Towards Crowd Density-Aware Video Surveillance Applications

Hajer Fradi, Jean-Luc Dugelay

EURECOM, Campus Sophia Tech 450 route des Chappes
06410 BIOT-SOPHIA ANTIPOLIS, FRANCE

Abstract

Crowd density analysis is a crucial component in visual surveillance mainly for security monitoring. This paper proposes a novel approach for crowd density measure, in which local information at pixel level substitutes a global crowd level or a number of people per-frame. The proposed approach consists of generating automatic crowd density maps using local features as an observation of a probabilistic density function. It also involves a feature tracking step which excludes feature points belonging to the background. This process is favorable for the later density estimation as the influence of features irrelevant to the underlying crowd density is removed. Since the proposed crowd density conveys rich information about the local distributions of persons in the scene, we employ it as a side information to complement other tasks related to video surveillance in crowded scenes. First, since conventional detection and tracking methods are hard to be scalable to crowds, we use the proposed crowd density to enhance detection and tracking in videos of high density crowds. Second, we employ the local density together with regular motion patterns as crowd attributes for high level applications such as crowd change detection and event recognition. Third, we investigate the concept of crowd context-aware privacy protection by adjusting the obfuscation level according to the crowd density. In the experimental results, our proposed approach for crowd density estimation is evaluated on videos from different datasets, and the results demonstrate the effectiveness of feature tracks for crowd measurements. Moreover, the employment of crowd density in other applications demonstrate good performances for detection, tracking, behavior analysis, and privacy preservation.

Keywords: Crowd density, local features, detection, tracking, behavior analysis, privacy protection

1. Introduction

Studying crowd phenomenon is becoming of great interest mainly with the increasing number of popular events that gather many people such as in markets, subways, religious festivals, public demonstrations, sport events, and high density moving objects like car traffic. In this context, crowd analysis has emerged as a major topic for crowd monitoring and management in visual surveillance field. In particular, the estimation of crowd density is receiving much attention for safety control. It could be used for developing crowd management strategies by measuring the comfort level in public spaces. Also, its automatic monitoring is extremely important to prevent disasters by detecting potential risk and preventing overcrowd. Many stadium tragedies could illustrate this problem, as well as what happened in 2010, in the Love Parade stampede in Germany and the Water Festival stampede in Colombia. To prevent such deadly accidents, early detection of unusual situations in large scale crowd is required and appropriate decisions for safety control have to be taken to insure assistance and emergency contingency plan.

Many recent works in the field of automatic video surveillance have been proposed to address the problem of crowd density analysis. Typically, given a video sequence the objective is to estimate the number of people, or to alternatively estimate the crowd level. For people counting problem, significant progress has been recently made to handle that by using features regression methods [1, 2, 3]. This paradigm is proposed as an alternative solution to detection-based methods because of the partial occlusions that occur in the crowd, and that make delineating people a difficult task. In addition to person counts, level of the crowd is another indicator in crowd density analysis. According to the classification introduced in [4], the crowd density can be categorized into 5 levels: free, restricted, dense, very dense, and jammed flow. Early attempts to handle this problem generally made use of local texture features. Especially the use of some variants of Local Binary Pattern (LBP) [5], has been an active topic of research

for this problem [6, 7, 8, 9].

Although these categories of people counting and crowd level classification are commonly used in the field of crowd analysis, they have the limitation of providing a global information of the whole image, and discarding local information about the crowd. We therefore resort to crowd measure at local level by computing crowd density maps. This alternative solution is indeed more appropriate since it enables both the detection and the localization of potentially crowded areas. The proposed crowd density map is based on using local features as an observation of a probabilistic density function. Also, a feature tracking step is involved in the estimation of crowd density. In fact, considering all extracted local features brings an inconvenience to the density function estimation as a substantial amount of components are irrelevant to the underlying crowd density. Therefore, we propose to use motion information to alleviate this effect.

In addition to the estimation of local crowd density, we intend to explore in this paper the usefulness of such crowd measure as additional information to other video surveillance tasks, mainly because common capabilities of automated surveillance systems are of limited success in high-density scenes. This is due to the challenging characteristics of crowded scenes such as the small size of objects in crowds, the occlusions caused by inter-object interactions, and the constant interactions among individuals in the crowd which make them indiscernible from each other. Given these difficulties, visual analysis of high density scenes remains a challenge compared to scenes with fewer people. As the density of people increases in the scene, a substantial deterioration in performances of automatic video surveillance tasks such as person detection, tracking, and behavior analysis is observed [10]. In this paper, we mainly focus on three major representative set of problems which are: (1) detection and tracking of people in crowded scenes, (2) modeling crowd behaviors and detecting anomaly (or change), and (3) studying privacy aspects in crowds.

The problems of detection and tracking in crowds have been addressed in the literature by learning motion patterns in order to constraint the tracks. In [11], global motion patterns are learned and participants of the crowd are assumed to follow a similar pattern. Rodriguez *et al.* [12] extended this approach in unstructured environments

to cope with different crowd behaviors by studying overlapping motion patterns. Although these solutions have shown promising results, they operate in off-line mode and the learned patterns are tied to a particular scene. Also, they impose constraints to the crowd motion, thus, trajectories not following the common patterns are penalized. Moreover, some of these methods include additional constraints; in [12], Rodriguez *et al.* employed a limited descriptive representation of target motion by quantizing the optical flow vectors into 10 possible directions. Also, the *floor fields* proposed in [11] impose how a pedestrian should move based on scene constraints, which results in only one single direction at each spatial position in the video.

Crowd behavior analysis is another problem that has attracted research attention in the field. This problem covers different subproblems such as crowd change or anomaly detection [13, 14, 15], and crowd event recognition [16, 17, 18]. Usually the activity process in video sequence can be categorized into three main steps [16]: (1) detection, (2) tracking, and (3) behavior analysis. Given the difficulties encountered by analyzing crowded scenes, related works to crowd behavior analysis bypass the detection and the tracking of individuals and instead operate on local features [15], or particles [14, 17]. In general, these methods aim at detecting and categorizing crowd events using motion information. This latter could correspond to normal (frequent) behavior or abnormal (unusual) behaviors.

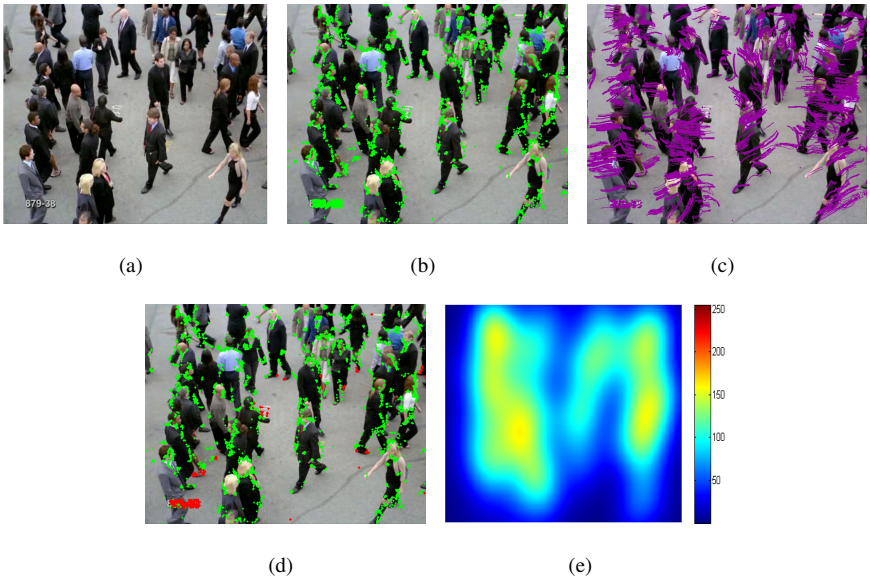
The last problem we intend to address in this paper, is about preserving privacy in crowded scenes. Actually, with the widespread growth in the adoption of digital video surveillance systems, several concerns have been raised related to the possibility of infringing the privacy rights of the subjects being monitored [19]. At the same time, the adoption of automated methods for the analysis of video surveillance data has raised additional concerns, since algorithms such as face recognition or people re-identification could potentially expose the identity of any individual under video surveillance at any time [20]. One big challenge related to privacy protection policies in crowded scenes is the identification of the correct trade-off between intelligibility of the video, which should be adequate for crowd monitoring tasks, and privacy protection itself. An attempt to deal with this problem is presented in [21], where a context-aware surveillance system is proposed by combining a number of contextual

information (based on the analysis of visual features such as global motion, and person counts) to determine an appropriate level of privacy protection.

To overcome all these problems, in this paper, we propose to incorporate the local crowd density measure in the three aforementioned applications: First, we propose a method for enhancing human detection and tracking in crowded scenes; it is based on applying a scene-adaptive dynamic parametrization using the crowd density measure. Compared to prior works, our approach does not depend on any learning step, and does not impose any direction to the crowd flow. It models the crowd in a temporally evolving system, which enables a large number of likely movements at each space-time location of the video. Second, we propose a novel approach to detect crowd change and to recognize crowd events. It is based on analyzing temporal and spatial distributions of persons using long-term trajectories within a sparse feature tracking framework. Our proposed approach employs the local density together with the commonly used motion patterns (speed and direction). The idea is motivated by the necessity of using local density to determine the ongoing crowd behavior since that helps to characterize the event, and to localize crowded regions. Finally, we investigate the usefulness of applying crowd density in privacy context. The concept of context-aware privacy protection has recently emerged, as the required amount of privacy protection is deeply linked to the context. In particular, we propose adaptive protection filters that select suitable level of privacy preservation according to the crowd density measure.

