100	100	100	100	100	100	100	100
UMN	only CNN	99.2	99.68	-	98.54	99.92	-	98.86	99.79	-	99.68
	Proposed Method	100	99.9	-	99.38	100	-	99.68	99.94	-	99.92

#### 4.3. Comparison with other state of the art methods

The proposed method is compared with other methods w.r.t. accuracy, F-measure and Area Under Receiver Operating Characteristics (AUC) on standard publically available UMN dataset. The comparison is presented in Table 3. It can be seen that accuracy, F measure and AUC of the proposed method are better than other methods. This means that, as compared to other methods, the proposed method can better distinguish between normal and abnormal situations. Thus, the proposed method achieves the best performance amongst all methods.

Table 3. Comparison of proposed method with other methods on UMN dataset

	Proposed	Abdullah <i>et al.</i> [14]	Wu <i>et al.</i> [7]	Guo <i>et al.</i> [12]	Fang <i>et al.</i> [21]	Smeureanu <i>et al.</i> [19]	Zhou <i>et al.</i> [16]	Ravanbaksh <i>et al.</i> [25]	Direkglu [24]
Features used	Handcrafted+ Deep	Handcrafted	Handcrafted	Handcrafted	Extracted by deep network	Extracted by deep network	Extracted by deep network	Feature Fusion (OF + AlexNet)	MII + Deep
Training: Testing Ratio	60:40	--	Not specified	--	Not specified	80:20	Not specified	---	Not Specified
Accuracy	99.92	87	99.03	NA*	NA*	NA*	NA*	98.8	99.08
F measure	99.68	NA*	NA*	NA*	98.81	NA*	NA*	NA	NA
AUC	99.94	NA*	NA*	NA*	NA	97.1	99.63	98.8	NA

## 5. CONCLUSION

In this paper, we have proposed a method for recognizing two unusual events-sudden dispersion and merging in crowded scenes. The proposed method assimilates benefits of deep networks and handcrafted features for abnormal event recognition. In the proposed method, we have employed convolutional neural network to extract higher level features from video frames. In addition, we propose a novel set of low level features to capture temporal variations in crowd density and motion patterns. Feature vector generated by fusion of handcrafted features and features extracted by CNN is used for event recognition. The method is validated on benchmark datasets as well as our private datasets. Experimental results show that our method successfully recognized abnormal events from small datasets having very few abnormal event instances. The

results obtained by fusion of deep and handcrafted features are more accurate than that of only deep features. This proves that the proposed set of handcrafted features provides additional information. Thus handcrafted features complement the features extracted by CNN and their fusion improves the performance of event recognition by significant amount. The method is also proved to be more effective as compared to other state of the art methods.




