International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Analyzing and Predicting Crowd Behavior using Machine Learning

Jignesh Haren Kumar Vaniya[*1], Nikunj Chunilal Gamit[2], Naimisha Shashikant Bhai Trivedi[3], Chetna Ganesh Chand[4], Dhaval Varia[5]

Submitted:10/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: Training a model of crowd behaviour using data taken from video sequences is essential in crowd behaviour comprehension. Most of the existing approaches rely only on low-level visual characteristics since crowd datasets do not include any ground-truth other than the crowd behaviour labels. But the conceptual gap between basic motion/appearance characteristics and the abstract idea of crowd behaviour is enormous. We provide an attribute-based approach to address this issue in this study. We believe our work is the first to demonstrate that crowd emotions may be used as features for understanding crowd behaviour, even if comparable approaches have been utilised for object and action detection in recent times. To reflect the movement of the crowd, the primary goal is to train a classifier set that is based on emotions. To that end, we amass a large collection of video clips and annotate them with "crowd behaviours" and "crowd emotions" tags. We provide the outcomes of the suggested approach on our dataset, which show that crowd emotions allow for the development of richer descriptions of crowd behaviours. Along with the publication, we want to share the dataset so that communities may utilise it as a standard.

Keywords: *Communities, Crowd, Behaviours, Emotions, Low-Level, Labels, Motion*

1. Introduction

Terrorism and criminal activity have both seen sharp increases in recent years. The use of video surveillance has grown in importance as a means of crime and violence prevention, particularly in large public gatherings. The last few years have seen a meteoric rise in the number of surveillance cameras set up in homes and businesses alike. [1] However, although technological advancements have enabled video quality to reach new heights, they have also increased processing demands. It becomes clear that computerised processing is necessary since manual analysis of large data quantities is not feasible.

An emerging area of study that is capturing the interest of many people in this setting is automated video-surveillance. There has been a lot of research into areas like facial recognition and identification and harmful item detection. [2] There has been a recent uptick in interest in automated video-surveillance systems, which can analyse crowd behaviour. Learning how people conduct in big groups is a primary goal of this field, as is deriving useful information from footage including such scenes. [3]actions of a big number of spectators at a sporting event. [4]latest technology The Deep Learning models, which include Deep Neural Networks (DNNs), have shown remarkable performance in many computer vision tasks and time series analysis. Using Deep Learning to analyse crowd behaviour is a new trend since these models work well with video sources that can be automatically analysed. [5,6]Deep features, which are one level of interaction between the characteristics obtained from any subsystem and applied via fusion functions to interpret the findings, are used in traditional methods to computer vision problems. Over time, hardware's computing capability has grown while its price has decreased dramatically. We have also seen the development of models that autonomously learn from data. [7] Feature vectors were the targets of this system's training. To find the generalised optimum solution in a high-dimensional feature space, structures are essential; to build such a system, we need the discriminant features of the given system. You can use these discriminant qualities to do a lot of operations on feature sets.[8] Then, you can use machine learning techniques or the old-fashioned way to fix the problems. A crucial part of the answer in both cases is identifying the unique features of the problems. Elements that are made by hand are called handmade elements. [9]Teaching computers these discriminant qualities with detailed issue codes are the most recent advancement and the fundamental

[1]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0009-0002-4121-0074
[2]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0009-0004-2821-4606
[3]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0000-0002-8395-1586
[4]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0009-0003-1876-8472
[5]*Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad*
ORCID ID: 0000-0003-4505-1509
** Corresponding Author Email: jignesh.apit@gmail.com*

component of optimum answers. In order to discover a solution, several deep learning approaches are used, making advantage of these attributes. [10] The use of feature vectors is fundamental to many deep learning algorithms. Deep learning is a relatively recent concept in the field of machine learning [11,12]. The ideas behind deep learning emerged from the breadth of machine learning inside the umbrella of AI [13]. Representation learning is a more recent kind of deep learning that has a hierarchical structure with layers. Information is sent up from lower layers to higher ones. Many sources [14] discuss three distinct kinds of architectures: generic, discriminative, and hybrid. Many consider convolutional neural networks (CNNs) to be the best model for image analysis. Convolutional neural networks (CNNs) include many layers that modify input using convolution filters with lower dimensions. In recent years, deep networks have been widely used in computer vision. [15]The most remarkable emerging technologies, like deep learning, may change the way people learn in every aspect of their life. The demonstration that DNNs could be trained in an unsupervised fashion (relating) layer by layer, with subsequent supervised fine-tuning of the stacked network, was a huge step forward. As a denoising approach, an autoencoder has been proposed, where the model is taught to reconstruct the input from a version distorted by noise. [16] Deep belief networks (DBNs) are one kind of stack-based autoencoder that uses unsupervised learning for its layers. Many researchers were intrigued by Hubel and Wiesel's [17] method of visual perception in animals and its layered neuronal architecture. This piqued their attention, and they set out to develop similar pattern recognition algorithms in computer vision. Some other interesting applications include medical research leading to the discovery of new drugs and outperforming human opponents in video games. [18]Organic compound in the realm of computer vision, cross ventilation refers to a tweaked form of convolution that is used whenever a CNN operation is performed on selected image inputs. While both AlexNet and LeNet-5 provide comparable designs, AlexNet's architecture is deeper since it uses a different number of layers. Deep learning's popularity has been on the rise in recent years. A continuous advancement in deep learning has been caused by the arrival of inexpensive and widely used GPU.[19] In comparison to central processing units (CPUs), graphics processing units (GPUs) are very fast. Therefore, learning on deep architectural components may be much faster than learning on CPUs. A further element contributing to the utilisation of GPU-based technology is the availability of complimentary deep learning software. [20] Effective neural network operations are provided by these libraries, including weight change, linear unit correction, pooling, and bias additions. Therefore, it is more prudent to create the top convolutional layers appropriately and utilise them to notice the flaws rather than rewriting the complete convolutional network.[21]