The remainder of the paper is organized as follows: In the next Section 2, we present our proposed approach for crowd density map estimation. In Section 3, we demonstrate how a prior estimation of crowd density could provide valuable information and could complement other applications in video surveillance. In particular, three applications are investigated: enhancing human detection in crowded scenes is presented in Section 3.1, studying crowd behaviors is presented in Section 3.2, and formulating contextualized privacy preservation filters is presented in Section 3.3. Detailed experimental results of the density map and the different applications follow in Section 4. Finally, we briefly conclude and give an outlook for possible future works in Section 5.

2. Crowd Density Map Estimation



(a) (b) (c)

(d) (e)

Figure 1: Illustration of the proposed crowd density map estimation using local features tracking: (a) exemplary frame, (b) FAST local features (c) feature tracks (d) distinction between moving (green) and static (red) features - red features at the lower left corner are due to text overlay in the video (e) estimated crowd density map

In this paper, we explore a new promising research direction which consists of using crowd density measures to complement some other applications in crowded scenes. For this, generating local crowd density measure is more helpful than computing only an overall density or a number of people in a whole frame. In the following, we present our proposed approach for crowd density estimation [22]. First, local features are extracted to infer the contents of each frame under analysis. Then, we perform local features tracking using the Robust Local Optical Flow algorithm from [23] and a point rejection step using forward-backward projection. To accurately represent the motion within the video, the estimation of the optical flow between consecutive frames is extended to trajectories. The generated feature tracks are thereby used to remove static features. Finally, crowd density maps are estimated using Gaussian symmetric kernel function. An illustration of the density map modules is shown in Figure 1. The remainder of this

section describes each of these system components.

2.1. Extraction of local features

One of the key aspects of crowd density measurements is crowd feature extraction. Under the assumption that regions of low crowd density tend to present less dense local features compared to high-density crowd, we propose to use local features as a description of the crowd by relating dense or sparse local features to the crowd size. Thus, the proposed crowd density map is estimated by measuring how close local features are.

For local features, we assess Features from Accelerated Segment Test (FAST) [24]. The reason behind selecting this feature for crowd measurement is as follows: FAST has been proposed for corner detection in a reliable way. It has the advantage of being able to find small regions which are outstandingly different from their surrounding pixels. In addition, FAST was used in [25] to detect dense crowds from aerial images and the derived results demonstrate a reliable detection of crowded regions.

2.2. Local features tracking

Using the extracted features to estimate the crowd density map without a feature selection process might incur two problems: First, the high number of local features increases the computation time of the crowd density. As a second and more important effect, the local features contain components irrelevant to the crowd density. Thus, we need to add a separation step between foreground and background entities to our system. This is done by assigning motion information to the detected features. Based on the assumption that only persons are moving in the scene, these can then be differentiated from background by non-zero motion vectors.

Motion estimation is performed using the Robust Local Optical Flow (RLOF) [23] [26], which computes accurate sparse motion fields by means of a robust norm[1]. A common problem in local optical flow estimation is the choice of feature points to be tracked. Depending on texture and local gradient information, these points often do not lie on the center of an object but rather at its borders and can thus be easily affected by

[1] www.nue.tu-berlin.de/menue/forschung/projekte/rlof

other motion patterns or by occlusions. While RLOF handles these noise effects better than the standard Kanade-Lucas-Tomasi (KLT) feature tracker [27], the process is still not prone against all errors. This is why we establish a forward-backward verification scheme where the resulting position of a point is used as input to the same motion estimation step from the second frame towards the first one. Points for which this "reverse motion" does not result in their respective initial position are discarded. For all other points, motion information is aggregated to form trajectories by connecting motion vectors computed on consecutive frames. This results in a set of p_k trajectories at each frame k:

$$\mathcal{T}_k = \{T_1^k, ..., T_{p_k}^k| \\ T_i^k = \{X_i(k - \Delta t_i^k), Y_i(k - \Delta t_i^k), ..., X_i(k), Y_i(k)\}\} \quad (1)$$

where Δt_i^k denotes the temporal interval between the start and the current frames of a trajectory T_i^k. $(X_i(k - \Delta t_i^k), Y_i(k - \Delta t_i^k))$, and $(X_i(k), Y_i(k))$ are the coordinates of the feature point at its start and current frames respectively. The advantage of using trajectories in our system instead of computing the motion vectors only between two consecutive frames is that outliers are filtered out and the overall motion information is more reliable and less affected by noise.

2.3. Kernel density estimation

After generating trajectories, our goal is to remove static features. It proceeds by comparing the overall mean motion Γ_i^k of a trajectory T_i^k to a certain threshold ζ which is set according to image resolution and camera perspective. Moving features are then identified by the relation $\Gamma_i^k > \zeta$ while others are considered as part of the static background. We mean by "others", trajectories of zero length because they are stationary along frames, or of small length because of the noise in video acquisition or dynamic background. Using trajectories, the separation between foreground and background entities is improved and the number and positions of the tracked features undergo an implicit temporal filtering step which makes them smoother.

After filtering out static features, the crowd density map is defined via kernel density estimate based on the positions of moving local features. Starting from the as-

sumption of a similar distribution of feature points on the objects, the observation can be formulated as: the more local features come towards each other, the higher crowd density is obtained. For this purpose, a probability density function (pdf) is estimated using a Gaussian kernel density. At a frame I_k, if we consider a set of m_k moving local features extracted at their respective locations $\{(x_i, y_i), 1 \leq i \leq m_k\}$, the corresponding density map C_k is defined as follows:

$$C_k(x,y) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{m_k} \exp -(\frac{(x-x_i)^2 + (y-y_i)^2}{2\sigma^2}) \tag{2}$$

where σ is the bandwidth of the 2D Gaussian kernel. The resulting crowd density map characterizes the spatial distributions of pedestrians in the scene which could complement others tasks in crowd analysis.

3. Crowd Density-Aware Applications

In this section, we intend to demonstrate how the proposed density presented in the last section, could provide valuable information and complement other tasks related to crowd analysis. Precisely, three applications are investigated, see Figure 2.

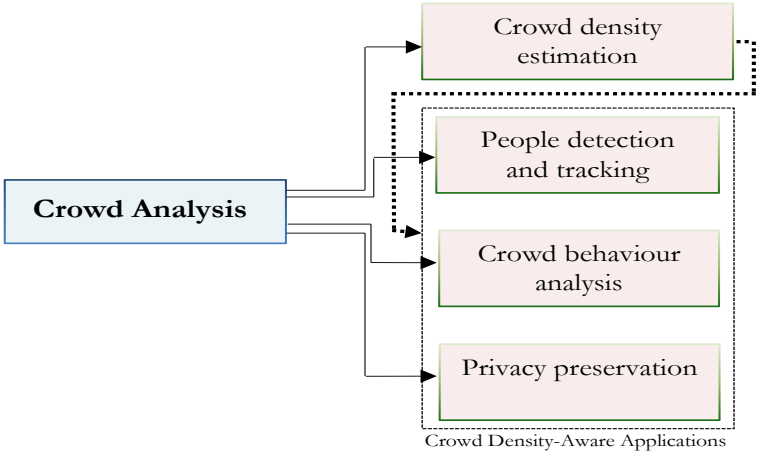


Figure 2: Schematic representation of the topics tackled in this paper in crowd analysis field. The dotted line shows that the crowd density map is used to complement other tasks (people detection and tracking, crowd behavior analysis, and privacy preservation).

3.1. Enhancing human detection and tracking in crowded scenes

Human detection is a common problem in computer vision as it is a key step to provide semantic understanding of video data. Accordingly, it has been intensively studied and different approaches have been proposed. In this context, the deformable part-based models [28] has recently shown good performances. It is an enriched version of Histograms of Oriented Gradients (HOG) [29], that achieves much more accurate results and represents the current state-of-the-art.

Although this human detector has become a quite popular technique, its extension to crowded scenes is of limited success. In fact, the density of people substantially affects their appearance in video sequences. Especially in dense crowds, people occlude each other and only some parts of each individual are visible. Therefore, accurate detection in such scenarios with dynamic occlusions and high interactions among the targets remains a challenge. In order to adapt the detector to such situations, it is important to include additional information about the crowd.

In this section, we present our proposed extension of human detection to crowded scenes. As a major improvement, we propose to employ the crowd density map described in Section 2 as context information to adaptively optimize the behavior of the human detector by selecting dynamic detection threshold. This is especially important in heterogeneous scenes with crowded and non-crowded regions since the detection results are highly dependent on the crowd size (i.e. the higher the crowd density is, the more difficult is to detect persons). As a result, low detection thresholds would be suitable in crowded scenes and higher values ensure less false positives in non-crowded spaces. It is therefore desirable to find a way of automatically setting the detection threshold according to the probability that people are present in a certain position of the image.