## REFERENCES

- [1] M. Bendali-Braham, J. Weber, G. Forestier, L. Idoumghar, and P.-A. Muller, "Recent trends in crowd analysis: A review," *Machine Learning with Applications*, vol. 4, p. 100023, Jun. 2021, doi: 10.1016/j.mlwa.2021.100023.
- [2] G. Tripathi, K. Singh, and D. K. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: a survey," *The Visual Computer*, vol. 35, no. 5, pp. 753–776, May 2019, doi: 10.1007/s00371-018-1499-5.
- [3] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, Mar. 2015, doi: 10.1109/TCSVT.2014.2358029.
- [4] K. Rangasamy, M. A. As'ari, N. A. Rahmad, N. F. Ghazali, and S. Ismail, "Deep learning in sport video analysis: A review," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 4, pp. 1926–1933, Aug. 2020, doi: 10.12928/TELKOMNIKA.V18I4.14730.
- [5] Y. Benabbas, N. Ihaddadene, and C. Djeraba, "Motion Pattern extraction and event detection for automatic visual surveillance," *EURASIP Journal on Image and Video Processing*, vol. 2011, pp. 1–15, 2011, doi: 10.1155/2011/163682.
- [6] Guohui Li, Jun Chen, Boliang Sun, and Haozhe Liang, "Crowd event detection based on motion vector intersection points," in *2012 International Conference on Computer Science and Information Processing (CSIP)*, Aug. 2012, pp. 411–415, doi: 10.1109/CSIP.2012.6308881.
- [7] Y. Cong, J. Yuan, and Y. Tang, "Video anomaly search in crowded scenes via spatio-temporal motion context," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pp. 1590–1599, Oct. 2013, doi: 10.1109/TIFS.2013.2272243.
- [8] S. Wu, H. S. Wong, and Z. Yu, "A bayesian model for crowd escape behavior detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 85–98, Jan. 2014, doi: 10.1109/TCSVT.2013.2276151.
- [9] A. S. Rao, J. Gubbi, S. Marusic, and M. Palaniswami, "Crowd event detection on optical flow manifolds," *IEEE Transactions on Cybernetics*, vol. 46, no. 7, pp. 1524–1537, Jul. 2016, doi: 10.1109/TCYB.2015.2451136.
- [10] H. Mousavi, M. Nabi, H. Kiani, A. Perina, and V. Murino, "Crowd motion monitoring using tracklet-based commotion measure," in *Proceedings - International Conference on Image Processing, ICIP*, Sep. 2015, vol. 2015-December, pp. 2354–2358, doi: 10.1109/ICIP.2015.7351223.
- [11] J. Wang and Z. Xu, "Spatio-temporal texture modelling for real-time crowd anomaly detection," *Computer Vision and Image Understanding*, vol. 144, pp. 177–187, Mar. 2016, doi: 10.1016/j.cviu.2015.08.010.
- [12] Q. Wang, Q. Ma, C. H. Luo, H. Y. Liu, and C. L. Zhang, "Hybrid histogram of oriented optical flow for abnormal behavior detection in crowd scenes," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 30, no. 2, p. 1655007, Feb. 2016, doi: 10.1142/S0218001416550077.
- [13] C. Guo, H. Lin, Z. He, X. Shu, and X. Zhang, "Crowd abnormal event detection based on sparse coding," *International Journal of Humanoid Robotics*, vol. 16, no. 4, p. 1941005, Aug. 2019, doi: 10.1142/S0219843619410056.
- [14] M. Pathade and M. Khambete, "Unsupervised detection of dispersion and merging activities for crowded scenes," in *Advances in Intelligent Systems and Computing*, vol. 1082, 2020, pp. 595–604.
- [15] F. Abdullah, Y. Y. Ghadi, M. Gochoo, A. Jalal, and K. Kim, "Multi-person tracking and crowd behavior detection via particles gradient motion descriptor and improved entropy classifier," *Entropy*, vol. 23, no. 5, p. 628, May 2021, doi: 10.3390/e23050628.
- [16] K. Schairi, C. Benbouchama, K. El Houari, and C. Fatima, "A real-time implementation of moving object action recognition system based on motion analysis," *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, vol. 5, no. 1, pp. 44–58, Mar. 2017, doi: 10.52549/ijeie.v5i1.261.
- [17] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing: Image Communication*, vol. 47, pp. 358–368, Sep. 2016, doi: 10.1016/j.image.2016.06.007.
- [18] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017, doi: 10.1109/TIP.2017.2670780.
- [19] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "ResNetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2017, pp. 1–7, doi: 10.1109/AVSS.2017.8078482.
- [20] S. Smeureanu, R. T. Ionescu, M. Popescu, and B. Alexe, "Deep appearance features for abnormal behavior detection in video," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10485 LNCS, 2017, pp. 779–789.
- [21] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning Deep representations of appearance and motion for anomalous event detection," in *Proceedings of the British Machine Vision Conference 2015*, 2015, pp. 8.1-8.12, doi: 10.5244/C.29.8.
- [22] Z. Fang *et al.*, "Abnormal event detection in crowded scenes based on deep learning," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 14617–14639, Nov. 2016, doi: 10.1007/s11042-016-3316-3.
- [23] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, Jan. 2017, doi: 10.1016/j.neucom.2016.09.063.
- [24] M. Gutoski, N. M. R. Aquino, M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes, "Detection of video anomalies using convolutional autoencoders and one-class support vector machines," in *Proceeding XIII Brazilian Congress on Computational Inteligence*, Jan. 2019, pp. 1–12, doi: 10.21528/cbic2017-49.
- [25] C. Direkoglu, "Abnormal crowd behavior detection using motion information images and convolutional neural networks," *IEEE Access*, vol. 8, pp. 80408–80416, 2020, doi: 10.1109/ACCESS.2020.2990355.
- [26] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sanginetto, and N. Sebe, "Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection," in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, Mar. 2018, vol. 2018-January, pp. 1689–1698, doi: 10.1109/WACV.2018.00188.




