A Comparison of Crowd Commotion Measures from Generative Models

Sadegh Mohammadi Hamed Kiani Alessandro Perina Vittorio Murino

Pattern Analysis and Computer Vision Department (PAVIS)
Istituto Italiano di Tecnologia
Genova, Italy

Abstract

Detecting abnormal events in video sequences is a challenging task that has been broadly investigated over the last decade. The main challenges come from the lack of a clear definition of abnormality and from the scarcity, often absence, of abnormal *training samples. To address these two shortages, the computer vision community made use of generative models to learn normal behavioral patterns in videos. Then, for each test observation, a (crowd) commotion measure is computed quantifying the deviation from the normal model. In this paper, we evaluated two different families of generative models, namely topic models, representing the standard choice, and the most recent Counting Grids which have never been considered for this task. Moreover, we also extended the 2D Counting Grid, introduced for the analysis of images, to three dimensions, making the model able to capture the spatial-temporal relationships of the videos. In the experimental section, we compared all the approaches on five challenging sequences showing the superiority of the 3-D counting grid.*

1. Introduction

Video surveillance systems are becoming a ubiquitous feature in our cities. Their value, however, is often questioned, especially for preventing crimes [2]. Indeed, the deployment of a multitude of cameras has little value without trained personnel or automatic algorithms to support them. This has fueled research in algorithms for the automatic detection of abnormal behavior in surveillance, feeds which in a real scenario, may trigger human intervention and help crime prevention. The biggest challenge lies in the definition of abnormality as it is strongly context dependent. Although panic and violence are mainly considered as two common examples of abnormality in the context of video surveillance, people running or walking in some areas of a scene may be considered as an abnormal event in particular contexts. This observation demands an in-the-box viewpoint about the abnormality in which introducing a context-dependent commotion measure seems crucial. For this purpose, the abnormal behaviors usually appear as commotion. Therefore, anomaly detection in general reduces to the definition or the learning of an *unsupervised commotion measure* with an associated abnormality threshold τ.

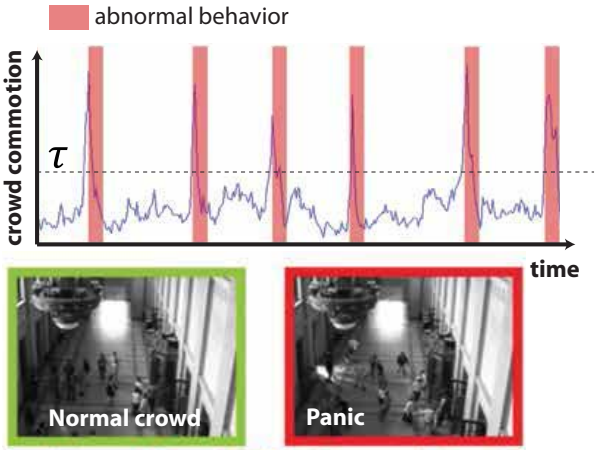

Figure 1. Negative likelihood as measure of crowd commotion for a sequence taken from the UMN dataset. As when an abnormality occurs we have a spike in commotion.

Although in the literature there exist several approaches that define ad-hoc measures starting from low-level video cues [21, 20], a popular approach is to employ generative models negative log-likelihood as a measure of abnormality [11, 12]. Generative models are built to explain how data can be generated and they are particularly attractive for our goal because *i)* they do not require abnormal data at train time, *ii)* it is easy to encode prior knowledge about the particular surveillance context and *ii)* they support online learning [9], therefore they can update the model as new footage comes in. In such data driven approaches, one exploits normal footage to automatically learn a model of ordinary behavioral patterns, and define abnormality as

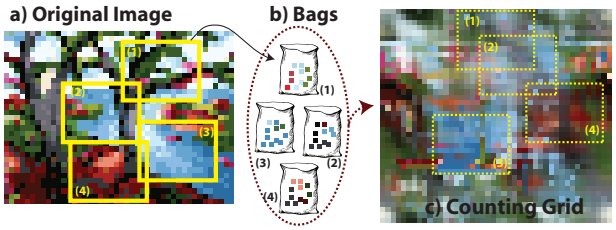

Figure 2. Layout reconstruction. a) Original image. b) Patches described by bags of features. c) Counting grid.

something which has low likelihood or deviates from what has been learned. This is illustrated in Fig. 1.