For a given threshold τ, $\mathcal{D}_k(\tau) = \{d_1^k, ..., d_{n_k}^k\}$ denotes a set of candidate RoIs at a frame k, d_j^k is the j^{th} detection at this frame and is defined as $d_j^k = \{x_j^k, y_j^k, w_j^k, h_j^k\}$, where (x_j^k, y_j^k) is the upper left position and w_j^k, h_j^k are the respective width and height. Using a pre-defined range of detection thresholds given by upper/lower boundaries τ_{max}/τ_{min}, we apply the following linear density-scaled adaptive rule to automatically

select acceptance threshold value of the detector:

$$\tau_{dyn} = \tau_{min} + (\tau_{max} - \tau_{min}) \cdot \hat{C}_k(d_j^k), j \in \{1...n_k\} \tag{3}$$

with

$$\hat{C}_k(d_j^k) = \frac{\sum_{p=0}^{h_j^k-1} \sum_{q=0}^{w_j^k-1} C_k(x_j^k + p, y_j^k + q)}{w_j^k \cdot h_j^k} \tag{4}$$

as the average crowd density of a detection d_j^k.

To obtain the dynamic threshold τ_{dyn} for every candidate d_j^k in $\mathcal{D}_k(\tau_{max})$, the average crowd density $\hat{C}_k(d_j^k)$ is computed as in (4) and inserted into (3) for all regions.

In addition to the crowd context constraints, we propose applying geometrical constraints in a filtering step. This is important due to the nature of the part-based model that may comprise certain human parts from *different* persons and match them together in *one* candidate RoI. If the score of such detection is higher than the scores of the individual objects' detections, the non-maximum suppression (NMS) step will keep it instead of the correct individual detections which might be recognized otherwise. Accordingly, in this case a false positive detection and a number of missed detections are generated which decrease the detection performance. To filter out inaccurate detections of inappropriate size, we propose to apply geometry-based pre-filters using the perceived height and the aspect ratio.

Since the perceived size of a person in a given image is affected by perspective distortions, we design a filter that uses the height of a candidate RoI to indicate the likelihood of human presence. Also, as some detections could include multiple persons at once, we propose to use the aspect ratio as a correction measure. Given a set of candidate RoIs $\mathcal{D}_k$, following [30] we assume that the relationship between a person's position and his/her perceived height to be:

$$h_j^k = \alpha_{k-1} \cdot y_j^k + \beta_{k-1}, j \in \{1...n_k\} \tag{5}$$

where α_{k-1} and β_{k-1} parameters are computed using a standard regression. Also, the aspect ration is defined as:

$$\gamma_{k-1} = median \left\{ \frac{w_j^i}{h_j^i} \right\}_{1 \le i \le (k-1), 1 \le j \le n_i} \tag{6}$$

α_{k-1}, β_{k-1}, and γ_{k-1} parameters are computed over all accepted detections $\{\mathcal{D}_1, ..., \mathcal{D}_{k-1}\}$ and are updated at each frame.

These proposed correction filters use the previous detections to predict the height and the ratio of a new detection candidate, allowing the algorithm to operate on-line without any preliminary learning step. By applying these two geometrical filters simultaneously, a detection candidate is accepted only if it fits the aspect ratio and the height according to the y-coordinate of its center. As the used NMS step is greedy and overlap-oriented, it is now possible to filter out any unlikely large or small region and to detect other objects in the same area which would have been suppressed otherwise.

3.2. Crowd behavior analysis

To achieve an improved overall performance for crowd behavior analysis, we consider that the crowd density measure could provide rich source of information about the spatial distributions of persons in the scene, mainly to localize and to recognize crowd events such as evacuation, crowd formation, and splitting. Therefore, in our approach we simultaneously consider these both cues of crowd dynamics: appearance (density) and motion (velocity, and direction).

To achieve this goal, the feature tracks (defined in (1)) used in a first step to estimate crowd density maps, are employed in a second step to extract crowd motion information. This extraction proceeds as follows: we consider only long-term trajectories, while other short-term trajectories of small length (because of tiny movement of crowd) are filtered out to not affect the computation of speed and orientation. Once the set of useful trajectories is determined, we restrict the history of each 2D trajectory over last few frames. Without such restriction, an augmentation in the speed will not be early detected, also the flow direction could be less precise. After that, the speed is computed as the quotient of the trajectory length divided by the number of frames being tracked. For flow direction, we consider the orientation of motion vectors formed by the start and the current position of each trajectory.

Overall, the spatio-temporal crowd measures introduced by density maps and motion vectors give fundamental information about the distributions and the movements of pedestrians in the scene which are strongly related to their behaviors. To model the

crowd, we encode each attribute by 1D-histogram. Given the crowd density map C_k at a frame k, the local density is quantized into N_d bins. We have chosen $N_d = 5$ according to the definition of 5 crowd levels [4]. Then, to group together motion vectors of the same direction, we quantize the orientation Θ into N_Θ bins. N_Θ is set to 8 bins, resulting in orientation bin of size $\Delta_\Theta = 45$ degrees. As proposed in [13], the speed is quantized into $N_s = 5$ classes: very slow, waking, walking fast, running, and running fast. Also, since speed changes can be affected by perspective distortions (due to the fact that when people are getting away from the camera, their motion vectors are of smaller lengths), we rectify these effects in the computation of speed.

After modeling crowd attributes by histograms, their application to crowd behavior analysis is demonstrated in three steps: First, the variation in time of a stability measure (using the histograms) is employed to detect changes or abnormal event, see paragraph 3.2.1. Second, a feature vector concatenating these histograms is used for event recognition, see paragraph 3.2.2. Third, the variations of these crowd attributes in time are used to characterize crowd events, see paragraph 3.2.3.

3.2.1. Crowd change detection

According to the procedure described so far, at each frame k, we have $H_d(k)$, $H_\Theta(k)$, and $H_s(k)$ which denote, respectively, the histograms of density, orientation, and speed. If a change occurs in the crowd behavior, that would generate dissimilarities between the histograms. For this, we compare histograms in time following the same strategy as in [13]: we compute the temporal stability $\sigma_i(k)$ of each histogram $H_i(k)$ as the weighted average of a similarity vector $S_i(k)$:

$$\sigma_i(k) = \omega^T S_i(k),$$
$$\omega = \frac{1}{\sum_{j=1}^{n} e^{\lambda \Delta t_j}} (e^{-\lambda \Delta t_1}, e^{-\lambda \Delta t_2}, ..., e^{-\lambda \Delta t_n}) \quad (7)$$

λ denotes the decay constant, $\Delta t_j = j\Delta t$ (Δt is a constant). $S_i(k)$ is computed using histogram correlation metric between each histogram $H_i(k)$ and histograms of n previous frames $H_i(k - \Delta t_1)$, ..., and $H_i(k - \Delta t_n)$.

In our approach, a change is detected if the temporal stability for one crowd attribute is low. For this, we compare each temporal stability $\sigma_i(k)$, $1 \leq i \leq 3$ to an

adaptive threshold $\tau_i(k)$ computed as the half average of σ_i between $(k - \Delta t_1)$ and $(k - \Delta t_n)$:

$$\tau_i(k) = \frac{1}{2n} \sum_{j=1}^{n} \sigma_i(k - \Delta t_j) \tag{8}$$

3.2.2. Crowd event recognition

The proposed crowd attributes are also used to recognize crowd events. In particular, 6 crowd events are tested namely, walking, running, evacuation, local dispersion, crowd formation and crowd splitting. In testing step, given a new frame $\mathbf{x}$, we aim at classifying it into one of the events $v^* \in \mathcal{V}$, which maximizes the conditional probability:

$$v^* = \arg\max_{v \in \mathcal{V}} P(v|\mathbf{x}, \theta^*) \tag{9}$$

where θ^* are learned from the training data. This can be performed by SVM classification, for the feature vector, we concatenate the 3 histograms $H_d(k)$, $H_\Theta(k)$, and $H_s(k)$ into $\mathcal{H}_k$. For classification, we use Chi-Square kernel:

$$K(\mathcal{H}_i, \mathcal{H}_j) = \sum_I \frac{\mathcal{H}_i(I) - \mathcal{H}_j(I))^2}{\mathcal{H}_i(I) + \mathcal{H}_j(I)} \tag{10}$$

3.2.3. Crowd event characterization

Local density is an important cue to characterize crowd events; it provides additional information about the density of people participating to a detected event, and it enables the localization of the event as well. The characterization of crowd events is as follows:

Walking/Running: Walking event corresponds to a number of persons moving at low speed. If the speed is high, running event is detected.

Evacuation: Evacuation is defined as a sudden dispersion of the crowd in different directions. To recognize this event, direction, speed, and crowd density attributes can be used. This event is characterized by detecting more than 4 principal directions which have to be distant from each others. Also, a degradation in the crowd density and an increase in the speed and in the motion area have to be detected.

Crowd formation/Splitting: Crowd formation (or merging) event is recognized when we detect a merge of many individuals coming from different directions towards the

same location. For this purpose, distance between main directions can be used. Also, this event is characterized by an increase in the crowd density and a decrease in the motion area. The opposite of crowd formation is splitting event.

