Spatio-Temporal Crowd Density Model in a Human Detection and Tracking Framework

Hajer Fradi[a], Volker Eiselein[b], Jean-Luc Dugelay[a], Ivo Keller[b], Thomas Sikora[b]

[a]*Multimedia Department, EURECOM, Sophia Antipolis, France*
[b]*Communication Systems Group, Technische Universität Berlin, Germany*

Abstract

Recently significant progress has been made in the field of person detection and tracking. However, crowded scenes remain particularly challenging and can deeply affect the results due to overlapping detections and dynamic occlusions. In this paper, we present a method to enhance human detection and tracking in crowded scenes. It is based on introducing additional information about crowds and integrating it into the state-of-the-art detector. This additional information cue consists of modeling time-varying dynamics of the crowd density using local features as an observation of a probabilistic function. It also involves a feature tracking step which allows excluding feature points attached to the background. This process is favourable for the later density estimation since the influence of features irrelevant to the underlying crowd density is removed. Our proposed approach applies a scene-adaptive dynamic parametrization using this crowd density measure. It also includes a self-adaptive learning of the human aspect ratio and perceived height in order to reduce false positive detections. The resulting improved detections are subsequently used to boost the efficiency of the tracking in a tracking-by-detection framework. Our proposed approach for person detection is evaluated on videos from different datasets, and the results demonstrate the advantages of incorporating crowd density and geometrical constraints into the detection process. Also, its impact on tracking results have

*Corresponding author
Email address: fradi@eurecom.fr (Hajer Fradi)

been experimentally validated showing good results.

Keywords: Crowd density, local features, human detection, tracking, crowded scenes

1. Introduction

Automatic detection and tracking of people in video data is a common task in the research area of video analysis and its results lay the foundations of a wide range of applications such as video surveillance, behavior modeling, security applications, and traffic control. Many tracking algorithms use the "Tracking-by-detection" paradigm which estimates the tracks of individual objects based on a previously computed set of object detections. Tracking methods based on these techniques are manifold [1, 2, 3, 4], but all of them rely on efficient detectors which have to identify the position of persons in the scene while minimizing false detections (clutter) in areas without people. Techniques based on background subtraction such as [5] are widely applied thanks to their simplicity and effectiveness but are limited to scenes with few and easily perceptible components. Therefore, the application of these methods on videos containing dense crowds is more challenging.

Crowded scenes exhibit some particular characteristics rendering the problem of multi-target tracking more difficult than in scenes with few people: Firstly, due to the large number of pedestrians within extremely crowded scenes, the size of a target is usually small in crowds. Secondly, the number of pixels of an object decreases with a higher density due to the occlusions caused by inter-object interactions. Thirdly, constant interactions among individuals in the crowd make it hard to discern them from each other. Finally and as the most difficult problem, full target occlusions that may occur (often for a long time) by other objects in the scene or by other targets. All the aforementioned factors contribute to the loss of observation of the target objects in crowded videos. These challenges are added to the classical difficulties hampering any tracking algorithm such as: changes in the appearance of targets related to

the camera view field, the discontinuity of trajectories when the target leaves the field of view and re-appears later again, cluttered background, and similar appearance of some objects in the scene.

Because of all these issues, conventional human detection or multi-target tracking paradigms are not scalable to crowds. That is why, some current solutions in crowd analysis field bypass the detection and the tracking of individuals in the scene. Instead, they focus on detecting and tracking local features [6, 7], or particles [8, 9]. The extracted local features are employed to represent the individuals present in the scene. By this way, tracking of individuals in crowds which is a daunting task is avoided. Likewise, alternative solutions that operate on particles tracking, observe that when persons are densely crowded, individual movement is restricted, thus, they consider members of the crowd as granular particles. For instance, in [6], Ihaddadene *et al.* propose to detect sudden change and abnormal motion variations using motion heat maps and optical flow. The proposed approach is based on tracking points of interest in the regions of interest (masks that correspond to areas of the built motion heat map). Then, the variations of motion are used to detect abnormal events. For this purpose, an entropy measure that characterizes how much the optical flow vectors are organized, or cluttered in the frame is defined in terms of a set of statistical measure. Another study that addressed the problem of abnormal crowd event detection is the social force model proposed by Mehran *et al.* in [8]. This method is based on putting a grid of particles over the image frame and moving them with flow field computed from the optical flow. Then, the interactions forces are computed on moving particles to model the ongoing crowd behaviors. In the same context of crowd behavior analysis, other methods [10, 11] studied the dynamic evolution of the crowd using biological models.