The main challenge here is in choosing a feature representation which provides good generalization while being fast to extract and handle. One common efficient choice is to simply represent a frame or a clip, as disordered "bags" of features, where a clip is a short contiguous sequence of frames. In a nutshell after extracting a low-level video description, a number of spatio-temporal patches are sampled from a frame or clip, clustered in K centers (e.g., codebook) and a discrete "codeword" is assigned to each feature descriptor. Then, an image is described by a histogram over the codebook entries. This is the so called *Bag of Words* paradigm, where each frame or clip is represented by a histogram over orderless or exchangeable features [6]. In particular, in the context of crowd behavior analysis the bag of word representation has already been considered by several works [7, 13, 11], such simplification makes this approach more robust to geometric variation of scenes, occlusions, and image transformations. As for the low-level features, popular options are optic flow [7], interaction force [11], trajectories [12] or dynamic textures [10].

Unorganized bag of features are often modeled compactly using admixture or topic models, such as latent Dirichlet allocation [3], in short LDA. Topic models learn a small number of topics, which correlate semantically-related "motion patterns" within a video sequence by establishing associations between those patterns that co-occur within the same frame or clip. LDA's log-likelihood has been often employed to define a commotion measure for anomaly detection in crowded scenes and it offered very good results of far [11, 12].
Although effectiveness of LDA is not questioned, several attempts to model some of the spatial-temporal information has been made because of ignoring these natural constraints may have negative consequences in classification or clustering tasks. These attempts aimed at extending LDA or proposing alternatives to it.
According to the recent studies on natural scene recognition [14, 15, 16], the bag of features extracted from images still have an *imprint* of the image spatial structure, more importantly, some of the spatial structure can be recovered if the bags are considered together. This is achieved by a generative model called the *Counting Grid*, which jointly maps the bags in the training set into windows of size W in a grid of counts recovering some spatial structures on the grid: each location $\mathbf{k}$ of the grid is a normalized distribution over the features $\pi_{\mathbf{k},z}$ indexed by z. To illustrate this property, we considered a drawing (available in Matlab: load trees - Fig. 2-a) and we extracted 50 patches of area roughly 10% of the original image, taken at random in the image. Subsequently, we discarded all the spatial structure, describing each patch with an histogram over $Z = 64$ colors, which in this experiment acts as features. Finally, using these bags-of-colors, we learned a counting grid model. For visualization, each grid location $\mathbf{k}$ was assigned the color equal to the average of the $Z = 64$ colors in color map, weighted by the normalized local feature counts $\pi_{\mathbf{k},z}$. The result is illustrated in Fig. 2-c). Remarkably, significant amount of the spatial structure in feature distributions was reconstructed from these 50 histograms. The algorithm discovered that the dark, red and brown tones go together and that they are bordered by green. Elongated dark structures against the blue background are discovered, as is the coast/island boundary. In this sense, the counting grid provides a good model for interpolating among the original 50 histograms, as the histograms from the original image are also likely under the inferred counting grid.

Inspired by the aforementioned idea, in this paper, we advocate the use counting grids to generate compact representations of videos from which we extract commotion measure for abnormality detection. We will employ here a 3D generalization of the original counting grid, which in our experiments outperformed its 2D counterpart. This is rather intuitive as bags possess a spatio-temporal imprint, which can be recovered by a 3-dimensional counting grid if enough data is provided. Finally, we labeled new surveillance sequences which we will make them available as additional test-bed, and we thoroughly compared LDA, the original CGs and its 3D extension, varying the amount of training data, the number of overlaps between clips, and the number of patches sampled from videos to generate the bags.

