Crowd-11: A Dataset for Fine Grained Crowd Behaviour Analysis

Camille Dupont* Luis Tobías* Bertrand Luvison
CEA, LIST, Vision and Content Engineering Laboratory,
Point Courrier 173, F-91191 Gif-sur-Yvette, France
{camille.dupont, joseluis.tobiasquiroz, bertrand.luvison}@cea.fr

Abstract

Crowd behaviour analysis continues to be a challenging task in computer vision, mainly due to the high complexity of the interactions between groups and individuals. This task is crucial given the magnitude of manual monitoring required for effective crowd management. Specifically, it has received a lot of attention among the video surveillance community to detect potentially dangerous situations and to prevent overcrowding. Within this context, a key challenge for research is to conceive a highly generic and fine characterisation of crowd behaviours according to their appearance and motion, given any camera perspective and in any context. Since current datasets answer only partially to this problem, a new dataset is generated and labelled accordingly to solve it. This dataset defines a total of 11 crowd motion patterns and it is composed of over 6000 video sequences with an average length of 100 frames per sequence. In order to establish the first baseline of crowd characterisation on the newly created dataset, an extensive evaluation on shallow and deep methods is performed. This characterisation is expected to be useful in multiple crowd analysis situations. We present a new deep architecture for crowd characterisation and demonstrate its application in the context of anomaly classification.

1. Introduction

Video-surveillance for crowd monitoring is becoming a problem of real interest for authorities, especially in big cities. Lately, more and more cameras are deployed in urban areas or public gathering spots such as train stations, airports, etc. All these cameras could provide precious information for crowd monitoring or abnormal event detection. Unfortunately, the humongous amount of information is so complex to analyse that it is exploited very sparsely for either forensic activities or partial live monitoring. Automatic processing of such data remains a challenging task that must be addressed. State of the art approaches for such problems have shown that one of the key issues lies in a proper crowd modelling. Early work focused only on partially modelling crowd behaviour, in order to detect specific movements such as panic escape, wrong way movement, etc. More recent approaches target a more exhaustive crowd behaviour modelling.

Different from a single person, a crowd is a significantly richer and less structured organism and does not express itself as a single entity: its behaviour is not merely the sum of each of its members individual behaviour, nor one collective will, but an inter-reacting mixture of collective behaviour coupled with heterogeneous individual goals. As a consequence, the variability of the crowd behaviour manifold in the video space is tremendously larger than for a single person. This multi-modal property of crowds renders the association between a crowd and a unique crowd behaviour in a video an ill-posed problem. Finally, the definition of a crowd itself is rather ambiguous. Up until which distance and motion similarity individuals are considered to be together, i.e. belonging to a same group? How to quantify the influence of their behaviour on the crowd dynamics? This notion of group is strongly linked to the density of the scene and the definition of a crowd and its elements. All these properties inherent to crowds make the problem of crowd behaviour characterisation extremely hard to solve. Despite this fact, finding such a characterisation would lead to improvements in multiple crowd analysis problems, for example crowd anomaly classification.

The contribution of this paper is threefold. First, a new dataset for fine grained crowd behaviour analysis is proposed. This dataset is substantially larger than existing datasets trying to solve this paradigm. Second, baseline results of recent methods for crowd characterisation on the proposed dataset are provided, one shallow and three deep approaches are compared. Third, to demonstrate the genericness of the fine-grained characterisation, an application of deeply learned models is employed in the context of supervised anomaly classification.

*Authors contributed equally

1.1. Related work

Crowd analysis comprises various sub-problems and applications. Works tackling these problems provided not only interesting approaches to improve crowd analysis and modelling, but also contributed with compelling datasets to the public domain. Among these applications can be cited:

a) Counting or density estimation. This problem is mostly considered from a static perspective; approaches like [25, 26] use Convolutional Neural Networks (CNN) while the approach from [10] relies on head detections to perform the density estimation.

b) Crowd segmentation, which aims at defining crowd boundaries. Likewise, this problem is considered from a static perspective, using Fully Convolutional Neural Networks (FCNN) in [12].

c) Crowd video context description. This application proposed by [18, 19] aims at classifying videos according to context information that could be used as tags, and is solved using deep architectures.

d) Crowd behaviour analysis. Early works on the topic usually focused on a specific behaviour such as panic movement [14], wrong way displacement [13], violence, etc. Most of the datasets used have been shot especially for the specific behaviour which is analysed, therefore, are not fully representative of what a crowd can look like in everyday life situations. Furthermore, they are relatively small datasets and the performances of the derived models are not truly evocative of their generic nature (see section 1.2 for more details on existing crowd behaviour datasets). More recent works [7, 20] tackle the problem of characterising the motion properties of the crowd to distinguish types of crowd motion (i.e. bottleneck, laminar mainstream, etc).