Most of the proposed works to takle multi-target tracking in crowded scenes use motion pattern information as priors to tracking. Some of these methods are applied in unstructured crowd scenes [12], while most of them focus on structured scenes [13, 14, 15], where objects do not move randomly, and exhibit clear motion patterns. In [12], a tracking approach in unstructured environments,

where the crowd motion appears to be random in different directions over time is presented. Each location in the scene can represent motion in different directions using a topical model. In [13], a Motion Structure Tracker is proposed to solve the problem of tracking in very crowded scenes. In particular, tracking and detection are performed jointly and motion pattern information is integrated in both steps to enforce scene structure constraints. In [14], a probabilistic method exploiting the inherent spatially and temporally varying structured pattern of crowd motion is employed to track individuals in extremely crowded scenes. The spatial and temporal variations of the crowd motion are captured by training a collection of Hidden Markov Models on the motion patterns within the scene. Using these models, pedestrian movement at each space-time location in a video can be predicted. Also motion patterns are studied in [15], where floor fields are proposed to determine the probability of moving from one location to another. The idea is to learn global motion patterns and participants of the crowd are then assumed to move in a similar pattern. Finally, in [16] a spatiotemporal viscous fluid field is proposed to recognize large-scale crowd event. In particular, a spatiotemporal variation matrix is proposed to exploit motion property of a crowd. Also, a spatiotemporal force field is employed to exploit the interaction force between the pedestrians. Then, the spatiotemporal viscous fluid field is modeled by latent Dirichlet allocation to recognize crowd behavior.

Although these solutions have shown promising results, they impose constraints on the crowd motion. In particular, targets are often assumed to behave in a similar manner, in such a way that all of them follow a same motion pattern, consequently, trajectories not following common patterns are penalized. Certainly, this constraint works well in extremely crowded scenes, such as in some religious events or demonstrations, where the movement of individuals within the crowd is restricted by others and by the scene structure as well. Thus, a single object can be tracked by the crowd motion because it is difficult, if not impossible, to move against the main trend. However, the aforementioned methods are not applicable in cases where individuals can move in different directions. Furthermore, some of these methods include other additional constraints. For

example, in [12], Rodriguez *et al.* use a limited descriptive representation of target motion by quantizing the optical flow vectors into 10 possible directions. Such a coarse quantization limits tracking to only few directions. Also, the *floor fields* [15] used by Ali *et al.* impose how a pedestrian should move based on scene-wide constraints, which results in only one single direction at each spatial position in the video.

In addition to these solutions based on exploiting global level information about motion patterns to impose constraints on tracking algorithms, similar ideas have been proposed using crowd density measures. In [17], Hou *et al.* use the estimated number of persons in the detection step, which is formulated as a clustering problem with prior knowledge about the number of clusters. This attempt to improve person detection in crowded scenes includes some weaknesses. At least two problems might incur: Firstly, the idea of detection by clustering features can only be effective in low crowded scenes. It is not applicable in very crowded cases because of the spatial overlaps that make delineating individuals a difficult task. Secondly, using the number of people as a crowd measure has the limitation of giving only global information about the entire image and discarding local information about the crowd.

We therefore resort to another crowd density measure, in which local information at pixel level substitutes a global number of people per frame. This alternative solution based on computing crowd density maps is indeed more appropriate as it enables both the detection and the location of potentially crowded areas. To the best of our knowledge, only one work [18] has investigated this idea. In the referred work, a system which introduces crowd density information into the detection process is proposed. Using an energy formulation, Rodriguez *et al.* [18] show how it is possible to obtain better results than the baseline method [19]. Although it is a significant improvement of multi-target tracking in crowded scenes, the referred work employs confidence scores from person detection as input to the density estimation. It means that the detection scores are used twice, to detect persons and to estimate crowd density maps which does not introduce any complimentary information in the process. In addition,

the proposed crowd density map in [18] involves a training step with large data. Thus, human-annotated ground truth detections are required, and the system is not fully automatic.