The rest of the paper is organized as follows: in Sec. 2, we introduce the notation and we detail the generative models, in particular the extension of 2D "spatial" CG to the 3D "spatio-temporal" CG. In Sec. 3 we will report the experimental results and finally, we will draw conclusions in Sec. 4

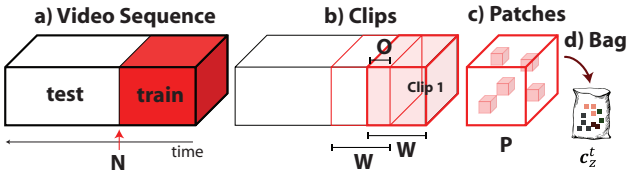

Figure 3. Notation used throughout the paper. a) We used the first N frames for training. b) We highlighted the first 2 clips, with their overlap of O frames. c-d) The process of bag formation.

2. Generative Methods

In this section we will give an overview about the generative approaches. The basic notation we will use throughout is summarized in Fig. 3. Let $\mathcal{V}$ be a video-sequence of N_{total} frames. In the experiments, we will use the first N (normal) frames are training data. Each sequence $\mathcal{V}$ is divided into a number of overlapping clips with W frames length, and O frames overlap. For each clip, we extracted P three dimensional cuboids or patches of fixed size $5 \times 5 \times 5$, we assigned to each patch a discrete codeword (feature) $z = 1 \dots Z$. Each clip is then described with an histogram over codewords $\mathbf{c}^t = \{c_z^t\}_{z=1}^Z$, where c_z^t represents the number of times the codeword z appears in the clip t. Finally, we use $\mathcal{C}^t$ for the crowd commotion measure for the t^{th} clip.
It is important to note that the notation is general and works for both frame-level ($O = W - 1$) or clip-level abnormality detection. In the experimental section we will thoroughly vary all the parameters and we will discuss the effect that they have on the three models tested.

2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation [3] has been widely used for text document analysis in order to uncover semantic topics from documents. In particular, LDA learns a set of topics β, which captures the co-occurence of the words and it mixes a limited amount of topics to explain a single bag. Under this model, the likelihood of a sample is:

$$\mathcal{L}(\mathbf{c}^t|\alpha,\beta) = \prod_{k=1}^{K} \hat{\theta}_k^{\alpha_k - 1} \cdot \prod_{z=1}^{Z} \sum_{k=1}^{K} \left(\hat{\theta}_k \cdot \beta_{k,z}\right)^{c_z^t} \quad (1)$$

where k indexes one of the K topics, $\hat{\theta}$'s are ML estimates of the topics proportions of each bag and $\alpha = \{\alpha_1, \dots, \alpha_K\}$ is the parameter vector of a Dirichlet prior imposed over θ.

In the past years, LDA has been translated and successfully employed to solve several computer vision tasks. For instance, Wang et al. [19] used LDA model to classify global behaviors through studying co-occurrences of trajectory-based motion from multi-view cameras. As main limitation, since LDA model discard the spatial structure of visual features, their method is not able temporal dynamics among activities within camera network. Mehran et al. [11] adapted the Social Force Model (SFM, [8]) to compute the interaction force later used as base feature descriptor in a standard application of the bag-of-words paradigm. LDA was then used to learn a model of normal crowd behavior patterns, thus to define the crowd commotion measure

$$\mathcal{C}_{LDA}^t = -\log \mathcal{L}(\mathbf{c}^t|\alpha,\beta) \quad (2)$$

2.2. The Counting Grid Model

The counting grid model (CG) instead of looking for co-occurrence of features, enforces that bags are generated taking windows (a "scene" or "view") in a larger scene (the "visual world" or "panorama"). The intuition upon which this model is built is that the space of all possible count combinations is constrained by the properties of the larger scene as well as by the size and the location of the window into it.

Formally, the basic counting grid π is a set of distributions over z on a 2-dimensional discrete grid indexed by $\mathbf{k} = (x, y)$. $E = (E_x, E_y)$ describes the extent of the counting grid. Since each element of the grid is a normalized distribution, $\sum_z \pi_{\mathbf{k},z} = 1$ everywhere on the grid.