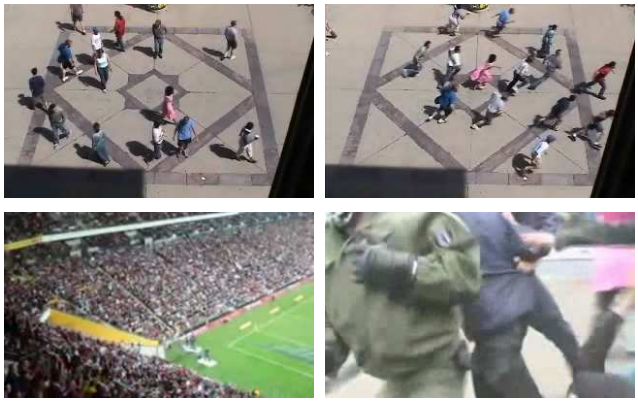
Additionally, approaches using CNN's are knowing a very recent interest. However, the number of architectures tackling crowd behaviour analysis is not yet abundant. Two approaches on crowd attributes can be mentioned: Shao et al. [18] use a derived version of [21] with custom motion features deduced from their previous work [20], and [19] approximate 2D+t analysis with successive 2D CNN's on xy, xt and yt slices extracted from a given video. Another interesting approach that exploits Coherent Recurrent Neural Networks (CRNN) is introduced in [22], which uses as input trajectory descriptors.

1.2. Existing datasets and their limitations

Describing precisely the way a crowd behaves is challenging because, as mentioned before, the variability in video space is extremely vast. Several datasets have been proposed to answer this problem.

UMN [2] focuses on a single behaviour, panic movement. It contains only three scenes, each containing few videos, which is far too restrictive. It also tends to exaggerate the behaviour which is to be recognised (i.e. abnormality defined only as an acute singularity: non-smooth change of the velocity field) (Fig. 1). In spite of presenting a wider range of crowd behaviours (bottleneck, dispersion, crossing flows, etc), PETS [1] and MED [16] (Fig 2) datasets also lack of diversity in appearance (with only one or two scenes) and Agoraset [4] suffers from lack of realism due to its basic computer-generated textures (Fig 2).

In Violent-Flows [8] and Hockey Fight and Movies [15]



(a) Normal (b) Abnormal

Figure 1: Specific and exaggerated abnormal/normal behaviour present in UMN (top) and Violent-Flows (bottom). Similar behaviours are found in the Hockey Fight and Movies dataset.

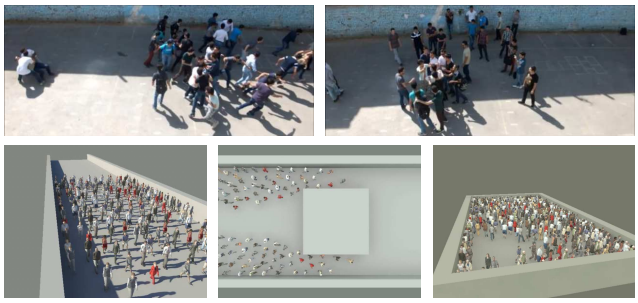


Figure 2: Behaviour lacking appearance diversity in MED (top) or lacking realism in Agoraset (bottom).

(a) Intervened escalator traffic (b) Crowd merge

Figure 3: Image (a.1) is attributed the same class as (a.2) due to the presence of escalators, albeit being closer to (b) in terms of behaviour. In CUHK.

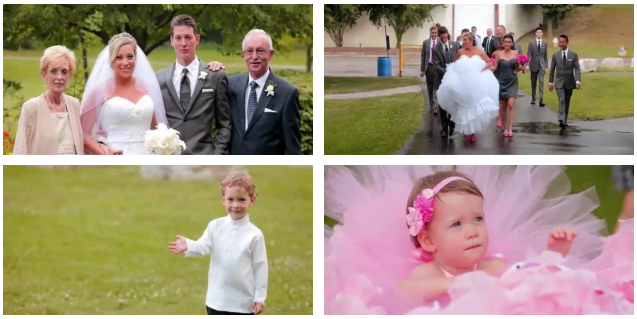


Figure 4: Samples of a same video, thus of same behaviour labels, found in the WWW dataset.