Local dispersion: This event is recognized when people moves locally away from a threat. The same attributes of crowd formation and splitting can be employed.

3.3. Improving the compliance between privacy and surveillance

In this section, we propose to apply the crowd density measure described in Section 2 in privacy context by adjusting the level of privacy protection according to the local needs. The crowd density is selected as a criterion for privacy protection for the following reasons: crowded areas have to be constantly monitored as they are common places for crimes or for dangerous overcrowding situations. At the same time, people in a crowd exhibit a smaller amount of information to a video operator, thus they do not have to be filtered by the same degree as for an isolated person who is entirely visible. We therefore propose to lower the level of privacy protection within a crowded area compared to non-crowded area.

A simple way to do that could be to use the crowd density map as input to choose the obfuscation level. Since this method could substantially decrease the visibility of potentially important information because all crowded areas would be obscured, we restrict the application of privacy preservation filters to some regions of interest, i.e. only regions that contain personal information are obfuscated. These could include face, clothing, skin/hair color or even gait depending on the scene context. Given this variety and considering that these information is not perceivable under all circumstances (e.g. heavy crowding, different lighting conditions, distance, low resolution...), in our work we consider head obfuscation as the most visible part of a human in a crowd. However, once a person leave the crowd and is perceived as an isolated subject, more information such as clothing or skin color has to be hidden from the viewer.

The flowchart of the proposed contextualized privacy protection filters is shown in Figure 3. First, for RoIs detection step, we employ the extension of the part-based models to crowded scenes described in Section 3.1 (dotted line in Figure 3). Then, for people obfuscation, we apply adaptive privacy preservation filters to the head part or to

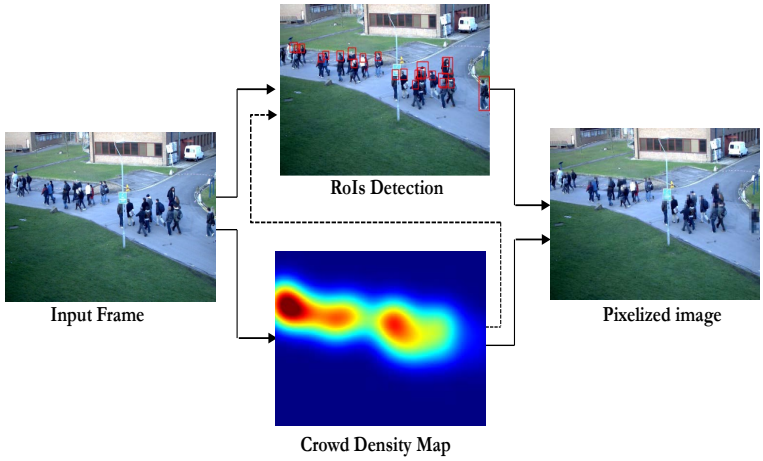


Figure 3: Flowchart of the proposed contextualized privacy preservation filters using an examplary frame from PETS [31], the dotted line in this figure shows that the crowd density map is also used to improve the robustness of the detection in crowded scenes.

the whole body depending whether the target is isolated or within the crowd. Finally, the level of privacy protection is adapted according to the crowd density. Given a set of filter parameters representing different obfuscation levels $P = \{P_{min}, ..., P_{max}\}$, for each detection d_j^k, its average crowd density value $\hat{C}_k(d_j^k)$ is used to choose the respective filter parameter that has to be applied.

As the visibility of a person in the scene is also sensitive to his/her distance from the camera because of perspective effects, we use this distance as second parameter to choose a suitable obfuscation level. A simple method to approximate the distance is to use the resolution of the detected bounding box. Since this information could be subject to errors, a more accurate method is to compute the aspect ratio and the perceived height of a person from all accepted detections (this information can be obtained from the detection step). Using this method, we are able to predict the height $\widetilde{h}_j^k$ and the ratio γ_{k-1} of a detection from the previous detections. Thus, the estimated size of a bounding box d_j^k is $\widetilde{S}_j^k = (\widetilde{h}_j^k)^2 * \gamma_{k-1}$ which is more robust than $w_j^k * h_j^k$.

In this work, we use two typical privacy protection filters which are:

3.3.1. Gaussian Blurring

This privacy filter essentially consists of removing details in a region of interest by applying Gaussian low pass filtering.

$$I^k_{blur}(x,y) = I_k(x,y) * \frac{1}{2\pi {\sigma_{k,j}}^2} e^{\frac{(x^2+y^2)}{2{\sigma_{k,j}}^2}} \tag{11}$$

The bandwidth $\sigma_{k,j}$ of the Gaussian is adapted according to the crowd density level and the predicted size.

3.3.2. Pixelization

This filter is based on decreasing the resolution of any region of interest by replacing each block of pixels in this area with its respective average.

$$I^k_{pix}(x,y) = \frac{1}{b^2_{k,j}} \sum_{i=0}^{b_{k,j}-1} \sum_{j=0}^{b_{k,j}-1} I\left(\left\lfloor \frac{x}{b_{k,j}} \right\rfloor + i, \left\lfloor \frac{y}{b_{k,j}} \right\rfloor + j\right) \tag{12}$$

As for the blurring process, the filter size $b_{k,j} \propto (\hat{C}_k(d^k_j), \widetilde{S}^k_j)$.

4. Experimental Results

4.1. Results of crowd density estimation

The proposed crowd density map is evaluated within challenging crowded scenes from multiple video datasets. In particular, we select some videos from PETS [31], UCF dataset [32], and the Data Driven Crowd Analysis dataset [33]. Regarding the nature of the videos, PETS sequences are all taken from the same view (View 1), however, they still pose different problems such as lighting conditions, shadows, and different crowd densities between the test sequences. The UCF-879 sequence is even more challenging due to higher crowd density and the tilted camera view. For the INRIA 879-38 sequence, the camera view is almost completely downward and people are walking very near to the camera. The proposed method was developed using an Intel Core i5-2500 CPU, 8 Go of RAM, Windows 7 running PC. The software used to perform the experiments was Matlab. The calculation of the density map per frame takes 2.71 seconds (for 509 local features).

For evaluating crowd density maps, the following methodology is adapted: we consider that accurate estimation of density maps can adequately represent the spatial distributions of people in the scene. For this purpose, we define the ground truth density function as a kernel density estimate based on annotated person detections. Then, we assume that an optimal feature representation can be produced by simple linear weighting of the ground truth density. Hence, given a set of annotated detections $\phi_k = \{\varphi_1^k, ..., \varphi_{l_k}^k\}$, $\varphi_i^k = \{xc_i^k, yc_i^k, h_i^k, w_i^k\}$, where (xc_i^k, yc_i^k), h_i^k, w_i^k denote, respectively, the center coordinates, the height, and the width. The corresponding ground truth density G_k is defined as:

$$G_k(x,y) = \sum_{i=1}^{l_k} \frac{1}{\sqrt{2\pi}\sigma_i^k} \exp -(\frac{(x-xc_i^k)^2 + (y-yc_i^k)^2}{2{\sigma_i^k}^2}) \tag{13}$$

σ_i^k corresponds to the size of the bounding box φ_i^k.

Given the estimated density maps $\{C_1, ..., C_N\}$ and their corresponding ground truth density maps $\{G_1, ..., G_N\}$, we aim at estimating the linear transformation mapping C_i to G_i, $1 \leq i \leq N$, with the least mismatches between them. The parameter vector Ω of this linear transformation [34] is defined as:

$$\begin{gathered} \Omega = \underset{\omega}{\operatorname{argmin}}(\omega^T\omega + \lambda \textstyle\sum_{i=1}^{N} Dist(G_i(.), C_i'(.|\omega))), \\ C_i'(.|w) = w^T C_i(.) \end{gathered} \tag{14}$$

where λ is a scalar hyperparameter controlling the regularization strength while $Dist$ is the distance measuring the loss. Since we aim at evaluating the local distribution of density, an appropriate choice of $Dist$ could be an L_p metric, which turns (14) to a typical linear regression problem, see Figure 4.