2. Review of literature

Huang et al. (2018) [22] suggest use Convolutional Restricted Boltzmann Machines to extract a set of characteristics. Visual patches, energy patches, and motion patches were data collected using three separate models: one from the original frames, one from feature maps derived by applying Gaussian filters to input patches, and one from Optical Flow calculations. Feature extraction was followed by feeding all characteristics into a one-class support vector machine (SVM) that was trained to recognise the typical pattern.

Yang M(2019) [23]the authors presented a two-stage paradigm. First, in order to extract semantic information from scene objects, a supervised version of Fast R-CNN was trained using large-scale datasets; this version was tailored for multitask learning. The class, action, and characteristics of each item were detailed. Phase two involves training the generic semantic extractor. Following this, an anomaly detector is brought in to learn the unique normal pattern for each dataset. In inference time, it reports an anomaly score for each piece of information. After trying out a number of different anomaly detectors, one-class SVM proved to be the most effective.

Girshick (2023) [24] suggested using two Stacked DAEs (SDAEs) as an ensemble approach. Using the Kanade-Lucas-Tomasi description, the first SDAE was fed the original video, and the second one the foreground sequence. To feed the SDAE, a motion map was computed for each input. For both models, a one-class support vector machine (SVM) was used for final classification after feature extraction and dimensionality reduction using a Deep Belief Network. After that, the total anomaly score was determined by adding the two outputs linearly. We choose to put this pipeline in this section since the dataset used to evaluate anomalous action detection is not publically accessible, even though the authors use it both for motion and action anomaly detection.

Fang et al. (2022) [25] trained a deep network named PCANet to do Cascaded Principal Component Analysis on the input data using two traditional features: saliency maps and Multi-scale Histograms of Optical Flow (MHOF). A normal pattern was learnt using a one-class SVM after data reduction. This detection model can run at approximately 20 FPS, making it appropriate for real-time applications, and achieves performance close to

deeper models. It does this by using a pretrained VGG-f network as a rapid feature extractor and then training an SVM over the retrieved features. Similarly, integrated a CNN with a one-class SVM layer and trained the whole model from beginning to finish.

3. Objectives

- To examine the attribute-based strategy that uses crowd emotions as attributes to connect the semantic gap between the low-level visual features and the high-level crowd behavior concepts.
- To gather a big dataset of video clips with both crowd behaviors and crowd emotions annotated, and then release this dataset as a benchmark for crowd behavior understanding research.

4.Statement of the problem

In crowd behavior analysis, most of the existing methods depend only on the low-level visual features that are extracted from the video sequences. Nevertheless, there is a big semantic gap between these low-level features and the high-level concepts of crowd behaviors that has to be modelled. Thus, it is difficult to accurately recognize and understand the crowd behaviors using only low-level visual cues. Moreover, there is a shortage of big datasets with both the abnormal crowd behaviors and the crowd emotion information, which could give the more descriptive representations to connect this semantic gap. New methods that can make use of the emotional context and the large annotated datasets are required for the improvement of crowd behavior analysis.

5. Significance of the study

This study is important for many reasons. Initially, it provides a new large-scale dataset with annotations for both the abnormal crowd behaviors and the crowd emotions. The availability of this large dataset annotated with emotional context in addition to behavior labels creates a new benchmark for the computer vision community. The study also suggests an attribute-based strategy that uses crowd emotions as mid-level attributes to fill the semantic gap between low-level visual features and high-level crowd behaviour concepts. This emotional attribute representation is the basis for the building of more descriptive models for crowd behaviour understanding. Through the combined use of crowd behaviour and emotion recognition, the proposed method surpasses the baseline approaches that only depend on the low-level visual features. In general, this work provides new research paths for using emotional context to enhance crowd behaviour analysis.