Label	Class names	# videos (N_{V_i})
0	Gas Free	529
1	Gas Jammed	520
2	Laminar Flow	1304
3	Turbulent Flow	892
4	Crossing Flows	763
5	Merging Flow	295
6	Diverging Flow	184
7	Static Calm	737
8	Static Agitated	410
9	Interacting Crowd	248
10	No Crowd	390

Table 2: List of crowd video classes.

datasets, most of the footages come from realistic environments, videos seem to be issued from mobile phones: the camera motion is very characteristic, the resolution is limited and the action variability is restricted. Indeed, it either consists of few individuals fighting or of far-viewed stadium crowds (Fig. 1). However, they present events that are particularly interesting for crowd management system, that is to know, fighting situations.

CUHK [20] is a compelling dataset and contains many definitions of how a crowd can evolve. Yet, it suffers from two main drawbacks. The first one is a small amount of videos for some classes. For example, the class crowd merge is formed by only 9 videos of 30 frames, to be distributed in training and testing sets. This could lead to a poor modelling of the unbalanced classes. The second drawback is its class definition itself that is sometimes ambiguous. As shown in Fig. 3, CUHK tends to preponderate the spatial context in which the crowd is evolving (e.g. escalators) over the actual crowd behaviour. In addition, the class definition presents many inaccuracies that can be very detrimental to any learning algorithm requiring precise indexing.

From all existing crowd behaviour datasets, one can conclude that: (a) despite their alluring characteristics, datasets tend to be either unrealistic or (b) rather small in terms of observed behaviours and quantity of video sequences. Thus, there is the need to create a larger database from which complex learning approaches requiring considerably large amounts of data such as Deep Learning (DL) would benefit.

To clearly analyse the behaviour of the crowd in any situation, an exhaustive characterisation of its dynamics is needed. To that end, we propose a substantial dataset, which similarly to [20] aims to solve the crowd characterisation but in a finer manner. This is accomplished by defining a wide set of classes describing the ways a crowd can evolve across time, without having any a priori over the people present in the video nor the area where the action takes place (context independence).

Note that while other crowd datasets are large enough for DL, they are dedicated to solve other problems: density estimation and counting (Worldexpo'10 Crowd Counting Dataset [25], UCSD [5], UCF_CC_50 [10] which usually consist of images rather than video sequences and provide mainly head location), and crowd attributes (WWW [18]) which answers the question *Who does What and Where?*. The latter being the largest public crowd dataset, we must stress that such a classification focuses mainly on providing context information rather than the actual behaviour. Even the action attribute corresponding to the *What* question is centred on the main goal of the crowd, which is deduce solely by the context, rather than on the way the crowd evolves (moves). For example in Fig. 4, although very different crowd behaviours are comprised in the video, they are all associated to the same labels (outdoor, street, newly-wed, couple, walk). Such an annotation is relevant for global crowd understanding but not for genuine and precise crowd behaviour understanding.

2. Proposed Dataset

The proposed dataset intends to define characteristic crowd motion patterns that are representative of everyday life behaviour. Because many breakthroughs in computer vision have recently been achieved thanks to Deep Learning, a specific motivation while creating this dataset was to provide a dataset large enough for learning CNN's, which are known to be data driven approaches. Hereafter, the proposed dataset will be named Crowd-11 in view of the number of its classes. A comparison with other datasets in terms of size and properties is provided on table 1. The majority of the videos in Crowd-11 have been manually selected and extracted from the web using keywords such as commuters, transit crowd, rush time, subway, airport, etc. Some of the existing datasets have also been partly used (WWW [18], CUHK [20], Violent-Flows [8], Worldexpo'10 Crowd Counting Dataset [25], Agoraset [4], PETS [1], UMN [2], Hockey Fight and Movies [15]).