A given bag $\mathbf{c}^t$ is assumed to follow the distribution of visual words found in a window $W = (W_x, W_y)$ of the counting grid. In particular, each bag can be generated by first selecting a position $\mathbf{k}$ on the grid and placing a window of dimension W in that position. Then, all counts in this window are averaged and normalized

$$h_{\mathbf{k},z} = \frac{1}{W_x \cdot W_y} \sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z} \quad (3)$$

and finally the set of features in the bag are generated from $h_{\mathbf{k},z}$. In the equation above $W_\mathbf{k}$ represent a window placed at location $\mathbf{k}$.
The position of window $\mathbf{k}$ in the grid is the only latent variable and once it is given, the probability of the bag of features can be computed as

$$\mathcal{L}(\mathbf{c}^t|\pi) = \sum_{\mathbf{k}=1}^{E} \prod_{z=1}^{Z} \left(h_{\mathbf{k},z}\right)^{c_z^t} \quad (4)$$

Computing and maximizing the log likelihood of the data is an intractable problem and to solve it the iterative EM algorithm is needed.
Starting from a random but symmetry-breaking initialization of π, the E-step aligns all bags on the grid windows computing the posterior probabilities over the mapping locations for each bag, e.g., $q_\mathbf{k}^t$, while the M-step re-estimate the counting grid π, is re-estimated, by firstly distributing

the counts within the appropriate windows, and then averaging across the samples.
Likewise the previous case, if we learn a counting grid with normal footage, we can define a generative crowd commotion measure as

$$\mathcal{C}^t_{CG} = -\log \mathcal{L}(\mathbf{c}^t|\pi) \tag{5}$$

3-Dimensional Counting Grids: Inspired by the recent success of counting grids to analyze images, we generalized the original spatial counting grids to three dimensions. In this way the model better adheres to the intrinsic characteristics of sequences and recovers some of their spatio-temporal relations among features. It is important to note, however, that a three dimensional grid in principle is not necessary. Therefore, in the experiments we will compare 2- and 3-dimensional grids.

The extension of the model straightforward and the EM algorithm stays the same, expect that now generic positions are indexed by a triplet $\mathbf{k} = (x, y, t)$ thus $E = (E_x, E_y, E_t)$. The only relevant difference is that requires a modification of the numerous sums-in-window (e.g., see Eq. 3 as well as in the M step [14]), which in the original paper were computing using Viola-Jones integral images [18]. To generalize we used *summed area tables*, a generalization of integral images for D-dimensions. In particular if we consider the corners of the hyper-window W as k_d with d in $\{0,1\}^D$, we first compute the cumulative table of $c(\mathbf{x})$,

$$\mathbf{C}(x_1, x_2, \ldots, x_D) = \sum_{\{x'_i \leq x_i\}} c(x_1, x_2, \ldots, x_D) \tag{6}$$

Then, the cumulative sum of window in grid can be computed as follows:

$$\sum_{d \in \{0,1\}^D} (-1)^{D-\|d\|_1} \cdot \mathbf{C}(k_d) \tag{7}$$

In the case of 2D the notation, the corner of a window located in the grid are $x^{(0,0)}$, $x^{(0,1)}$, $x^{(1,0)}$, and $x^{(1,1)}$ and one recovers the original update.

As further consideration we observed that the update for π given in the original paper and reported below (Eq. 8) is multiplicative, making the EM algorithm prone to slow convergence. This is especially true in higher dimensions when windows become larger

$$\pi_{\mathbf{i},z} \propto \hat{\pi}_{\mathbf{i},z} \sum_t c^t_z \cdot \sum_{\mathbf{k}|\mathbf{i}\in W_\mathbf{k}} \frac{q^t_\mathbf{k}}{\hat{h}_{\mathbf{k},z}} \tag{8}$$

To fasten the convergence we rewrite Eq. 8, to highlight the contribution of the data $f_{\mathbf{i},z}$, which can be computed once after the E-Step. This allows us to iterate the updates for