The evaluation is performed using MAE (mean-absolute-error) between the ground truth densities G_k and the estimated densities C_k' after applying linear transformation. Additionally, we split the image regions to crowd ($\mathcal{C}$) / non-crowd ($\overline{\mathcal{C}}$) regions using the reference image and the ground truth density. In Table 1, we report the results in terms of normalized MAE to the range of data in order to ensure scale-independence. In particular, three evaluation metrics E, $E_{\mathcal{C}}$ and $E_{\overline{\mathcal{C}}}$, are computed which denote normalized MAE in the whole video, in crowded areas, and non-crowded areas, respec-

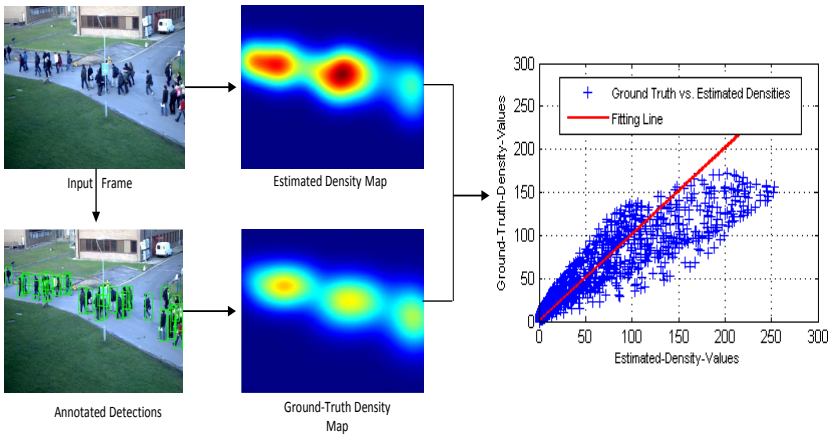


Figure 4: Flowchart of the evaluation methodology of crowd density map: The ground truth density is estimated using annotated person detection. These ground truth values are plotted vs. the estimated density values to approximate the linear transformation mapping the estimated to the ground truth values.

tively. Also, a comparison of FAST to other local features (namely Scale-Invariant Feature Transform (SIFT) [35], Good Features to Track (GFT) [36], Maximal Stable Extremal Regions (MSER) [37], and Speeded Up Robust Features (SURF) [38]) is shown and values for a GMM-based crowd density-estimation (which consists of substituting the feature-tracking step by foreground segmentation) are given. These comparisons clearly show that the feature tracking step achieves substantial improvement over using foreground segmentation. That highlights the advantage of using trajectories; our estimate is more robust to noise and the overall motion is more accurate. As a result, the number and position of the tracked features undergo an implicit temporal filtering step which improves the consistency compared to the separation between foreground and background entities.

By comparing different local features, the results show that the choice of local features in general have limited impact on the performance if we consider all image regions, even if a small improvement of FAST features is noted compared to other features. However, a more significant margin in the results between FAST and the other features is shown in crowded regions (using E_C quality metric) which demonstrates good performance of FAST for crowd measurements. For this reason, this feature will

sequence name	E	$E_{\overline{\mathcal{C}}}$	$E_{\mathcal{C}}$
S1.L1.13-57 (FAST):	**0.07** / 0.20	0.05 / 0.18	**0.30** / 0.44
S1.L1.13-57(SIFT):	**0.07** / 0.15	0.05 / 0.13	0.32 / 0.38
S1.L1.13-57 (GFT):	0.08 / 0.17	0.06 / 0.14	0.34 / 0.40
S1.L1.13-57 (MSER):	**0.07** / 0.15	0.05 / 0.13	0.36 / 0.40
S1.L1.13-57 (SURF):	**0.07** / 0.21	**0.04** / 0.19	0.36 / 0.47
S1.L1.13-59 (FAST):	**0.04** / 0.12	**0.04** / 0.11	**0.13** / 0.30
S1.L1.13-59 (SIFT):	**0.04** / 0.09	**0.04** / 0.09	0.18 / 0.27
S1.L1.13-59 (GFT):	**0.04** / 0.11	**0.04** / 0.10	0.18 / 0.31
S1.L1.13-59 (MSER):	**0.04** / 0.12	**0.04** / 0.11	0.30 / 0.37
S1.L1.13-59 (SURF):	0.05 / 0.13	**0.04** / 0.13	0.39 / 0.44
S1.L2.14-31 (FAST):	0.09 / 0.24	0.07 / 0.21	**0.21** / 0.41
S1.L2.14-31 (SIFT):	0.09 / 0.20	0.07 / 0.17	0.24 / 0.41
S1.L2.14-31 (GFT):	0.10 / 0.22	0.08 / 0.18	0.27 / 0.41
S1.L2.14-31 (MSER):	**0.07** / 0.18	0.05 / 0.14	0.26 / 0.43
S1.L2.14-31 (SURF):	**0.07** / 0.20	**0.04** / 0.16	0.26 / 0.44
S2.L3.14-41 (FAST):	0.04 / 0.23	0.03 / 0.20	0.23 / 0.54
S2.L3.14-41 (SIFT):	**0.03** / 0.17	**0.02** / 0.13	0.21 / 0.60
S2.L3.14-41 (GFT):	**0.03** / 0.18	**0.02** / 0.15	0.21 / 0.58
S2.L3.14-41 (MSER):	**0.03** / 0.11	**0.02** / 0.07	0.19 / 0.69
S2.L3.14-41 (SURF):	**0.03** / 0.14	**0.02** / 0.10	**0.18** / 0.66
UCF-879 (FAST):	**0.10** / 0.28	0.10 / 0.28	**0.09** / 0.23
UCF-879 (SIFT):	0.26 / 0.37	0.25 / 0.36	0.33 / 0.38
UCF-879 (GFT):	0.14 / 0.31	0.14 / 0.31	0.17 / 0.33
UCF-879 (MSER):	0.15 / 0.42	0.14 / 0.42	0.25 / 0.41
UCF-879 (SURF):	**0.10** / 0.47	**0.08** / 0.47	0.21 / 0.47
INRIA-879-42(FAST):	**0.11** / 0.36	0.09 / 0.38	**0.21** / 0.30
INRIA-879-42 (SIFT):	0.16 / 0.33	0.13 / 0.34	0.28 / 0.31
INRIA-879-42 (GFT):	0.13 / 0.34	0.10 / 0.36	0.24 / 0.31
INRIA-879-42 (MSER):	0.12 / 0.37	0.12 / 0.39	**0.21** / 0.38
INRIA-879-42 (SURF):	**0.11** / 0.34	**0.08** / 0.35	0.23 / 0.36

Table 1: Results of crowd density estimation for five different local feature types (FAST, SIFT, GFT, MSER and SURF) and for different test videos in terms of normalized MAE (E, $E_{\mathcal{C}}$ and $E_{\overline{\mathcal{C}}}$). Val1/Val2 are the results of our proposed approach using feature tracks, and the results using GMM foreground segmentation.

be used in the experiments for the three following applications.

4.2. Results of person detection and tracking

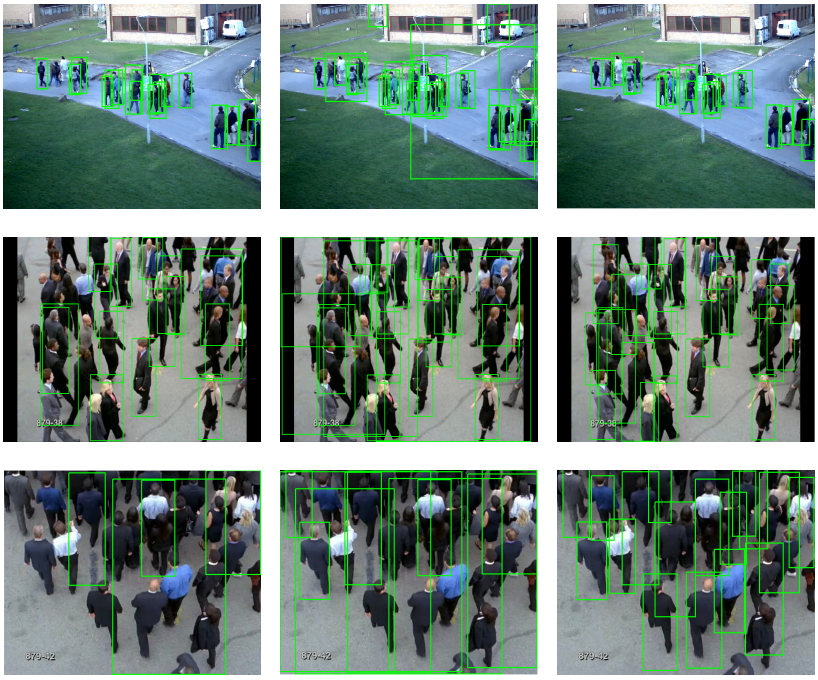
For quantitative evaluations of detection results, we use the CLEAR metrics [39]: the Multi-Object Detection Accuracy (MODA) and the Multi-Object Detection Precision (MODP). To demonstrate the effectiveness of the proposed detection algorithm, we compare the baseline method [28] using two detection thresholds (τ_{min} and τ_{max}) to the proposed method using a dynamically chosen threshold $\tau_{dyn} \in \{\tau_{min} ... \tau_{max}\}$ according to the crowd density. Additional tests are conducted to assess the impact of the correction filters.

sequence name	τ_{min}	τ_{max}	τ_{dyn}	Filtering	τ_{dyn}+ Filtering
S1.L1.13-57	0.48 / 0.65	0.36 / 0.57	0.59 / 0.59	0.48 / 0.66	0.63 / 0.63
S1.L1.13-59	0.56 / 0.68	0.25 / 0.61	0.60 / 0.67	0.56 / 0.69	0.60 / 0.68
S1.L2.14-31	0.33 / 0.63	0.09 / 0.57	0.40 / 0.59	0.32 / 0.65	0.47 / 0.63
S2.L3.14-41	0.29 / 0.54	0.04 / 0.56	0.34 / 0.56	0.29 / 0.54	0.35 / 0.57
UCF-879	0.44 / 0.58	0.34 / 0.54	0.41 / 0.55	0.41 / 0.62	0.59 / 0.58
INRIA879-42	0.27 / 0.54	0.06 / 0.55	0.35 / 0.55	0.20 / 0.42	0.42 / 0.47

Table 2: MODA / MODP results for FAST features used in the crowd density estimation and for different test videos.