6. Research methodology

Typical behaviour identification algorithms in laboratories still fail miserably when it comes to comprehending crowd behaviours in the real world, despite the high need for such knowledge. The main one is that most algorithms that have been suggested have only been tested on non-standard datasets that include a small number of sequences collected under controlled conditions and have a restricted number of behaviour classes.

An essential feature of any dataset is the number of samples it contains. Having a large enough collection of recorded videos is useful for assessment and training with more samples. Among the many significant dataset criteria is the annotation level. It technically represents the annotation richness of a dataset and may be expressed at three different levels: pixel-level, frame-level, and video-level. “There is also the matter of crowd density to think about. Scenes with a large number of people make it more difficult to distinguish between various types of behaviours due to increased occlusion and clutter.

Another important aspect of a dataset is the kind of scenarios, which represents the events that occur in each video sequence. More complex datasets provide more difficulties, since the suggested algorithms need to be able to handle a wider range of situations (i.e., the actual world). Lighting, background clutter, occlusions, and other factors might be significantly affected by the Indoor/Outdoor criterion, which pertains to the area where the video sequences were captured.

We stress the importance of meta-data in this work since it is a crucial component of every dataset. Additionally, it offers researchers the chance to go towards more abstract interpretations of the video sequences, which is one of the things that makes our dataset special. As an additional annotation, we included "crowd emotion" into our dataset. We detail all of the aforementioned crowd behaviour datasets according to the qualities that were discussed. The absence of further annotation is a common deficiency across all of them, limiting their use to low-level characteristics for behaviour class discrimination. Prior datasets also suffered from a lack of variety in terms of both participants and situations, as well as a low volume crowd density and a small number of video sequences.

- **Potential Dataset**

A total of 35 video sequences, including around 40,000 normal and aberrant video clips, make up the dataset that is being presented. The movies, which have a resolution of 554 × 235, were shot at a rate of 30 frames per second using a stationary video camera that was positioned at a height to capture each walkway. The crowd's density fluctuated, going from very sparse to quite packed. In order to make the situations more realistic, we capture

both typical and odd behaviour, as well as a few instances involving strange things that may be seen as dangers to the audience. Illustrations of this kind of situation include "a suspicious backpack left by an individual in the crowd," "a motorbike which is left between many people," "a motorbike crossing the crowded scene," and many more.

Our suggested dataset includes five common kinds of crowd behaviour. We drew each scenario topology with the conditions often seen in crowding problems in mind. They line up with a neutral scene with people moving freely, an obstacle scene with people clumped together and strange things, a panic scenario with people running away, a fight scene with people becoming violent and finally, a congested scene with people swarming together. We captured footage from two distinct angles for every behaviour category, varying the crowd densities from very sparse to very dense. Frames of normal behaviour begin each video in the collection, whereas frames of deviant behaviour conclude each one. The commonly accepted psychological definition of emotion, "a feeling evoked by environmental stimuli or by internal body states," is what we use for emotion annotation.

In terms of environmental activities or changes in the body's internal state, this might influence human behaviour. Based on the face and gesture datasets, we constructed six categories of fundamental emotions in our dataset: "Angry," "Happy," "Excited," "Scared," "Sad," and "Neutral." Since there is no existing crowd anomalous behaviour dataset with emotion annotation in the computer vision research, we continued with these categories. Given the inherent subjectivity in perceiving any of these emotions, we have many annotators provide their input independently before reaching a consensus by majority vote. We performed agreement research to guarantee annotators' consistency and found a total agreement of 92% with a Kappa value of 0.81.

The highest discrepancy was seen between Happy and Excited, with confused cases occurring 4% of the time. Emotion annotations, as anticipated, may provide a more illustrative depiction of crowd behaviour, as is evident simply from the annotations. A person's or a group's emotional state may significantly impact their decision-making in many contexts. Compared to the raw, low-level motion data of films, emotional data is far more robust. A public release of the films, ground-truth annotations, and baseline scripts is imminent after the paper's publication. We anticipate that this dataset will serve as a standard for studies pertaining to the identification of deviant behaviour and the identification of emotions in the future.

- **Crowd Representation Based on Emotions**

We can solve a regular classification issue using the ground-truth emotion labels as input if they are accessible throughout testing and training as well. However, without a ground-truth emotion name to refer to, things get more complicated during testing. We provide a formal explanation of the latter in this section.