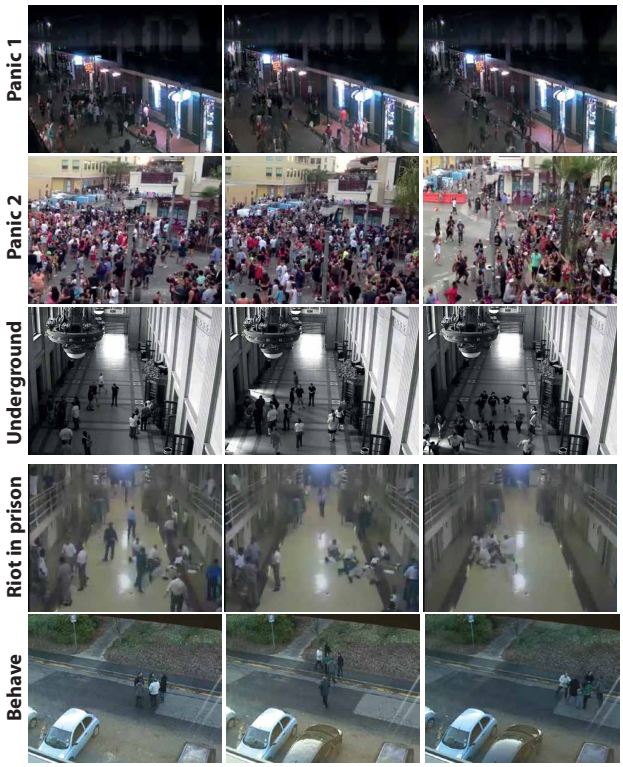

Figure 4. Sample frames from video sequences of Panic1 (first row), Panic2 (second row), the RIP (third row), the UMN (fourth row), and the Behave (fifth row). The two first column of each dataset are reflect the norma situations while the last two column reflect the panic or fight situations.

π and h, in the M-step to reach a faster convergence and a better placement of the features.

$$\pi_{\mathbf{i},z} \propto \hat{\pi}_{\mathbf{i},z} \cdot \underbrace{\left(\sum_t c^t_z \sum_{\mathbf{k}|\mathbf{i}\in W_\mathbf{k}} q^t_\mathbf{k}\right)}_{f_{\mathbf{i},z}} \cdot \sum_{\mathbf{k}|\mathbf{i}\in W_\mathbf{k}} \frac{1}{\hat{h}_{\mathbf{k},z}} \tag{9}$$

Please note that despite a loop is introduced in the M-step, the biggest computational burden is the E-step which dominates the complexity of the algorithm. This technique proved to be effective especially for large W and when a large amount of data is present (e.g., Behave dataset).
The crowd commotion of 3D-CG is the same as 2D-CG, Eq. 5.

3. Experiments

We extensively compared the generative models on two standard datasets, including UMN[1] and Behave dataset [4], and three new datasets collected from web namely Panic1, Panic2 and Riot In Prison (RIP). Some normal and abnormal frames of these dataset are shown in Fig. 4; as visible these videos are characterized by different situations

in terms of size of crowd, varying resolutions, full/partial occlusions, camera motion, illumination and abnormality.

Unless differently specified, we considered optic flow as base-feature and we divided each video-sequence in temporally overlapping clips of length $W = 15$-frames with $O = 5$-frames overlap (see Fig. 3). For each clip, we randomly sampled P patches of size $(5 \times 5 \times 5)$. Following standard procedure, we run the K-means algorithm with the training data to compute codebook; we set the number of centers to $Z = 500$. A given clip is then described with a histogram c_z^t over the entries. Finally, we learned the generative models, namely LDA, 2D-CG and 3D-CG for the task of abnormality recognition. In all the sequences we used half of the normal clips to train the models. At test time, we computed the commotion $\mathcal{C}$ of each test clip with Eq. 2 and Eq. 5 and as performance measure we record the area under the ROC curve. In all the experiments, we repeated this process 10 times averaging the results.

In all the tests, we varied the model complexity and we used AICc [17] to select one (although consistently to previous work, we did not observe much variation in the results for all the models). For LDA we varied number of topics in range of $T = \{10, 20, \ldots, 200\}$, while for counting grids we considered all the combinations between $E = \{(8,8), (10,10), (15,15), (20,20), \ldots, (50,50)\}$ and $W = \{(3,3), (4,4), (5,5)\}$.

3.1. Datasets

Panic1. We collected this sequence form youtube; it consists of 3293 frames, including 1723 normal followed by 1570 abnormal (panic situations of gun shot). The video was recorded with a fixed surveillance camera in a outdoor scene with challenging conditions of low light and perspective distortion.

Panic2. We collected this sequence form youtube; it consists with 2207 frames including 1962 frames of normal and 245 abnormal (police attack situations). The video was recorded with moving camera in outdoor with day light illumination.