As it is shown in Table 2, we set τ_{min} to (-0.5) and τ_{max} to (-1.2), these values have been found empirically suitable for lowly-resp. highly crowded scenes. The second column of this table shows that using τ_{min} as detection threshold does not provide satisfactory results, also by decreasing the threshold to τ_{max} in the third column, the results are globally worse. However, as shown in the fourth column, the automatic choice of the detection threshold gives better results than both configurations of the baseline method. Regarding the final results (in the last column), the proposed method using a dynamically chosen detection threshold together with filtering gives the best results for all test videos. These results demonstrate that using both steps (filtering and dynamic threshold) performs favorably better than implementing them separately which justifies that filtering has to be performed first to suppress false detections and to emphasize correct ones.

Figure 5 shows exemplary visual results which also indicate that the performance increases by the proposed method. Although the PETS sequences pose different problems to the detector, in all cases the proposed method improves the detection results compared to the baseline method. The UCF-879 sequence is even more challenging, however, the proposed method still enhances the detection considerably compared to the baseline method. For INRIA 879-38 sequence, people are walking very near to the camera which significantly changes their aspect ratio for different positions. Additionally, for this specific perspective, many detection candidates comprising the head of one person and the body of another are generated. As the correction filter does not apply any prior-knowledge about the shape of a person but is only estimated from



(a) (b) (c)

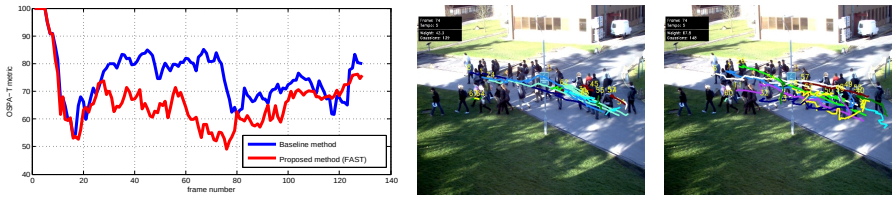
Figure 5: Exemplary visual results comparing the performance of crowd-sensitive threshold to the baseline method: (a) baseline algorithm at τ_{min}, (b) baseline algorithm at τ_{max}, (c) proposed method using dynamically chosen τ_{dyn} and correction filter according to aspect ratio and perceived height. From Top to bottom: Frames from PETS, UCF 879, and INRIA 879-38.

previous detections, it is misled in this situation. Accordingly, in this special case, its contribution is smaller.

To demonstrate the impact of improving detection results on tracking, we use Probability Hypothesis Density (PHD) filter [40] in a tracking-by-detection framework. The results in terms of OSPA-T distance [41] that are generated using the same tracker configuration for all videos are shown in Table 3. In all cases, our results using a dynamical detection threshold and correction filtering are better compared to the baseline method. These results are consistent with our expectations as the tracker relies on improved detections and lower clutter. As the tracker can deal with clutter and also with missed detections to some extent, detection improvements enhance the tracking performance

sequence name	original ($\tau = 0.5$)	proposed method
S1.L1.13-57	65.26	63.64
S1.L1.13-59	64.81	62.36
S1.L2.14-31	75.27	66.39
S2.L3.14-41	88.19	87.65
UCF-879	89.92	86.89
INRIA-879-42	81.15	73.22

Table 3: Averaged OSPA-T values for test sequences. We use a cut-off parameter $c = 100$, $\alpha = 30$ and a distance order of $d = 2$.



(a) (b) (c)

Figure 6: (a) OSPA-T distance over full sequence PETS S1.L2.14-31. (b)-(c) Exemplary visual tracking results for this scene. (b) baseline method, (c) proposed method using FAST features

but not with the same effect.

In Figure 6 (a), the OSPA-T metric over a complete scene (PETS S1.L2.14-31) is shown. For this scene with hard lighting conditions and medium crowd density, the detection performance is considerably increased by the proposed method. The diagram shows that the tracking performance of our method is mostly better than using the baseline algorithm. Visual examples are given in Fig. 6 (b)-(c) where it can be seen that our method using FAST is visibly able to track objects for a longer time and also maintains more tracks than the baseline method.

4.3. Results of crowd behavior analysis

For crowd change detection, we test our proposed approach on the publicly available UMN dataset [42], which has been widely used to distinguish between normal and abnormal crowd activities. The dataset comprises 11 videos from three indoor and outdoor scenes. Each of these videos can be divided into normal and abnormal parts.

More precisely, they illustrate different scenarios of escape events such as running in one direction, or people dispersing from a central point.

For the ground truth, as noticed in [13, 15], the labels of abnormal events shown in the videos are not accurate; there are some lags in the ground truth labels. To overcome this conflict, we use the labels of change detection of some videos from [13], and [15], for the other videos we follow the same annotation strategy; we manually label the frame in which people start running.

For quantitative evaluation, we employ the relative mean frame error [17]. As

seq. UMN	nb. frames	ground truth	our det. changes	e_F
Video1	625	484	493	0.0144
Video2	828	665	669	0.0048
Video3	549	303	319	0.0291
Video4	685	563	582	0.0277
Video5	769	492	512	0.0260
Video6	579	450	466	0.0276
Video7	895	734	754	0.0223
Video8	667	454	471	0.0255
Video9	658	551	551	0
Video10	677	570	577	0.0103
Video11	807	717	722	0.0062

Table 4: Comparison of our detection results to the ground truth labels using error frame metric

shown in Table 4, the comparison of our detection results to the ground truth labels demonstrates accurate detections in most videos. The delay in the detection of some frames after the event occurs is because of our strategy of detection, in which an abnormal event is detected only if the temporal stability is below the dynamic threshold. This requires some times to be detected, which justifies the delay. At the same time, this strategy is suitable to avoid false alarms.

Moreover, we compare our results to other methods, namely, the Social Force Model (SFM) [14], the adjacency-matrix based clustering (AMC) [15], and the similarity metric based on 2D-histograms decoupling speed and orientation in [13], see Figures 7, and 8. In these figures, the green bar indicates normal events, and the red color denotes abnormal events. These results show that our method gives better results

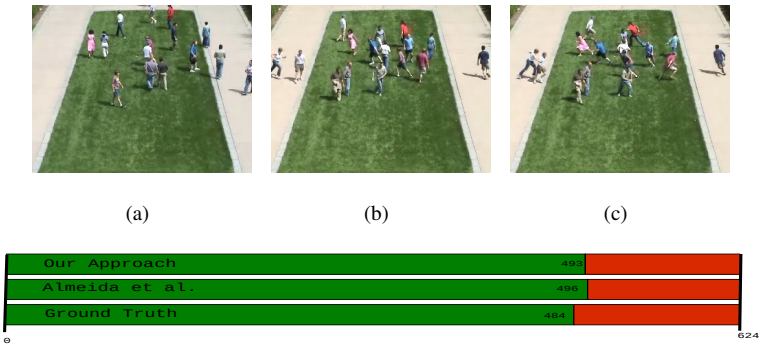
than SFM and comparable results regarding the two other methods. It is important to note that UMN dataset does not include events such as crowd formation/splitting, that could justifies the satisfactory results achieved by methods based only on motion information.

For evaluating crowd event recognition, we test our method on PETS. S3 dataset [31], which is used to assess flow analysis and event recognition algorithms. For event recognition, this dataset depicts 6 classes of crowd events: walking, running, formation (merging), splitting, evacuation, and dispersion. We randomly split this dataset into (75%) for training and (25%) for testing. Following one-vs-one strategy, we obtain (99.54%) as classification accuracy. In addition, we report the classification accuracy on the test set for each class separately, following one-vs-rest strategy, see Table 5. As

Events	Walking	Running	Splitting	Dispersion	Evacuation	Formation
accuracy	99.41	99.21	100.00	99.87	99.80	99.54

Table 5: Classification accuracy of our proposed crowd event recognition method on test set from PETS. S3 dataset following one-vs-rest strategy

it is shown in this table, we obtain good results for all crowd events including crowd formation/splitting, which justifies the relevance of our proposed crowd attributes.