In contrast to the previous work, we intend to demonstrate in this paper, how it is possible to enhance detection and tracking results using fully automatic crowd density maps that characterize the spatial and temporal variations of the crowd. The proposed crowd density map is typically based on using local features as an observation of a probabilistic density function. A feature tracking step is also involved in the process to alleviate the effects of feature components irrelevant to the underlying crowd density. Compared to the 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 implies a large number of likely movements in each space-time location of the video. This additional information is incorporated in a detection and tracking framework: First, the proposed space-time model of crowd density is used as a set of priors for detecting persons in crowded scenes, where we apply the deformable part-based models [19]. A filtering step based on the aspect ratio and the perceived height of a person precedes the fusion of the crowd density and the detection filter in order to deal with false positive detections of inappropriate size. Second, we extend our approach to tracking using Probability Hypothesis Density (PHD) filter based on the improved detections.

As in many video surveillance setups, we consider the camera to be static. This assumption may appear strict but in fact reflects a high number of real surveillance setups and applications (e.g. numerous applications using background subtraction). However, there exist approaches for feature tracking on moving/PTZ cameras (e.g. using global motion estimation / compensation) and future work could use them to extend the system to non-static camera views.

The remainder of the paper is organized as follows: In the next Section, we introduce the human detector we use. Details about our proposed crowd density map are given in Section 3. In Section 4, we explain how to use this crowd density information together with a correction filter in order to improve

the detection results. In Section 5, an extension of the improved detection results to tracking using PHD methodology for data-association is presented. A detailed evaluation of our work follows in Section 6. Finally, we briefly conclude and give an outlook of possible future works.

2. Human detection using Deformable Part-Based Models

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 studied intensively and different approaches have been proposed (e.g. [20], and [19]) which are often gradient-based. In most of the proposed methods, the problem is formulated via binary sliding window classification, where an image pyramid is built and a fixed window size is scanned at all locations and scales to localize individuals. In this context, the deformable part-based models [19] has recently shown excellent performance. It is an enriched version of Histograms of Oriented Gradients (HoG) [20], that achieves much more accurate results and represents the current state-of-the-art. The detector uses a feature vector over multiple scales and a number of smaller parts within a Region of Interest (RoI) to get additional cues about an object (see Figure 1).

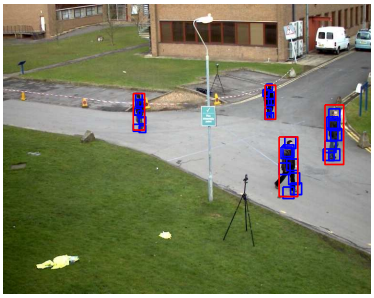


Figure 1: Exemplary human detections using the part-based models [19]: Blue boxes describe object parts which also contribute to the overall detection (red).

In this framework, an object hypothesis specifies the location of each filter in a feature pyramid $z = (p_0, ..., p_n)$ with $p_i = (x_i, y_i, l_i)$ as the position and level

of the i-th filter. The detection score is given as the score of all filters minus a deformation cost plus a bias b:

$$score(p_0, ..., p_n) = \sum_{i=0}^{n} F_i' \cdot \Psi(H, p_i) \quad - \sum_{i=1}^{n} d_i \cdot \Psi_d(dx_i, dy_i) + b \tag{1}$$

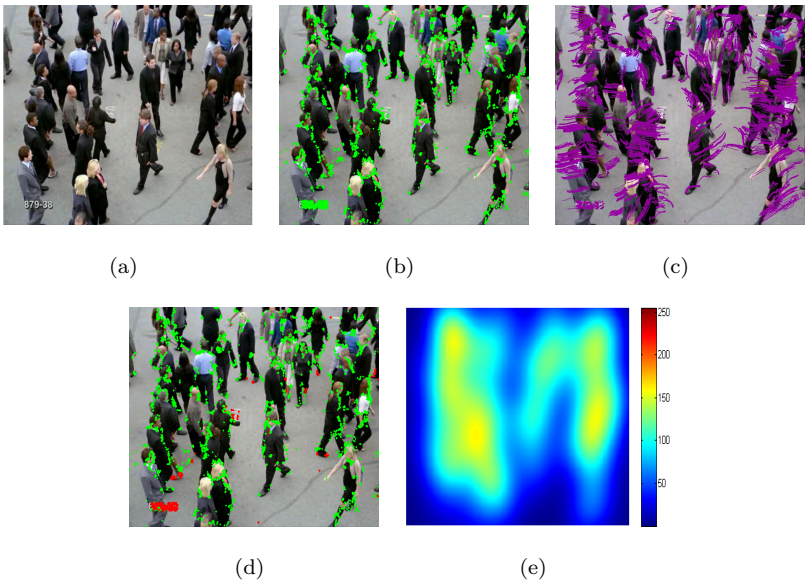
with (dx_i, dy_i) as the displacement of the i-th part relative to its anchor position and $\Psi_d(dx, dy)$ as deformation features weighted by the vector d_i.