Several approaches modelling crowd movements use fluid dynamics descriptions [9]. While this analogy makes

dataset	CUHK	Violent-Flows	UMN	Worldexpo'10	WWW	MED	**Crowd-11**
goal	**behaviour**	violence	**behaviour**	counting	attributes	**behaviour**	**behaviour**
# videos	474	246	11	1,132	10,000	31	**6,272**
# scenes	215	246	3	108	8,257	3	**3,005**
# frames	60,384	22,074	7,739	3,980	$>$ 8 million	45,000	**621,196**
resolution	multiple	320×240	320×240	720×576	640×360	640×480	**multiple**

Table 1: Comparison between the proposed and the existing datasets for crowd analysis.

sense for high crowd densities as the motion equations often resemble the two-dimensional Navier-Stokes equations, crowds with lower densities do not strictly obey viscous fluid or gas dynamics and a more thorough analysis of crowd phenomenology is needed. From this partial connection with fluid dynamics and recent representations of crowd behaviour [20], we construct 11 classes named in table 2. We distinguish two types of crowds:

a) Crowds that are dynamic, dense and structured enough to follow flows, i.e. groups that can be segmented. Note that unless they cause interference amongst themselves or are of different behaviour, the number of flows in the scene is irrelevant to the way the global motion is described.

b) Crowds with no perceivable streams. These crowds do not form collective groups, individuals act independently of the global crowd motion.

For the sake of clarification, we explicit the criteria of each class, illustrated in Fig. 5. Like in fluid dynamics, a laminar flow occurs when the individuals of a crowd follow a smooth stream. The notion of laminarity being inherently related to the viscosity of the fluid, it can only be applied to dense crowds, i.e groups in which the individuals have a very small leeway before disrupting their neighbours and causing the stream to not flow well. Hence the necessity to keep a structure and stable velocity field (in magnitude and direction) over time. Scattered individuals rarely meeting this criterion as they move independently from each others, we say nevertheless that they follow a laminar flow when they follow an easily distinguishable stream which flows well (i.e. the individuals are not disrupted from their trajectories). On the contrary, a flow that undergoes a disturbance is said to be turbulent. A dissimilarity of the velocity field does not necessarily implies a disturbance (e.g. in an almost empty corridor, some individuals going faster than others do not cause a disturbance). Crossing flows are intertwined streams in opposite directions. The streams, too thin to be individually segmented, form a global stream of a same orientation but different direction. Merging (diverging) flows are characterised by a compression (expansion), e.g. in a bottleneck situation. Such classes can be seen as a transitive phase between laminar and static (see paragraph below).

In the following classes, the individuals act too independently to form perceivable streams. Gas free refers

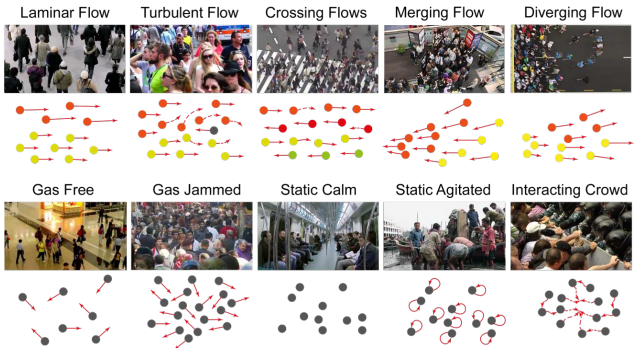


Figure 5: Illustration of the proposed classification. The flows of the first 5 classes are illustrated by a same color. In gray are represented the individuals that do not belong to any group. Best viewed in colour.

to very scattered individuals whose rectilinear trajectories are not disturbed despite the dissimilarity of their velocity/direction. In gas jammed, the scene is so crowded that the trajectories of the individuals disturb one another. Two classes cover the behaviour of static crowds (i.e. no spatial progression over time in the scene), one with no movement (crowd waiting or watching) and one with moving individuals (applauding, dancing, etc.). An interacting crowd contains individuals who move towards each others in a violent and erratic way. An 11th class is added as background, *i.e.* void of crowd (can contain nonetheless other dynamic entities such as cars). As videos sometimes exhibit several behaviours in time and space, they were cropped and trimmed to fit a unique well-defined behaviour for training purposes. Hence a wide variability of resolutions, ranging from 220×400 to 700×1250. To ensure a dataset design without any assumption on the field of view or on the context of the scene, several clips of unique behaviour were often extracted from a same scene.

Sources URLs from which this dataset has been drawn, as well as annotation over original videos to get the proposed dataset can be asked by e-mail at crowd11-dataset@cea.fr.

3. Learning crowd characterisation

In order to provide the baseline in the proposed dataset, several state of the art methods have been chosen to be meticulously evaluated and compared on the Crowd-11 dataset: an *ad hoc* approach and several variants of three deep network architectures.