Riot in Prison. We collected this sequence form youtube; it is recorded with a surveillance camera inside a prison. After several normal frames frames, we have a person-on-person fight, then the number of participants increase gradually, and finally the sequence ends again with a normal situation controlled by security guard. The dataset contains 3728 frames consists of 2568 normal and 1160 abnormal (fight situations).

UMN. This dataset is publicly available provided by University of Minnesota [1], which contains normal and abnormal panic scenarios. The original dataset comprises 11 different scenarios of three different scenes recorded under controlled circumstances with surveillance camera in outdoor and indoor scenes. In our experiment, however, we only used 6 scenarios of one indoor scene scenarios, including 3433 of 3003 and 430 normal and panic frames, respectively. We discarded the other two scenes as they are not realistic.

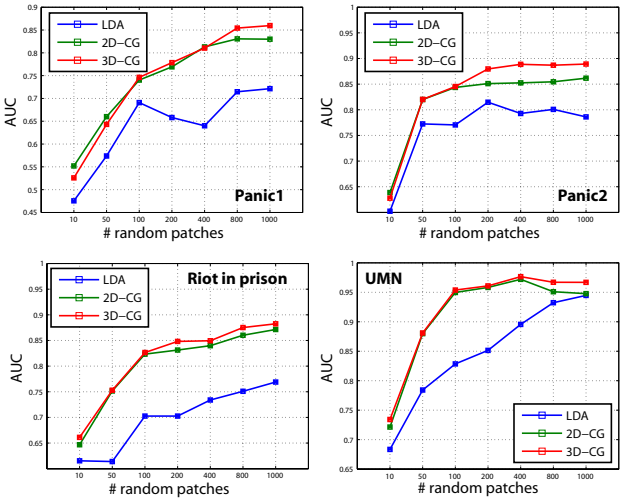

Figure 5. Comparison of average AUCs (y-axis) on Panic1, Panic2, RIP, and UMN sequences varying the number of sampled patches P (x-axis).

Behave. Behave dataset [4] contains of 9 different activities, including approaching, walking together, meeting, splitting, ignoring, chasing, following, running together, and fighting with various number of participants. We divide the dataset into two classes, considering fighting and running actions as abnormal behaviors and the rest as normal acts. In total, it contains 33373 frames (excluding frames with no individual) including 30533 of normal and 2840 of abnormal activities.

3.2. Parameter Evaluation

Effect of sample patches In the first experiment, we examine the effect of selecting number of random patches P on performance of the generative models. We varied the number of sampled patches in the range of $P \in \{10, 50, 100, 200, 400, 800, 1000\}$. The results of this setting are illustrated in Fig. 5 . The results show that the CG models outperform the LDA for all number of sampled patches over all the four datasets. This is caused by the drawback of LDA to model the spatial (temporal) relation of training bags. Moreover, AUCs obtained by 3D-CG is higher than 2D-CG, particularly when the number of sampled patches increases. This improvement caused by the ability of 3D-CG to capture the spatio-temporal structure

	Panic1			Panic2		
# of Training Clips	LDA	2D-CG	3D-CG	LDA	2D-CG	3D-CG
$N = 5$	**0.64 ± 0.015**	0.57 ± 0.002	0.58 ± 0.004	**0.73 ± 0.007**	0.68 ± 0.009	0.71 ± 0.010
$N = 40$	0.67 ± 0.110	0.71 ± 0.060	**0.73 ± 0.090**	0.76 ± 0.040	0.83 ± 0.030	**0.84 ± 0.012**
N = Half	0.71 ± 0.052	0.83 ± 0.042	**0.86 ± 0.076**	0.78 ± 0.015	0.86 ± 0.024	**0.87 ± 0.020**

Table 1. Average AUCs on Panic1 and Panic2 sequences varying the amount of training data. We also report the 95% confidence interval.