(a) (b) (c)

(d)

Figure 7: Results on Video1 of UMN [42] dataset (a) The first frame of the video sequence (b) The frame in which the crowd change occurs (c) The frame in which our method detects the crowd change (d) Comparisons of our result to [13] and to the ground truth

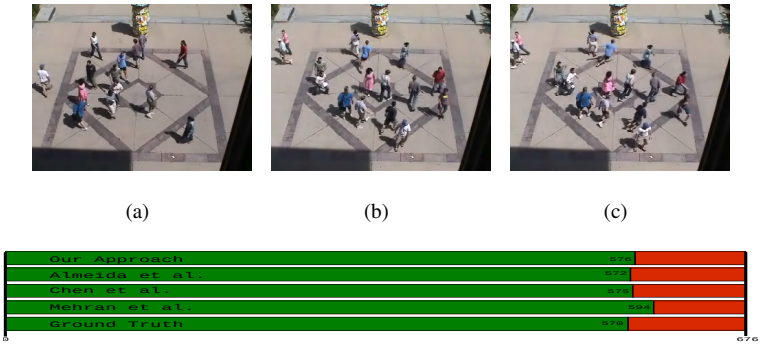
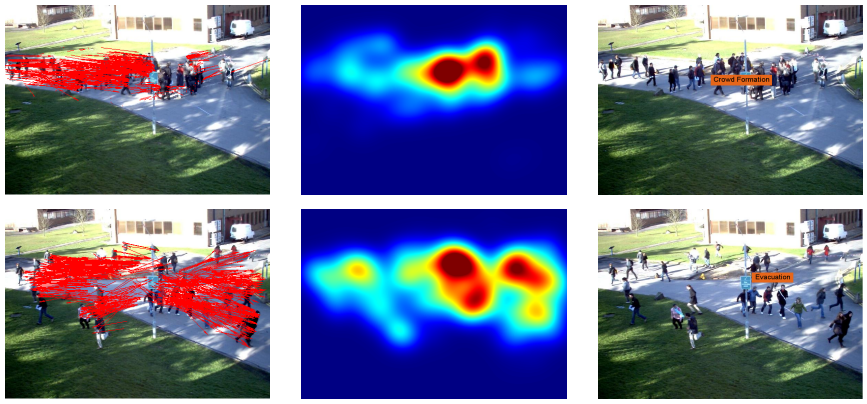


Figure 8: Results on Video10 of UMN [42] dataset (a) The first frame of the video sequence (b) The frame in which the crowd change occurs (c) The frame in which our method detects the crowd change (d) Comparisons of our result to [13, 15, 14] and to the ground truth

For evaluating our proposed method for crowd event characterization, we use PETS. S3 dataset. By following up some measures extracted from the crowd attributes (unsupervised method), we are able to monitor what is happening in the scene, to localize the event, and to have clear idea about the density of people participating to each event. Figure 9 illustrates some examples of event characterization on PETS. In the first row of this figure, a sample frame of crowd formation is shown. This event is characterized by people coming from different directions and they are moving towards the same location (as it is depicted in the first column, showing the direction of motion vectors). Also, this event is characterized by a decrease of motion area ratio in time (equal to 40.72% at this frame). In the second column, we show the estimated density map, which localizes where the crowd is formed. The area of dense regions is increasing in time, it reaches 6.10% at this frame. Given all the characteristics, crowd formation event can be recognized and localized as it is shown in the third column.

In the second row, an example of evacuation is shown. This event is characterized by the divergence of motion vectors as it is shown in the first column, because people are moving away from each others in different directions. In addition, this event is characterized by a sudden increase in the speed; the average of magnitude of all motion vectors at this frame is equal to 12.48 pixels. Evacuation event is also characterized by



a) Motion vectors (b) Density map (c) Recognized event

Figure 9: Results of event characterization from PETS dataset: examples of crowd formation and evacuation events.

in an increase in the motion area ratio (53.79%) and a decrease in time of dense areas (as it is shown in the second column).

4.4. Results of contextualized protection filters

The proposed context-dependent privacy protection filers are tested with challenging crowd scenes from PETS [31], UCF [32] and Data Driven Crowd Analysis [33] datasets. For evaluation, we adopt an objective evaluation framework, by studying the variation in performances of the commonly used algorithms in video surveillance analytics before and after applying the filters. We recall that one of the major challenges in defining privacy protection policies lies in identifying the appropriate balance between the two axis of intelligibility and privacy of the surveillance data.

On one side, we model the impact of privacy filters on intelligibility by evaluating the performances of a people counting-by-detection algorithm before and after applying the filters. We motivate our choice by observing that privacy protected video surveillance footage must at least retain those visual features necessary to perform very basic monitoring tasks such as people detection and counting. On the other side, we model privacy as the inverse score of a person matching algorithm based on local features. Such algorithm tries to identify an individual among a set of other subjects by

extracting and matching local features. This algorithm represents a common step for higher level tasks such as person re-identification, recognition or tracking, which could potentially reveal information on the identity of a subject. In our implementation, we use Hessian-Laplace interest point detector together with the SIFT descriptor and nearest neighbor matching. Based on such premise, an appropriate privacy filter should prevent the person matching algorithm to correctly detect and match local features.

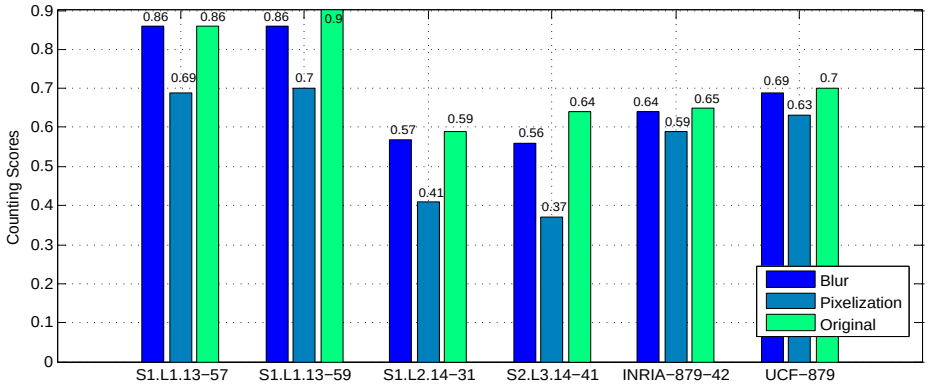


Figure 10: Counting scores on sequences protected by blur and pixelization, compared to original results.

Figure 10 reports the people counting results for blurring and pixelization protection techniques, compared to the counting scores when no protection filter is applied. The evaluation score is chosen as the percentage $p \in [0, 1]$ of correctly detected individuals with respect to the annotated ones in the ground truth. We can observe that the counting results do not decrease significantly after applying the protection filters. The score drop is 0.09, with the minimum loss observed for the blur filter. Consequently, we are still able to correctly perform people counting within a 9% error margin. We also notice that counting results are better using blurring filter compared to pixelization.

Matching results are displayed in Figure 11, following the same strategy as for counting. We can clearly observe a dramatic drop in performances of the person matching algorithm. On average, the drop in matching score is 0.39, with the maximum observed loss is for the blur filter. These results confirm that our approach for privacy protection behaves in accordance to the requirements, in terms of preservation of intelligibility and privacy of the original source. Our privacy protection filters generate a

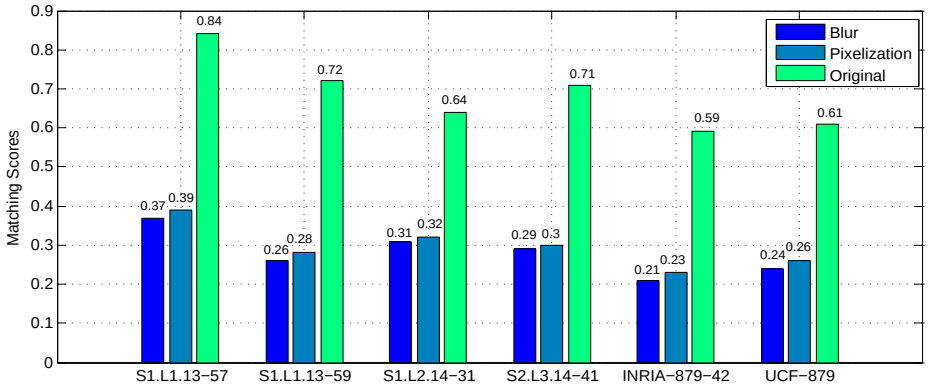


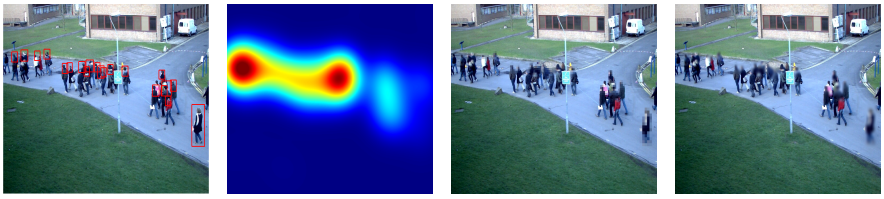
Figure 11: Matching scores on sequences protected by blur and pixelization, compared to original results.

relatively small loss in people counting score, and therefore in intelligibility, compared to the drop in performances of the matching step, and thus the gain in privacy protection level. We notice as well that in both counting and matching experiments, blurring filters provide better intelligibility and privacy levels compared to pixelization.

In Figure 12, we show the results on one frame from PETS. It is visible that the block size in the pixelization filter and the bandwidth of the Gaussian blurring are changed by our system according to the crowd density value and perceived size of the person. Comparing e.g. the woman in the lower right corner, to the persons walking in the crowd, it is well perceivable that the protection level is reduced within the crowd by a smaller block size or a smaller bandwidth respectively. At the same time, it can be seen that this woman does not have such a high density measure compared to the group of people walking in the crowd, consequently, the application of privacy protection filters is extended to the whole body.