In this work we use the implementation from [21] which is trained on samples of the INRIA and PASCAL person datasets. The output of the detector is a set of RoIs for a given detection threshold. These must then be processed by an additional non-maximum suppression (NMS) step which is essentially based on maintaining regions with high detection scores while removing detections overlapping with these more than a given threshold.

Although human detection using the deformable part-based models has become a quite popular technique, its extension to crowded scenes has a 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's body are visible. Therefore, accurate human detection in such scenarios with dynamic occlusions and high interactions among the targets remains a challenge.

To improve the detection performance in crowded scenes, some methods (e.g. [18], and [22]) rely only on head detections and discard the rest of the body. This is less error-prone but also focuses on a smaller amount of information characterizing a human. Although improved accuracy can be obtained using these solutions, the large amount of partial occlusions in high dense crowds still present big challenges to such detection methods. In order to adapt the detector to these situations, it is important to include additional information about crowds in the scene. In the following, we present details on our proposed local crowd density measure which conveys rich information about the spatial distributions of persons in order to enhance the detection process.

3. Crowd Density Estimation



(a) (b) (c)

(d) (e)

Figure 2: 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.

Crowd density analysis has been studied as a major component for crowd monitoring and management in visual surveillance systems. In this paper, we explore a new promising research direction which consists of using crowd density measures to complement person detection. For this, generating locally accurate crowd density maps is more helpful than computing only an overall density [23] or a number of people [24] in a whole frame. In the following, our proposed approach for crowd density estimation [25] is presented. 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 [26] and a point rejection step using forward-backward projection. To accurately represent the motion within the video, the estimation of optical flow between consecutive frames is extended to long-term 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 2. The remainder of this section describes each of these system components.

3.1. 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 a high-density crowd, we propose to use local feature 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 select Features from Accelerated Segment Test (FAST) [27], Scale-Invariant Feature Transform (SIFT) [28], and Good Features to Track (GFT) [29]. The reason behind selecting these features for crowd measurement is as follows: FAST was 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 [30] to detect dense crowds from aerial images and the derived results demonstrate a reliable detection of crowded regions. SIFT is another well-known texture descriptor, for which interest point locations are defined as maxima/minima of the difference of Gaussians in scale-space. Under this respect, SIFT is rather independent of the perceived scale of the considered object which is appropriate for crowd measurements. These two aforementioned features are compared to the classic feature detector GFT, which is based on the detection of corners containing high frequency information in two dimensions and typically persist in an image despite object variations.

3.2. 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, there is a 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 their non-zero motion vectors.

Motion estimation is performed using the Robust Local Optical Flow (RLOF) [26] [31], 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 other motion patterns or by occlusions. While RLOF handles these noise effects better than the standard Kanade-Lucas-Tomasi (KLT) feature tracker [32], it is still subject to 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 longterm trajectories by connecting motion vectors computed on consecutive frames. This results in a set of p_k trajectories in every time step k:

$$\mathcal{T}_k = \{T_1, ..., T_{p_k}|$$
$$T_i = \{X_i(k - \Delta t_i), Y_i(k - \Delta t_i), ..., X_i(k), Y_i(k)\}\} \quad (2)$$

where Δt_i denotes temporal interval between the start and the current frames of a trajectory T_i. $(X_i(k-\Delta t_i), Y_i(k-\Delta t_i))$, and $(X_i(k), Y_i(k))$ are the coordinates of the feature point in its start and current frames respectively.

The advantage of using trajectories in our system instead of computing the

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

motion vectors between two consecutive frames is that outliers are filtered out and the overall motion information is more reliable and less affected by noise.

3.3. Kernel density estimation

After generating trajectories, our following goal is to remove static features. These are identified by comparing the displacements of the generated trajectories to a small constant ζ. It proceeds by comparing the overall average motion Γ_i of a trajectory T_i to a certain threshold ζ which is set according to image resolution and camera perspective. Moving features are then identified by the relation $\Gamma_i > \zeta$ while the others are considered as part of the static background. Using long-term trajectories, the separation between foreground and background entities is improved and the number and position 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 local features. The observation can be formulated as the more local features come towards each other, the higher crowd density is perceived. 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 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:

$$C_k(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{m_k} \exp -(\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}) \tag{3}$$

where σ is the bandwidth of the 2D Gaussian kernel.