	UMN			Riot in Prison		
# of Training Clips	LDA	2D-CG	3D-CG	LDA	2D-CG	3D-CG
$N = 5$	**0.92 ± 0.540**	0.91 ± 0.180	0.90 ± 0.120	**0.59 ± 0.023**	0.51 ± 0.003	0.50 ± 0.003
$N = 40$	0.93 ± 0.020	0.93 ± 0.290	**0.94 ± 0.028**	0.74 ± 0.043	0.74 ± 0.038	0.72 ± 0.011
N = Half	0.94 ± 0.001	0.94 ± 0.018	**0.96 ± 0.027**	0.77 ± 0.052	0.87 ± 0.042	**0.88 ± 0.032**

Table 2. Average AUCs on UMN and RIP sequences varying the amount of training data. We also report the 95% confidence interval.

of bags of words, while this structure in 2D-CG is limited just to the spatial domain and the temporal relation of the bags are missed. Although we also observed that for the low number of sampled training patches, the AUC of 2D-CG and 3D-CG are very close, while for higher amount of sampled patches (e.g. more than 200 patches) the AUC of 3D-CG is fairly higher than 2D-CG. This can be justified in the way that 3D-CG requires enough training patches to model the temporal relation of bags. Otherwise, when the temporal relation of bags is not appropriately captured, 3D-CG performs same as 2D-CG.

Effect of training data In this experiment, we fix the number of random patches $P = 1000$, and we varied number of training between $N \in \{5, 40, \text{Half}\}$, where "Half" means to half of the initial normal clips (the standard).
The results are reported in Table 1 and Table 2; as expected LDA outperformed counting grids at $N = 5$. LDA in fact is equipped with priors which make is more robust to scarcity of training data. However, counting grids quickly catch up as we increase the amount of training data. For example, in Panic1, the improvement of LDA from 5 to 40 training clips is around 3% while for 2D-CG and 3D-CG is respectively around 14% and 15%. Finally, the result also shows that 3D-CG outperformed 2D-CG in all the situations.

Overlaps between clips In the third experiment we varied the amount of overlap between clips. We restricted this evaluation to "Panic2" (moving camera) and "Riot in Prison" (fixed camera) dataset, setting the number of random patches and training clips are respectively set to $P = 1000$ and the half of the normal clips and the amount of temporal overlapping varies in the range of $O \in \{0, 5, W - 1\}$. It is important to note that if $O = 0$ there is no overlap in the temporal domain, while if $O = W - 1$ we are computing a bag for each frame.
Results are reported in Table 3, and shows that increasing the amount of temporal overlapping improves the CGs performance. The reason is that counting grid recover some of the spatio-temporal structure by considering the bags jointly which must be characterized by a overlapping imprint, and increasing O help the algorithm to recover what the bags shares. This is not the case in LDA, and increasing the overlapping stride slightly degrades LDA performance. LDA in fact does not exploit the spatial/temporal structure of bags and having overlapped clips may result in information redundancy and affect its performance.
Results for the other sequences are similar.

Effect of the base feature. In this experiments we evaluate the effect of the based feature. We restrict to "Panic2" and "UMN" and we considered as base feature the interaction force [11]. Once again we set $P = 1000$, $N =$ Half. Results are presented in Table 4: regardless to the type of the base feature, the performance obtained by 3D-CG is slightly better than 2D-CG. Results for the other sequences are similar.

Comparison with state of the art on Behave dataset. As final experiment, we considered Behave dataset and we compared our generative models with the state of the art. Once again we used optic flow as base-feature and we set $P = 1000$, $N =$ Half.
Table 5 summarizes the results: Energy Potential method [5] outperforms our generative approaches which however reached very compelling result. One possible explanation for such a good performance is that [5] is a trajectory-based approach, therefore suitable for scenes with low pedestrian density as Behave. However as major drawback [5] is an SVM-based method and it *requires* abnormal data at training time. This is very difficult to have in real scenarios and we believe generative approaches should be preferred.
Among generative models, once again 3D-CG outperforms LDA and 2D-CG.