With a dataset consisting of N video clips labelled as {(x(n), e(n), y(n))}N, the objective is to train a model that uses emotion labels e to attribute the class label y to a test video clip x that has not been seen before. The d-dimensional low-level feature f ∈Fd that is taken from video clip x is represented as a tuple (f, e, y) during the training phase. The image's aberrant behaviour class label is denoted by y ∂ Y. A K-dimensional vector e = (e_1, e_2,..., e_k) represents the crowd emotion of the video clip x, where ek ∈ Ek (k = 1, 2,..., K) signifies the k-th emotion of the video clip. Take "Angry" as an example of the k-th emotion attribute; if it's 1, then the audience is "Angry," and if it's 0, then it's not. For the sake of our datasets' compatibility with conventional multi-label emotion detection systems, we assign a single non-zero value to each video clip's emotion attribute, denoted as Ek = {0, 1} (e = 1, 2,..., K), so that e 0 = 1. However, it is important to note that our suggested approach may easily be expanded to handle continuous valued (fuzzy) qualities as well as those with several emotions, not only those with binary values. We may skip training a classifier C :Fd → Y that uses emotion information and instead just transfers the feature vector f to a behaviour class label y. Instead, the classifier C is broken down into :

$$\mathcal{H} = \mathcal{B}(\mathcal{E}(f))$$
$$\mathcal{E}: \mathbb{F}^d \to \mathbb{E}_k \text{ and } \mathcal{B}: \mathbb{E}_k \to \mathbb{Y}$$

In this case, E involves K distinct emotion classifiers, and for each i-th emotion in E, each classifier C_{eimaps} f to a behaviour class label y from Y that corresponds to an emotion attribute En. As part of the training process, our dataset provides emotion annotations that the emotion classifiers use to learn. C_{ei} (f) is a binary linear support vector machine (SVM) classifier that has been trained to identify positive examples from all behaviour classes when ei = 1 and negative examples from other classes. We denote each video clip x by χ (x) ∈ Ek, assuming there is no emotion ground-truth information accessible during test time:

$$\phi(x) = [s_1(x), s_2(x), \ldots, s_K(x)]$$

in which the confidence score of the k-th emotion classifier is denoted as sk(x). To E., Cek. The presence

or absence of an emotion in a video clip may be captured by looking at this crowd representation vector, which is based on emotions. The last step in obtaining the mapping B is training a multi-class linear support vector machine (SVM) for behaviour classes using vectors that indicate crowd emotions.

Latent Emotion: In order to extract the intra-class change of each behaviour, we use crowd emotions as characteristics, which are discriminative. While videos may seem to be part of the same behaviour class, intra-class changes may lead to their being linked to various sets of emotional information. As an example, the behaviour class Congestion may have the Angry emotion attribute in certain video clips of a dataset and the Happy emotion attribute in others. We solve this issue by learning the model with the help of the latent support vector machine (SVM), which treats emotion characteristics as latent variables. In this testing phase, we want to train a classifier f_w to label an unknown video clip x as belonging to a certain behaviour class. One way to describe a linear model is as:

$$W^T\Psi(x,y,e) = W_x\psi_1(x) + \sum_{l\in\mathbb{E}} W_{e_l}^T\psi_2(x,e_l) + \sum_{l,m\in\mathbb{E}} W_{e_l,e_m}^T\psi_3(e_l,e_m)$$

Inside which the parameter vector *W*is$W = \{W_x; W_{el}; W_{el,em}\}$, and E sets of emotional characteristics. The coefficients learnt from the raw features x make up the template W_x, and the first term in Eq. gives the score that measures how well the raw feature ψ1(x) of a video clip fits this template.To determine if video clip x has an emotion, we may look at the second term in Eq., which gives the score for that emotion. During training, the behaviour label determines the initial value of el. In testing, a pre-trained emotion classifier provides this value. The recurrence of the pairs of feelings is captured by the third phrase.

In order to learn the model vector W from a collection of training examples, the following formulation is used as the learning objective function:

$$W^* = \min_W \lambda||W||^2 + \sum_{j=1}^{n} \max(0, 1 - y_j.f_w(x_j))$$

The quantity of regularisation is controlled by the trade-off parameter Λ, and a soft-margin is performed by the second term. A local optimum may be reached via the coordinate descent method because the objective function in Eq. is semi-convex. Belief propagation is used to identify the optimal combinations of emotions in our present implementation, where each emotion has two states {0} and {1}.

- **Baseline Methods**

Low-level Visual Feature Baseline: Using the algorithm provided by, we extracted the well-known packed trajectories for each video clip and used them as low-level features. To achieve this goal, we calculated state-of-the-art feature descriptors, such as dense trajectories inside space-time patches to take use of the motion data contained in dense trajectories, the histogram of oriented gradients (HOG), the histogram of optical flow (HOF), and the motion boundary histogram (MBH). With 15 frames, the patch is 32 × 32 pixels in size.

We construct a basic visual feature using the retrieved feature descriptors and the bag-of-words representations for every clip. Paying special attention to HOG, HOF, MBH, and Trajectories, we set the total amount of visual words to d=1000 and extract a codebook for each of these descriptors. We just used k-means clustering on a certain number of randomly chosen training characteristics to save time. Euclidean distance is used to assign descriptors to the words in the vocabulary that are closest to them.