The resulting crowd density map characterizes the spatial and temporal variations of the crowd. The spatial variation arises across the frame thanks to the probability density function and temporal variation occurs over the video by the motion information included in the process. Overall, this spatio-temporal crowd information introduced by density maps conveys rich information about the distributions of pedestrians in the scene which could complement the detection process.

4. Integration of crowd density and geometrical constraints into human detector

In this Section, we present our proposed extension of human detection algorithm described in Section 2 to crowded scenes. As a major improvement, we propose a variation of the standard non-maximum suppression (NMS) by using the crowd density measure presented in Section 3 to improve human detection performance in crowds. In addition, some geometrical constraints are introduced in a first filtering step to remove false positive detections. The remainder of this section is organized as follows: First, we present our proposed density-based NMS in crowd-context constraints (Section 4.1). Then, the geometrical constraints introduced in a filtering step are defined (Section 4.2). In Section 4.3, a summary of our proposed integration algorithm is presented.

4.1. Crowd Context Constraint

The usage of detection thresholds in many human detectors is problematic in real-world applications. Beforehand it is not always clear to the user how to adapt the algorithm to a new scene and how to choose the threshold value. While lower values usually increase the number of detections and allow recognizing more persons, they also increase the number of false positives. On the other hand, higher thresholds only detect more reliable candidate regions but might cause the detector to miss some people in the scene.

This is especially difficult in heterogeneous scenes with crowded and non-crowded regions and is due to the fact that high crowd scenes present many challenges that are not present in low-crowd scenes. These include the large number of persons, small target size, occlusions because of object interactions. The impact of these difficulties on the detection results is highly dependent on the crowd size i.e. the higher the crowd density, the more difficult it is to detect people. 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 important to find a way of automatically setting the

detection threshold τ according to the probability that people are present in a certain position of the image. As discussed in Section 3, crowd density maps provide this information. Therefore, we propose to use this local information regarding the crowd density in order to adjust the detection threshold.

In the detection step of a video sequence of N frames $\{I_1, ..., I_N\}$, we obtain a set of candidate RoIs for a given threshold τ: $\mathscr{D}(\tau) = \{\mathcal{D}_1, ..., \mathcal{D}_N\}$, where $\mathcal{D}_k = \{d_1^k, ..., d_{n_k}^k\}$ is the set of detections at frame k. d_j^k denotes 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 of the RoI d_j^k and w_j^k, h_j^k are the respective width and height. Using a pre-defined range of detection thresholds given by an lower/upper boundary τ_{min}/τ_{max}, we apply the following linear density-scaled adaptive rule to automatically select acceptance threshold value of the detector:

$$\tau_{dyn} = \tau_{max} + (\tau_{min} - \tau_{max}) \cdot \hat{C}_k(d_j^k), j \in \{1...n_k\} \tag{4}$$

with

$$\hat{C}_k(d_j^k) = \frac{\sum_{p=0}^{h_j^k - 1} \sum_{q=0}^{w_j^k - 1} C_k(x_j^k + p, y_j^k + q)}{w_j^k \cdot h_j^k} \tag{5}$$

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

To obtain the dynamic threshold τ_{dyn} for every candidate d_j^k in $\mathscr{D}_{min}$, the average crowd density $\hat{C}_k(d_j^k)$ is computed as in (5) and inserted into (4) for all regions.

4.2. Geometrical Constraints

Due to the part-based nature of the used human detector, it is possible that certain human parts which actually lie on *different* persons are matched together in *one* candidate RoI which then comprises all of them (highlighted in yellow in Figure 3 (a)) or that a region is chosen even though it is much too large to surround one person (shown in red in Figure 3 (a)). If the score of such detection is higher than the scores of the individual objects' detections, the NMS step will keep it instead of the correct individual detections which might otherwise

(a) (b)

Figure 3: Exemplary effects of the proposed correction filters on a frame from PETS dataset [33]: (a) detections without filtering, (b) filtering according to aspect ratio and perceived height. While the unfiltered detections might include too large candidates (red) and also detections comprising several persons at correct height (yellow), the aspect ratio and perceived height allow removing most of them.

be recognized. Accordingly, in this case a false positive detection and a number of missed detections are generated which result in a decrease in the detection performance. We propose to overcome this problem by applying geometry-based pre-filters in order to filter out inaccurate detections of inappropriate size. The design of geometrical correction filters is based on the perceived height and the aspect ratio.