3.1. Chosen approaches

Shao et al. [20] published one of the most recent *ad hoc* works on crowd behaviour analysis. However, this method cannot process static video streams. Indeed, this approach relies on point trajectories over the crowd and only models moving groups. This is why their model is not adapted for the analysis of loitering people, whether they are agitated or not: trajectories can be extracted but are confused with noise, and their implementation filters them. As a consequence, "static crowd", "agitated crowd" and most of the "no crowd" class cannot be processed with this method.

As mentioned in 1.1 more recent approaches for crowd characterisation use deep learning techniques to address this issue. The work of [18] aims to find attributes of crowds that can be used to infer the main crowd activity in a contextual manner. The fact that this approach strongly relies on context makes it not directly correlated with our objective, thus, cannot be compared.

Similar to [20], Su et al. [22] use trajectory descriptors as input of their CRNN, in contrast to an end-to-end raw pixels training. Moreover, the model of this architecture is not freely available yet, thus, this approach has not been tested on the proposed dataset.

On the other hand, deeply-learned human action recognition is meeting significant progress in recent years, thus, we chose to investigate how two recent approaches [21, 23] perform on the proposed dataset. Additionally we propose a third network architecture. These architectures were designed to process spatio-temporal information, as a consequence, they can also be applied to crowd video analysis. Multiple declinations of the three architectures have been learned either by domain adaptation, or by a complete learning based entirely on the Crowd-11 dataset, results and detailed descriptions are presented in section 3.3. The selected approaches are:

a) Two-stream architecture [21] which uses a motion feature, composed of the stack of 10 successive optical flow images, coupled with a color image to capture the appearance of the scene. Each stream of the network is based on the VGG architecture. In a late stage the two streams are fused together by either averaging their scores or concatenating them before a final fully connected layer.

b) C3D [23] which consists of a succession of 3D convolutions where the input's third dimension corresponds to a temporal stack of images that form a clip. A clip is defined as a pack of 16 consecutive RGB frames. The network follows a configuration of five 3D convolution + 3D Pooling layers, followed by three Fully Connected (FC) layers.

c) V3G is a network specifically proposed to tackle the crowd behaviour analysis problem. This approach combines the key ideas of C3D [23] and VGG [21] networks with Batch Normalisation (BN) [17]. The network follows the VGG network structure with the main differences of using 3D convolutional layers and the addition of BN layers after each 3D pooling layer, similarly the three FC layers are transformed in convolution layers with a size of $1 * 1$ and a high dropout ratio (0.8) to avoid over-fitting.

3.2. Technical details

The publicly available code from [20] was used to evaluate the performance of their approach in the Crowd-11 dataset. In order make a fair comparison this *ad hoc* approach was tested on the subset of valid data remaining after removing the static classes.

The networks [21,23] were trained using respectively the publicly available toolbox CAFFE [11] and a modified version supporting 3D convolutions [23]. Pre-trained models on UCF-101 provided by Wang [24] for two-stream model and on Sports-1M provided by Tran [23] for C3D were used. V3G network was trained using [3]. Data augmentation was applied (random crops, vertical mirrors and temporal overlaps). Optical flow for [21] is computed using the Dense Optical Flow algorithm [6] from the OpenCV toolbox. All networks are optimised using Stochastic Gradient Descent (SGD).

The Crowd-11 dataset being large enough for holding out a reasonable portion of it for reliable testing, it is split into three subsets so as to train the model, perform a cross validation and a test evaluation. For each class C_i, the test, train and validation splits consist respectively of m, $0.9 \times (N_{C_i} - m)$ and $0.1 \times (N_{C_i} - m)$ videos, where m is the fixed test size for all classes and N_{C_i} is the number of videos of this class. In this way we prevent an imbalanced class evaluation. Same-class scenes do not overlap on the three sets to ensure the effectiveness of the model's generalisation capacity. A scene is defined as a specific area viewed with a specific viewpoint.

3.3. Results and interpretation

Several variants of the two stream architecture have been tested on Crowd-11, both appearance and motion models are finetuned from the original model of [21]. Appearance-based descriptors surpass motion-based ones when learned independently. Motion entries contain far more information (30 channels vs. 3 channels) and are more complex to treat, generating an increased number of parameters in the model, thus, being prone to overfitting. Merging the spatial and temporal information yields to better results, especially

Variant	Per clip accuracy	Per video accuracy
Concatenation of RGB and OF	**54.8%**	**56.8%**
Average of RBG and OF scores	49.8%	50.7%
RGB	47.5%	49.8%
Optical Flow	27.5%	28.2%

Table 3: Results obtained from the multiple variants of Two-Stream network on the Crowd-11 database, best performance is presented in bold.