In the end, the video descriptor was the retrieved histograms of visualisation words. We use a conventional one-vs-all multi-class SVM classifier for video classification. The average accuracy of each low-level feature describer, which we test separately using ground-truth label information of the behaviour. The describers are HOG, HOF, MBH, Trajectory, and Dense Trajectory. When compared to four other feature descriptors, dense trajectory feature performed better and attained an accuracy of 387.1% in detecting crowd abnormalities. We can see how different combinations of dense trajectory feature-based confusion matrices for various kinds of behaviour categories compare in terms of performance. Compared to other behaviour classes, the "Panic" category clearly performs better, with a result of 74.82%. This is likely because it deals with a less complex job. The closeness in motion patterns between this category and "fight" (very sharp motions) explains why this term is most often confused with it.

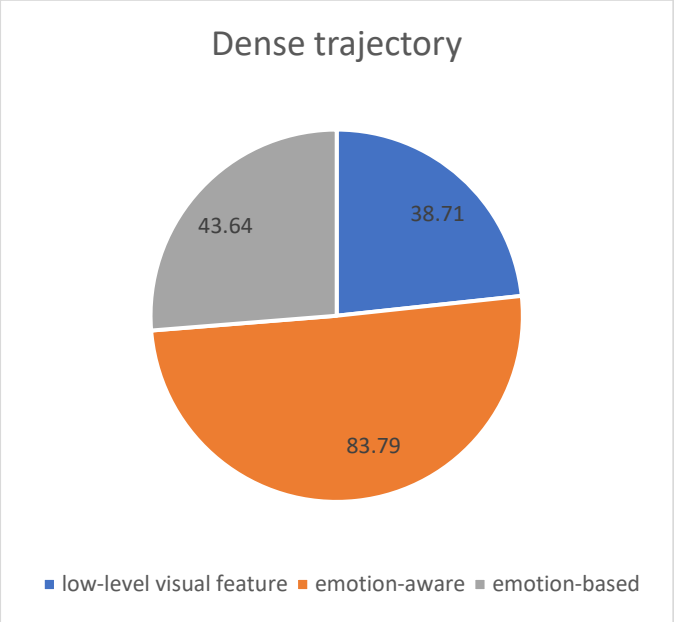

Fig 1: We compared the dense trajectory descriptor on our dataset's emotion-aware and emotion-based categories with low-level visual characteristics. For our dataset, we give the average accuracy across all classes.

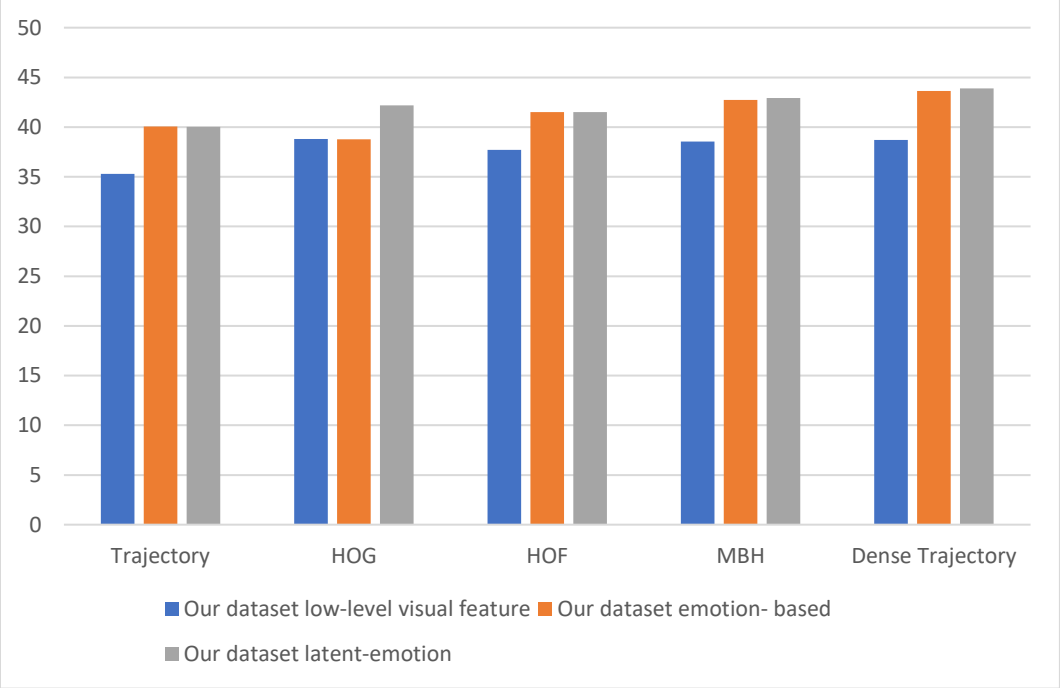

Fig 2: Our dataset's low-level visual feature, emotion-based, and latent-emotion categories compared using several feature descriptors (Trajectory, HOG, HOF, MBH, and Dense Trajectory). For our dataset, we give the average accuracy across all classes.