	Panic2			Riot in Prison		
Amount of overlap	LDA	2D-CG	3D-CG	LDA	2D-CG	3D-CG
$O = 0$ (no overlap)	0.77 ± 0.072	$\mathbf{0.8 \pm 0.065}$	0.79 ± 0.067	$\mathbf{0.87 \pm 0.026}$	0.85 ± 0.046	0.77 ± 0.117
$O = 5$	0.78 ± 0.062	0.86 ± 0.015	$\mathbf{0.88 \pm 0.028}$	0.77 ± 0.052	0.87 ± 0.042	$\mathbf{0.88 \pm 0.032}$
$O = W - 1$	0.93 ± 0.037	0.99 ± 0.037	$\mathbf{0.99 \pm 0.028}$	0.76 ± 0.013	0.86 ± 0.051	$\mathbf{0.87 \pm 0.089}$

Table 3. Comparison of average AUCs on Panic2 and Riot in prison sequences varying overlapping between clips. We also report the 95% confidence interval.

	UMN			Panic2		
Base feature	LDA	2D-CG	3D-CG	LDA	2D-CG	3D-CG
Optic Flow	0.94 ± 0.089	0.94 ± 0.063	$\mathbf{0.96 \pm 0.039}$	0.78 ± 0.0056	0.86 ± 0.052	$\mathbf{0.88 \pm 0.069}$
Interaction Force [11]	0.98 ± 0.023	0.98 ± 0.034	$\mathbf{0.99 \pm 0.043}$	0.91 ± 0.037	0.93 ± 0.027	$\mathbf{0.94 \pm 0.017}$

Table 4. Comparison of average AUCs on UMN and Panic2 sequences employing optical flow and interaction force [11] as base-feature. We also report the 95% confidence interval.

Method	Classifier	AUC
Interaction Force [11]	SVM	0.88
Energy Potential [5]	SVM	**0.94**
Violent Flows [7]	SVM	0.81
Method	Commotion	AUC
Optical Flow	LDA	0.90 ± 0.043
Optical Flow	2D-CG	0.91 ± 0.023
Optical Flow	3D-CG	$\mathbf{0.93 \pm 0.084}$

Table 5. Comparison of averge AUCs using optical flow on Behave dataset. We also report the 95% confidence interval.

4. Conclusions

This paper presented a comparison of crowd commotion measures from different generative models, We have evaluated the performance of generative models extensively on five datasets and have shown that 3-dimensional counting grids outperforms other generative approaches.

References

[1] Unusual crowd activity dataset of university of minnesota, availabel. 4, 5

[2] R. L. Akers. *Criminological theories: Introduction and evaluation*. Routledge, 2013. 1

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 2, 3

[4] S. Blunsden and R. Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4:1–12, 2010. 4, 5

[5] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. Abnormal detection using interaction energy potentials. pages 3161–3167, 2011. 6, 7

[6] B. S. Divakaruni and J. Zhou. Image categorization using codebooks built from scored and selected local features. 2

[7] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. pages 1–6, 2012. 2, 7

[8] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 3

[9] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. pages 856–864, 2010. 1

[10] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. pages 1975–1981, 2010. 2

[11] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. pages 935–942, 2009. 1, 2, 3, 6, 7

[12] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino. Analysing tracklets for the detection of abnormal crowd behaviors. 2015. 1, 2

[13] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar. Violence detection in video using computer vision techniques. pages 332–339, 2011. 2

[14] A. Perina and N. Jojic. Image analysis by counting on a grid. pages 1985–1992, 2011. 2, 4

[15] A. Perina and N. Jojic. Capturing spatial interdependence in image features: the counting grid, an epitomic representation for bags of features. *Transactions on Pattern Analysis and Machine Intelligence*, Accepted for pubblication, 2015. 2

[16] A. Perina, M. Zanotto, B. Zhang, and V. Murino. Location recognition on lifelog images via a discriminative combination of generative models. 2014. 2

[17] D. Posada and T. R. Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004. 5

[18] P. Viola and M. Jones. Robust real-time object detection. 2001. 4

[19] X. Wang, K. Tieu, and W. E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):56–71, 2010. 3

[20] G. Xiong, J. Cheng, X. Wu, Y.-L. Chen, Y. Ou, and Y. Xu. An energy model approach to people counting for abnormal crowd behavior detection. *Neurocomputing*, 83:121–135, 2012. 1

[21] Z. Zhong, W. Ye, S. Wang, M. Yang, and Y. Xu. Crowd energy and feature analysis. pages 144–150, 2007. 1