Network	Per clip accuracy	Per video accuracy
C3D Sports 1M + Finetuning	**61.6%**	**63.7%**
V3G UCF-101 + Finetuning	58.0%	59.6%
V3G Crowd-11	57.8%	59.4%
C3D UCF-101 + Finetuning	49.2%	51.3%
C3D Crowd-11	46.9%	47.4%

Table 4: Results obtained with the C3D and V3G Networks on the Crowd-11 database. Accuracy of the Top-1 prediction.

when the two stream are concatenated on late stages and then classified using a final fully connected layer, instead of averaging the scores of both streams individually [21]. This result can be explained, since appearance and motion are, to a certain extent, correlated information, and simply averaging scores does not reveal such correlation, whereas concatenating descriptors and using a last fully connected layer enable the modelling of such a correlation. Results are presented in table 3.

For the C3D convolution method two different strategies are employed: first a whole training in the Crowd-11 database and second a finetuning using as weight initialisation pre-trained models on the Sports-1M [23] or the UCF-101 datasets. The model learned using only the Crowd-11 data reaches a 46.9% accuracy score, UCF-101 finetuned model obtains 49.2%, and the Sports-1M finetuned model rises up to 61.6% in the per clip evaluation. Accuracy scores are also computed in a per video fashion averaging overlapped clip scores from the entire video, the results are 47.4%, 51.3% and 63.7% respectively. The confusion matrix of the best C3D architecture is shown in (Fig. 6). Notice that the network is able to handle both dynamic and static behaviour.

In the case of the V3G model, since it is first proposed in this work, we proceeded to first train it on the UCF-101 dataset, then finetune it on Crowd-11 obtaining 58.0% per clip accuracy and 59.6% per video accuracy. Finally, a version entirely trained on the Crowd-11 database reaches a per clip accuracy of 57.8% and a 59.4% per video accuracy, see table 4.

Best results are obtained with the C3D Sports-1M finetuned model, however, the deeper network V3G shows similar characteristics. Importantly, when trained on the split 1 of the UCF-101 dataset for action recognition, the V3G network outperforms C3D with 49.9% vs 47.3% accuracy on the test split of UCF-101. In addition, the V3G model is faster to train due to the batch normalisation and the reduction of parameters.

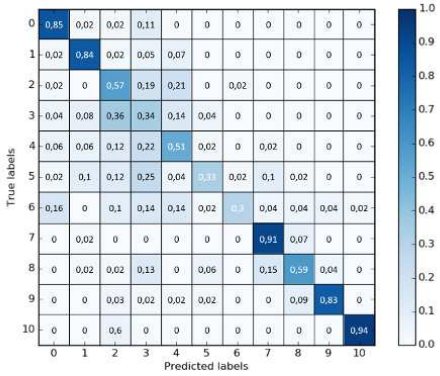


Figure 6: Confusion matrix on the Crowd-11 test set, using the C3D Sports-1M finetuned model.

When compared to the best C3D finetuned model, two-stream provides lower results. The intuition behind this result is a better modelling of action time variation thanks to the temporal-wise convolution done in the 3D space, a larger temporal window and an early fusion of appearance and temporal information that can be efficiently modelled by the network. Qualitative results of the finetuned C3D network are shown on Fig. 7.

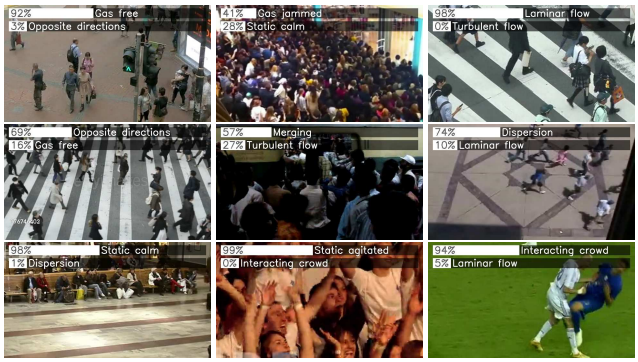


Figure 7: Results obtained with the finetuned C3D architecture, with top and second class probabilities. Best viewed electronically.