7. Results and Discussion

This section details the experiments that supported our emotion-based techniques. In the first trial, we use the premise that we have access to emotion labels throughout both training and testing times; in the second, we restrict this access to training time alone.

sensitive to Starting point: Here, we build features using the ground-truth label data of the crowd's emotions. Specifically, we start by standardising on a 6-dimensional binary feature vector for both the training and testing sets of data. For instance, a feature vector associated with a video clip with the emotion class "happy" is represented as $\{0,1,0,0,0,0\} \in E_6$, since there are a total of six emotion classes: "angry," "happy," "excited," "scared," "sad," and "neutral" correspondingly. We train a multi-class support vector machine classifier with the aberrant behaviour labels, taking created features into account. When it's time to test, we use test examples to see how well the trained classifier performed. Having an accurate emotion identification approach may be very useful for crowd behaviour

analysis, as shown by such substantial margins. These findings motivated us to use emotion as a middle-level representation for crowd movements in subsequent investigations.

An Experiment on Emotion-Based Crowd Representation: Here, we began by evaluating the aforementioned low-level feature descriptors independently using the emotion's groundtruth label information. Using dense trajectory characteristics, Compares the accuracy of various combinations of emotion categories in a confusion matrix, with an average accuracy of 34.13%.

The confusion matrix reports less-than-stellar findings, but it may still be useful to discover anomalous behaviour. Here, we learn a set of binary linear SVMs using the emotion labels in the training data, assuming that there is no emotion label available for the test data. Our name for this kind of classifier is $C_{ei}(f)$ emotion classifiers. Emotion classifiers provide a vector as its output, with confidence scores for emotion prediction shown on each dimension. This is something that we're looking at