- [27] A. J. Abdulelah, M. Al-Kubaisi, and A. M. Shentaf, "An efficient human activity recognition model based on deep learning approaches," *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, vol. 10, no. 1, pp. 177–186, Mar. 2022, doi: 10.52549/ijeie.v10i1.3438.
- [28] M. A. Alsaedi, A. S. Mohialdeen, and B. M. Albaker, "Development of 3D convolutional neural network to recognize human activities using moderate computation machine," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3137–3146, Dec. 2021, doi: 10.11591/eei.v10i6.2802.
- [29] S. Sharma, B. Sudharsan, S. Narahariseti, V. Trehan, and K. Jayavel, "A fully integrated violence detection system using CNN and LSTM," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 3374–3380, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3374-3380.
- [30] A. Al Mamun, P. P. Em, and J. Hossen, "An efficient encode-decode deep learning network for lane markings instant segmentation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, p. 4982, Dec. 2021, doi: 10.11591/ijece.v11i6.pp4982-4990.
- [31] A. Abozeid, H. Farouk, and S. Mashali, "Depth-DensePose: An efficient densely connected deep learning model for camera-based localization," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 3, pp. 2792–2801, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2792-2801.
- [32] M. Masadeh, A. Masadeh, O. Alshorman, F. H. Khasawneh, and M. A. Masadeh, "An efficient machine learning-based COVID-19 identification utilizing chest X-ray images," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, p. 356, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp356-366.
- [33] T. Mathu and K. Raimond, "A novel deep learning architecture for drug named entity recognition," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 19, no. 6, pp. 1884–1891, Dec. 2021, doi: 10.12928/TELKOMNIKA.v19i6.21667.
- [34] A. M. Alkababji and O. H. Mohammed, "Real time ear recognition using deep learning," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 19, no. 2, pp. 523–530, Apr. 2021, doi: 10.12928/TELKOMNIKA.v19i2.18322.
- [35] H. Elmannai and A. D. Al-Garni, "Classification using semantic feature and machine learning: Land-use case application," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 19, no. 4, pp. 1242–1250, Aug. 2021, doi: 10.12928/TELKOMNIKA.v19i4.18359.
- [36] I. Ahmed, C. Der, N. Jamil, M. A. Mohamed, "Improve of contrast-distorted image quality assessment based on convolutional neural networks," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 6, pp. 5604–5614, doi:10.11591/ijece.v9i6.pp5604-5614.
- [37] Z. Kadim, M. A. Zulkifley, and N. A. M. Kamari, "Training configuration analysis of a convolutional neural network object tracker for night surveillance application," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 282–289, Jun. 2020, doi: 10.11591/ijai.v9.i2.pp282-289.
- [38] G. A. Shadeed, M. A. Tawfeeq, and S. M. Mahmoud, "Deep learning model for thorax diseases detection," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 1, pp. 441–449, Feb. 2020, doi: 10.12928/TELKOMNIKA.v18i1.12997.
- [39] H. Prasetyo and B. A. Putra Akardihas, "Batik image retrieval using convolutional neural network," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 6, pp. 3010–3018, Dec. 2019, doi: 10.12928/TELKOMNIKA.v17i6.12701.
- [40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, et al., "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 40, no. 1, pp. 1–9.
- [41] N. Otsu, "Threshold Selection Method From Gray-Level Histograms.," *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979, doi: 10.1109/tsmc.1979.4310076.
- [42] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203, Aug. 1981, doi: 10.1016/0004-3702(81)90024-2.
- [43] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, Feb. 1994, doi: 10.1007/BF01420984.
- [44] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Dec. 2009, pp. 1–6, doi: 10.1109/PETS-WINTER.2009.5399556.

## BIOGRAPHIES OF AUTHORS



**Manasi Pathade**    is an assistant professor in Cummins College of Engineering, Pune, India. She received her M.E. degree from the department of E&TC, University of Pune, India. Currently, she is pursuing Ph.D. in computer vision domain under the guidance of Dr. Madhuri Khambete. She can be contacted at manasi.pathade@cumminscollege.in.



**Dr. Madhuri Khambete**    received her Ph.D. from the department of E&TC at College of Engineering, Pune, India. Currently, she works as the Principal in Cummins College of Engineering for Women, Pune, India. Her research interests are computer vision, image processing and pattern recognition. She can be contacted at madhuri.khambete@cumminscollege.in.