All of the tested CNN models have no problem recognising other entities, such as cars from persons because the "no crowd" class contains mainly scenes with moving vehicles and background. This property can be a great asset as vehicles can have types of motion which can be wrongly interpreted by trajectory based methods. The weakness of the 3D networks lies in the classification of transitional be-

haviour. Indeed, merging dense flows can be easily confused with other dense and/or static crowds as the movement can be extremely slow. The confusion between diverging flows and more common flows can be explained by the fact that those flows do not always present a constant growth of the velocity field magnitude but rather incremental changes which are hard to capture over the temporal windows size (10 frames for two-stream and 16 for C3D and V3G).

Finally, as stated earlier, *ad hoc* features of [20] have been evaluated on the Crowd-11 dataset without motionless classes. In the 8 remaining classes, due to implementation details, the features could only be generated on 72% of the test set. This is because of the lack of valid trajectories in some videos: due to either very far or close viewpoints of the videos. As a result, only 52% of the full test set was employed for evaluation of the group descriptors. We compare them with the best C3D model, for a fair comparison, we present the performances obtained on the same reduced testing set. Results are shown on Fig. 8 using confusion matrices.

Figure 8: Confusion matrices of the crowd video classification on only dynamic classes. Left: using group descriptors [20], average per video accuracy is 50,0%. Right: using finetuned C3D model, average per video accuracy is **56,7%.**

Apart from the global accuracy which is significantly higher on the finetuned C3D model, one can notice that group descriptors fail to classify videos consisting of disorganised trajectories which cannot be clustered (Gas Free), but are better at distinguishing laminar and turbulent flows thanks to the group segmentation and modelling of its dynamic steps. Table 5 presents a summary of the experiments.

4. Crowd characterisation for anomaly classification

The main objective of obtaining a fine-grained crowd characterisation is to be able to analyse how a crowd evolves across time. Here we show how the models learned on the Crowd-11 dataset using deep architectures can be easily repurposed for specific applications. In this case we employ two models to perform anomaly classification. As mentioned in 1.2, the MED dataset is not well suited for learning a crowd characterisation, moreover, it was designed to be a common ground for benchmarking anomaly classification [16]. This dataset consists of five behaviours: panic, fight, congestion, obstacle (abnormal object) and neutral.

Method	Per clip accuracy	Per video accuracy
Group Profiling Descriptors *	-	50.0%
Two-stream OF + RGB concatenation	54.8%	56.8%
V3G UCF-101 + Finetune	58.0%	59.6%
C3D Sports-1M + Finetune	**61.6%**	**63.7%**

Table 5: Accuracy of state-of-the-art *ad hoc* and neural network methods on the test set of the Crowd-11 database. Note: group descriptors were only evaluated on 52% of the test set, see technical details in section 3.2.

The V3G trained on Crowd-11 and the finetuned C3D Sports-1M models are used to generate features issued from the FC7 and FC8 layers on the MED dataset. Then, following the experimental set-up of [16] the leave-one-out strategy is carried out to learn *k-folds* linear Support Vector Machines (SVM). However, the dictionary creation and the quantification steps are omitted since the features issued by the CNN's are of a fixed size and of a higher hierarchical level than the features used in [16]. Table 6 presents the evaluation results.

	Panic	Fight	Congestion	Obstacle	Neutral	MeanAcc
V3G-FC7	80.72%	**37.41%**	**31.18%**	**47.25%**	71.35%	**53.58%**
C3D-FC7	**84.72%**	32.93%	16.16%	29.61%	**92.69%**	51.22%
MED	74.82%	30.47%	23.48%	27.94%	36.88%	38.71%
V3G-FC8	53.23%	29.89%	27.32%	42.35%	32.16%	36.99%
C3D-FC8	57.32%	25.89%	17.22%	25.51%	46.64%	34.50%

Table 6: Results of the anomaly classification on the MED database, best results are presented in bold

For both of the CNN's can be observed that features issued from the FC8 layer generate slightly lower results than the approach from [16], while the best FC7 features produce better results. The performance discrepancy between the FC7 and FC8 features can be attributed to their complexity: while the FC8 feature is a compact signature of dimension 11, the feature FC7 resides in a larger feature space of dimension 4096, and therefore, is expected to preserve more detailed information. Best accuracy is obtained using the V3G FC7 features, performing consistently across the five behaviours and outperforming the method proposed in [16], we assume this behaviour is strongly linked to the training of the characterisation model being done purely on crowd videos.