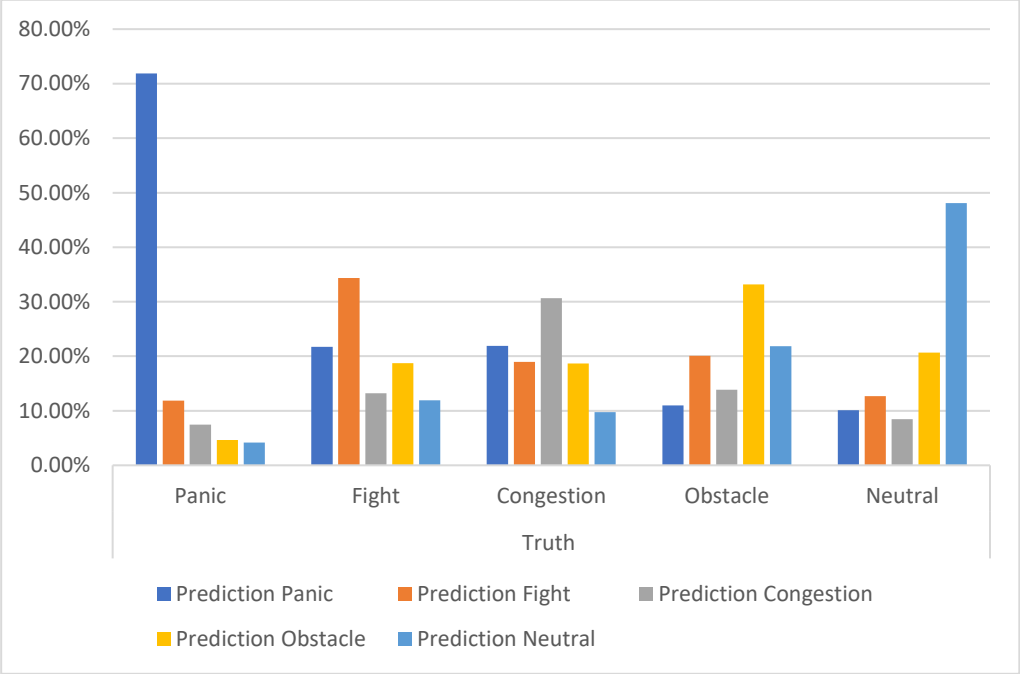

Fig 3: Misunderstanding matrix for every category based on emotions

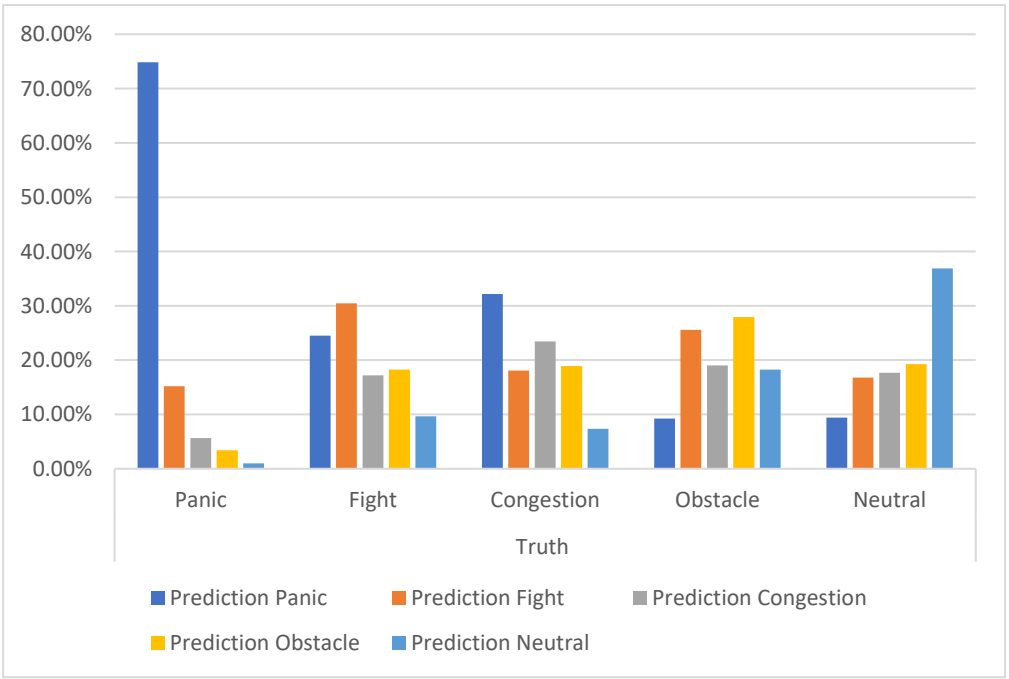

Fig 4: Matrix of misunderstandings for every category of basic visual features.

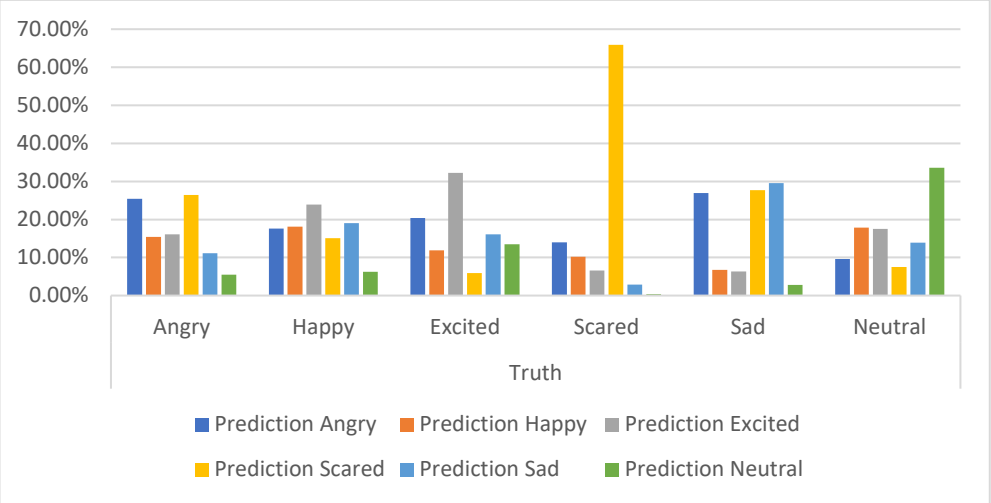

Fig 5: Matrix of misunderstandings for six sets of emotions.

use vector as a representation vector for behaviour categorization based on crowd emotions. Before training a multi-class support vector machine (SVM) with behaviour labels, we extract this vector for all test and train pictures. The last step in reporting the accuracy of behaviour categorization is to assess this classifier using test data.

Each of the low-level feature descriptors—HOG, HOF, MBH, Trajectory, and Dense trajectory—was treated independently using this technique. Out of all the low-level features, dense trajectory feature had the greatest accuracy at 43.64%. [26] Our technique achieves the best accuracy and increases it by over 7% when compared to two other baselines. The class of "fight" behaviours causes the greatest conflict with this class, accounting for 11.88% of the total, while the class of "Panic" behaviours has the best detection result at 71.87%. However, the "congestion" behaviour class had the poorest detection result, with a conflict rate of 21.92% with the "panic" behaviour class. [27,28] Our average accuracies for emotion-based classifiers and the emotion-aware baseline are consistent with these findings." These findings provide credence to the idea that we may improve performance with better emotion detection classifiers and more accurate emotion labels.[29]

A Study on the Representation of Latent Emotions in Crowds: Lastly, the model was learned using the latent SVM, with emotion labels being regarded as latent variables. 43.9% of them being from this experiment. This finding provides additional evidence that crowd emotion may be a useful mid-level representation for improving the identification of various kinds of crowd behaviour. [30]

8. Conclusion

We have put out a new crowd dataset that includes annotations for both aberrant crowd behaviour and crowd mood. The dataset has the potential to serve as a benchmark for the computer vision community and provide insights into the relationships between "crowd behaviour understanding" and "emotion recognition," according to our research. In our second contribution, we showcase a strategy that surpasses the baselines of both tasks by using the complementary knowledge of both. Specifically, the goal of future research is to define the emotion-to-behavior mapping function manually in order to identify a new class of deviant behaviours without training data.