5. Conclusion

Our contribution in this paper fits the context of crowd density estimation and its application to other video surveillance tasks. The crowd density information was represented as a new statistical model of spatio-temporal local features that varies temporally over the video and spatially across the frame. Our proposed approach was tested on videos from different datasets and the results highlighted the relevance of the feature



(a) RoIs detections (b) Crowd density map (c) Pixelized image (d) Blurred image

Figure 12: Results of adaptive protection filters using one frame from PETS: (a) RoIs detection, (b) estimated crowd density map, (c) application of pixelization filter, and (d) application of blurring filter

tracking process compared to the foreground segmentation. Furthermore, we included a comparative study between different local features in order to investigate their discriminative power to the crowd.

In addition, we approached some problems related to the crowd analysis field from a new perspective. Given the difficulties encountered by video analytic components in crowded scenes, we employed the proposed local space-time model of crowd density to complement the following applications: First, the crowd density was used to enhance human detection and tracking in crowded scenes by applying a scene-adaptive dynamic parameterization. Second, it was used with motion information for studying crowd behaviors by analyzing long-term trajectories. Finally, it was applied in privacy context to boost the compliance between privacy and surveillance concerns. The experimental results demonstrated the usefulness of the crowd density to improve detection and tracking results compared to the baseline methods. Also, its application to crowd behavior analysis showed good performance for early detection of crowd change, and accurate event recognition. Finally, the effectiveness of the proposed crowd density-dependent privacy preservation filters has been demonstrated by an objective evaluation assessing privacy and intelligibility trade-off.

There are several possible extensions of this work: First, for human detection and tracking more contextual information to improve the results in crowded scenes might be investigated. Also, since the incorporation of the crowd density model into the tracking is performed by providing improved detection results, a more elegant approach could

formulate both detection and tracking as a joint framework and crowd density information could be integrated in both steps to enforce scene constraints. For crowd behavior analysis, our proposed method succeeds to achieve accurate results for early detection and recognition once the change or the event occurs, however, it is important to investigate event prediction (before it happens). Finally, for privacy preservation, since we only used objective evaluation to assess our proposed contextualized protection filers; it could be advantageous to perform subjective evaluation of them as well.

References

[1] A. B. Chan, Z. S. J. Liang, N. Vasconcelos, Privacy preserving crowd monitoring: Counting people without people models or tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–7.

[2] D. Conte, P. Foggia, G. Percannella, F. Tufano, M. Vento, A method for counting people in crowded scenes, in: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010.

[3] H. Fradi, J. L. Dugelay, Low level crowd analysis using frame-wise normalized feature for people counting, in: IEEE International Workshop on Information Forensics and Security, 2012.

[4] A. Polus, J. L. Schofer, A. Ushpiz, Pedestrian flow and level of service, Journal of Transportation. Engineering 109 (1983) 46–56.

[5] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987.

[6] W. Ma, L. Huang, C. Liu, Advanced local binary pattern descriptors for crowd estimation, Computational Intelligence and Industrial Application 2 (2008) 958–962.

[7] Z. Wang, H. Liu, Y. Qian, T. Xu, Crowd density estimation based on local binary pattern co-occurrence matrix, IEEE International Conference on Multimedia and Expo Workshops.

[8] H. Yang, H. Su, S. Zheng, S. Wei, Y. Fan, The large-scale crowd density estimation based on sparse spatiotemporal local binary pattern, IEEE International Conference on Multimedia and Expo (2011) 1–6.

[9] H. Fradi, X. Zhao, J. L. Dugelay, Crowd density analysis using subspace learning on local binary pattern, in: ICME 2013, IEEE International Workshop on Advances in Automated Multimedia Surveillance for Public Safety, 2013.

[10] M. Rodriguez, J. Sivic, I. Laptev, Analysis of crowded scenes in video, in: J. Y. Doufour (Ed.), Intelligent Video Surveillance Systems, Wiley, 2012, pp. 251–272.
URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1848214332.html

[11] S. Ali, M. Shah, Floor fields for tracking in high density crowd scenes, in: ECCV (2), 2008, pp. 1–14.

[12] M. Rodriguez, S. Ali, T. Kanade, Tracking in unstructured crowded scenes, in: ICCV, 2009, pp. 1389–1396.

[13] I. R. de Almeida, C. R. Jung, Change detection in human crowds, in: Conference on Graphics, Patterns and Images, no. 26, 2013.

[14] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: CVPR, 2009, pp. 935–942.

[15] D. Y. Chen, P. Huang, Motion-based unusual event detection in human crowds, J. Visual Communication and Image Representation 22 (2) (2011) 178–186.

[16] C. Garate, P. Bilinski, F. Bremond, Crowd Event Recognition using HOG Tracker, in: Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter), IEEE, 2009, pp. 1–6. doi:10.1109/PETS-WINTER.2009.5399727.
URL http://hal.inria.fr/inria-00515197

[17] V. Kaltsa, A. Briassouli, I. Kompatsiaris, M. G. Strintzis, Timely, robust crowd event characterization, in: ICIP, 2012, pp. 2697–2700.

[18] R. Emonet, J. Varadarajan, J.-M. Odobez, Temporal analysis of motif mixtures using dirichlet processes, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

[19] V. Norris, M. McCahill, D. Wood, Editorial: The growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space, Surveillance and Society 2(2/3) (2004) 110–135.

[20] A. Senior, S. Pankanti, A. Hampapur, L. Brown, Y.-L. Tian, A. Ekin, J. Connell, C.-F. Shu, M. Lu, Enabling video privacy through computer vision, Security Privacy, IEEE 3 (3) (2005) 50–57. doi:10.1109/MSP.2005.65.

[21] S. Moncrieff, S. Venkatesh, G. West, Context aware privacy in visual surveillance, in: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008, pp. 1–4. doi:10.1109/ICPR.2008.4761616.

[22] H. Fradi, J.-L. Dugelay, Crowd density map estimation based on feature tracks, in: MMSP 2013, 15th International Workshop on Multimedia Signal Processing, September 30-October 2, 2013, 2013.

[23] T. Senst, V. Eiselein, T. Sikora, Robust local optical flow for feature tracking, Transactions on Circuits and Systems for Video Technology 09 (99).

[24] E. Rosten, R. Porter, T. Drummond, Faster and better: A machine learning approach to corner detection, IEEE Trans. Pattern Analysis and Machine Intelligence 32 (2010) 105–119.

[25] M. Butenuth, F. Burkert, F. Schmidt, S. Hinz, D. Hartmann, A. Kneidl, A. Borrmann, B. Sirmacek, Integrating pedestrian simulation, tracking and event detection for crowd analysis, in: ICCV Workshops, 2011, pp. 150–157.

[26] T. Senst, V. Eiselein, R. H. Evangelio, T. Sikora, Robust modified l2 local optical flow estimation and feature tracking, in: IEEE Workshop on Motion and Video Computing (WMVC), 2011, pp. 685–690.

[27] C. Tomasi, T. Kanade, Detection and tracking of point features, Technical report CMU-CS-91-132, CMU (1991).

[28] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1627–1645.

[29] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, Vol. 2, 2005, pp. 886–893.

[30] D. Hoiem, A. A. Efros, M. Hebert, Putting objects in perspective, International Journal of Computer Vision 80 (1) (2008) 3–15.

[31] J. Ferryman, A. Shahrokni, Pets2009: Dataset and challenge, in: PETS, 2009, pp. 1–6. doi:10.1109/PETS-WINTER.2009.5399556.

[32] S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: CVPR 07, 2007, pp. 1–6.

[33] M. Rodriguez, J. Sivic, I. Laptev, J.-Y. Audibert, Data-driven crowd analysis in videos, in: ICCV, 2011.

[34] V. Lempitsky, A. Zisserman, Learning to count objects in images, in: J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (Eds.), Advances in Neural Information Processing Systems 23, 2010, pp. 1324–1332.

[35] D. G. Lowe, Distinctive image features from scale-invariant keypoints, in: Int. J. Comput. Vision, 2004, pp. 91–110.

[36] J. Shi, C. Tomasi, Good features to track, in: CVPR, 1994, pp. 593 –600. doi:10.1109/CVPR.1994.323794.

[37] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: Proc. BMVC, 2002, pp. 36.1–36.10, doi:10.5244/C.16.36.

[38] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Underst. 110 (3) (2008) 346–359. doi:10.1016/j.cviu.2007.09.014.
URL http://dx.doi.org/10.1016/j.cviu.2007.09.014

[39] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, P. Soundararajan, The clear 2006 evaluation, in: Multimodal Technologies for Perception of Humans, Vol. 4122, 2007, pp. 1–44. doi:10.1007/978-3-540-69568-4_1.

[40] B.-N. Vo, W.-K. Ma, The gaussian mixture probability hypothesis density filter, Signal Processing, IEEE Transactions on 54 (11) (2006) 4091 –4104. doi:10.1109/TSP.2006.881190.

[41] B. Ristic, B.-N. Vo, D. Clark, B.-T. Vo, A metric for performance evaluation of multi-target tracking algorithms., IEEE Transactions on Signal Processing 59 (7) (2011) 3452–3457.

[42] U. of minnesota crowd activity dataset, http://www.mha.cs.umn.edu/Movies/Crowd-Activity-All.avi.