When compared with the original method proposed in [16] the V3G features are not only largely better at de-

tecting normal behaviour, but also at identifying better the abnormal situations. Moreover, the method performs sub-optimally in two out of five classes: fight and congestion. This detriment on the classification can be explained from the high overlap in terms of both appearance and motion between the two classes. Genericness of the models learned from Crowd-11 can be observed in this application where features are directly employed to perform the classification of anomalies.

5. Conclusion

Fine grained crowd behaviour analysis is a challenging task. Deep supervised machine learning approaches not only need immense quantities of videos, but also accurate class definition. For this reason, we built a new dataset defining 11 classes that depict how a crowd can evolve across time within a video scene. Furthermore, we show that this problem can be encouragingly tackled with Spatio-temporal Convolutional Neural Networks. Specially, architectures that deal straightly with temporal stacks of raw pixels as inputs such as C3D deliver the best performance. A new spatio-temporal CNN, has been presented: the V3G model which shows potentially good results at crowd characterisation. Lastly, we have shown how this fine-grained characterisation can be employed to classify anomalies yielding to interesting results. Further improvements will be done to improve the crowd characterisation, i.e. coupling the existent solution to RNN to cover a longer and variable temporal observation window.

Acknowledge

This work was supported by the research project Fluid-Tracks (co-funded by French FUI-BPI France and Conseil Général de Seine et Marne).

References

[1] Dataset - PETS: Performance Evaluation of Tracking and Surveillance. http://www.cvg.rdg.ac.uk/slides/pets.html. 2, 3

[2] Unusual crowd activity dataset made available by the University of Minnesota. http://mha.cs.umn.edu. 2, 3

[3] Video-Caffe, a video friendly caffe. https://github.com/chuckcho/video-caffe. 5

[4] P. Allain, N. Courty, and T. Corpetti. AGORASET: a dataset for crowd video analysis. In *ICPR*, 2012. 2, 3

[5] A. B. Chan, Z.-S. John, and L. N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. *CVPR*, 2008. 3

[6] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *SCIA*, 2003. 5

[7] H. Fradi, B. Luvison, and Q. C. Pham. Crowd behavior analysis using local mid-level visual descriptors. In *TCSVT special issue on "Group and Crowd Behavior Analysis for Intelligent Multi-camera Video Surveillance"*, 2016. 2

[8] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *CVPR, Workshop on Socially Intelligent*, 2012. 2, 3

[9] R. Hughes. The flow of human crowds. *Transportation Research Part B: Methodological*, 2014. 3

[10] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, 2013. 2, 3

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ICM*, 2014. 5

[12] K. Kang and X. Wang. Fully convolutional neural networks for crowd segmentation. *arXiv:1411.4464*, 2014. 2

[13] B. Luvison, T. Chateau, P. Sayd, Q. C. Pham, and J. Laprest. Automatic detection of unexpected events in dense areas for videosurveillance applications. In *INTECH*, 2011. 2

[14] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009. 2

[15] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar. Violence detection in video using computer vision techniques. In *CAIP*, 2011. 2, 3

[16] H. Rabiee, J. Haddadnia, H. Mousavi, M. Kalantarzadeh, M. Nabi, and V. Murino. Novel dataset for fine-grained abnormal behavior understanding in crowd. *AVSS*, 2016. 2, 7

[17] I. Sergey and S. Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML-15*, 2015. 5

[18] J. Shao, K. Kang, C. C. Loy, and X. Wang. Deeply learned attributes for crowded scene understanding. In *CVPR*, 2015. 2, 3, 5

[19] J. Shao, C. C. Loy, K. Kang, and X. Wang. Slicing convolutional neural network for crowd video understanding. In *CVPR*, 2016. 2

[20] J. Shao, C. C. Loy, and X. Wang. Scene-independent group profiling in crowd. In *CVPR*, 2014. 2, 3, 4, 5, 7

[21] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *NIPS*, 2014. 2, 5, 6

[22] H. Su, Y. Dong, J. Zhu, H. Ling, and B. Zhang. Crowd scene understanding with coherent recurrent neural networks. In *IJCAI*, 2016. 2, 5

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 5, 6

[24] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. *CVPR*, 2015. 5

[25] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, 2015. 2, 3

[26] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016. 2