8.1 Findings of the study

The study suggests a new crowd dataset with annotations of abnormal crowd behavior and crowd emotion. This dataset can be used as a benchmark for computer vision tasks and will help to understand the relation between crowd behavior understanding and emotion recognition. The study also presents a method that uses the complementary information of these two tasks, thus, the method outperforms the baselines. Future research will be concentrated on the identification of the new abnormal behavior classes without the training samples by the manual definition of the emotion-to-behavior mapping function.

8.2 Scope for further research

The subsequent studies could delve into more advanced deep learning architectures that can at the same time learn emotional attributes and crowd behavior representations in an end-to-end manner. The attribute learning should be expanded to the recognition of a larger number of crowd emotions and behaviors which would be also useful. Few-shot or zero-shot learning is another promising way to recognize new abnormal crowd behaviors without re-training by using the emotion-behavior mappings. The creation of interpretable models to clarify the logic of crowd behavior predictions based on emotional cues could boost the trustworthiness. Besides, the performance on

the more diverse in-the-wild datasets with different densities, viewpoints, and the real-world challenges is the key point. Lastly, the combination of emotion-aware crowd analysis with other multi-modal perception and decision systems could lead to the creation of smarter crowd monitoring and management applications.

References

[1] Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2016)

[2] Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2019)

[3] Deng, L.: An overview of deep-structured learning for information processing. In: Asian-Pacific Signal and Information Processing Annual Summit and Conference (APSIPA-ASC), Oct. 2021

[4] Vicsek, T., Zafeiris, A.: Collective motion. Phys. Rep. **517**(3), 71–140 (2012)

[5] Hinton, G.: Deep neural networks for acoustic modelling in speech recognition. IEEE Signal Process. Mag. **29**(6), 82–97 (2022)

[6] Yu, D., Deng, L.: Deep learning and its applications to signal and information processing. IEEE Signal Process. Mag. **28**(1), 145–154 (2021)

[7] Arel, I., Rose, C., Karnowski, T.: Deep machine learning—a new frontier in artificial intelligence. IEEE Comput. Intell. Mag. **5**(4), 13–18 (2020)

[8] Deng, L.: A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Trans. Signal Inf. Process. **3**, e2 (2017)

[9] Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**(4), 193–202 (2021)

[10] Lo, S.-C., Lou, S.-L., Lin, J.-S., Freedman, M.T., Chien, M.V., Mun, S.K.: Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans. Med. Imaging **14**(4), 711–718 (2021)

[11] Lecun, Y.B.L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE (2020)

[12] Krizhevsky, A., Sutskever, I., Geoffrey, E.H.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS 2012), vol. 25 (2022)

[13] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 1–42 (2019)

[14] Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Comput. Vis. Image Underst. **81**(3), 231–268 (2021)

[15] Bishop, C.M.: Pattern Recognition & Machine Learning, vol. 128, 1st edn, pp. 1–58. Springer, New York (2016)

[16] Kephart, J.O., Chess, D.M.: The vision of autonomic computing. Computer **36**(1), 41–50 (2023)

[17] Lemley, J., Bazrafkan, S., Corcoran, P.: Deep learning for consumer devices and services: pushing the limits for machine learning, artificial intelligence, and computer vision. IEEE Consum. Electron. Mag. **6**(2), 48–56 (2017)

[18] Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.: Computer vision for assistive technologies. Comput. Vis. Image Underst. **15**, 1–15 (2017)

[19] Liu, D., Wang, Z., Nasrabadi, N., Huang, T.: Learning a mixture of deep networks for single image super-resolution. In: Asian Conference on Computer Vision (2017)

[20] Wing, J.M.: Computational thinking. Commun. ACM **49**(3), 33–35 (2016)

[21] Sun, Y., Fisher, R.: Object-based visual attention for computer vision. Artif. Intell. **146**(1), 77–123 (2022)

[22] Huang S., Huang D., Zhou X. Learning multimodal deep representations for crowd anomaly event detection. *Math. Probl. Eng.* 2018;2018

[23] Yang M., Rajasegarar S., Erfani S., Leckie C. *2019 International Joint Conference on Neural Networks (IJCNN)* 2019. Deep learning and one-class SVM based anomalous crowd detection; pp. 1–8.

[24] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2023, pp. 1440–1448.

[25] Fang Z., Fei F., Fang Y., Lee C., Xiong N., Shu L., Chen S. Abnormal event detection in crowded scenes based on deep learning. *Multimedia Tools Appl.* 2022;75(22):14617–14639.

[26] Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2016)

[27] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G.: Recent

advances in convolutional neural networks. eprint arXiv:1512.07108, Dec. 2018

[28] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2016)

[29] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2016)

[30] Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: International Conference on Neural Information Processing Systems (2017)