Since the perceived size of a person is affected by perspective distortions, any detected RoI for person farther away accounts for a smaller portion compared to closer persons. Based on that, we design a filter that uses the height of a candidate RoI to indicate the likelihood of human presence. Also, as some detections could comprise multiple persons at once, we propose to use the aspect ratio as a correction measure. Given the set of candidate RoIs $\mathcal{D}_k$, following [34] we assume the relationship between a person's position and his/her perceived height to be:

$$h_j^k = \alpha_{k-1} \cdot y_j^k + \beta_{k-1}, j \in \{1...n_k\} \tag{6}$$

where α_{k-1} and β_{k-1} parameters are computed using a standard regression.

Also, the aspect ration is defined as:

$$\gamma_{k-1} = median \left\{ \frac{w_j^i}{h_j^i} \right\}_{1 \leq i \leq (k-1), 1 \leq j \leq n_i} \tag{7}$$

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 updated at each frame.

These proposed correction filters use the previous detections to predict the height and the ratio of a new candidate, allowing the algorithm to operate on-line without any previous 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 an unlikely large or small region and to detect other objects in the same area which would have been suppressed otherwise. An example of these correction filters can be seen in Figure 3 (b) where false positive detections from the previous images are suppressed.

4.3. Summary of the integration algorithm

Algorithm 1 shows in pseudo-code an overview of our proposed human detection algorithm in crowds by integrating crowd density and geometrical constraints into the state-of-the-art human detector. The implementation of this algorithm can be efficiently done as follows: Firstly, a set of candidate RoIs $\mathscr{D}$ is computed for the minimal detection threshold τ_{min}. This set contains all possible detections which can be extracted for the given threshold range $[\tau_{min} ... \tau_{max}]$. To filter out inaccurate detections of inappropriate size, the two proposed geometrical filters are applied. A detection d_j^k is accepted only if it fits the predicted ratio and height with an error less than certain thresholds (Δ_γ, Δ_h) i.e. only if ($(w_j^k / h_j^k) \leq \gamma_{k-1} \pm \Delta_\gamma$) and ($h_j^k \leq \widetilde{h}_j^k \pm \Delta_h$), where $\widetilde{h}_j^k$ denotes the predicted height of the bounding box, computed from (6).

After applying these two geometrical filters, we obtain a set of new detections $\mathcal{D}'_k$ and their corresponding scores $\mathcal{S}'_k$. At this stage, we often get multiple overlapping detections, thus we use a greedy procedure for eliminating repeated

Algorithm 1 Proposed Human Detection in Crowds

Input:

- $\mathscr{I} = \{I_k\}_{1 \leq k \leq N}$, N frames of a given video sequence V and their corresponding crowd density maps $\mathscr{C} = \{C_k\}_{1 \leq k \leq N}$.
- $\mathscr{D} = \{\mathcal{D}_k\}_{1 \leq k \leq N}$: a set of preliminary candidate detections and their corresponding scores $\mathscr{S} = \{\mathcal{S}_k\}_{1 \leq k \leq N}$.

Output: Selected detections $\mathscr{D}''$

Initialize: Set (α_0, β_0, γ_0) parameters to $-\infty$

for $k = 1$ **to** N **do**

$\mathcal{D}_k = \{d_1^k, ..., d_{n_k}^k\}$, $\mathcal{S}_k^k = \{s_1^k, ..., s_{n_k}^k\}$

1. **Filtering**:
   if $(\alpha_{k-1} = -\infty)$
   $\mathcal{D}'_k \leftarrow \mathcal{D}_k$, $\mathcal{S}'_k \leftarrow \mathcal{S}_k$
   else
   $(\mathcal{D}'_k, \mathcal{S}'_k) \leftarrow$ Apply filtering $(\mathcal{D}_k, \mathcal{S}_k, \alpha_{k-1}, \beta_{k-1}, \gamma_{k-1})$
   end if
2. **nms-based-density:**
   $\mathcal{D}'{}_k = \{d_1'^k, ..., d_{m_k}'^k\}$, $\mathcal{S}'{}_k = \{s_1'^k, ..., s_{m_k}'^k\}$
   - $Index_1^k \leftarrow$ Sort confidence scores $\mathcal{S}'_k$
   - **for** each position $i \in Index_1^k$ **do**
     Compute ratio of overlap ϑ_{ij}^k between detections at
     $Index_1^k(i)$ and at $Index_1^k(j)$, $(i+1) \leq j \leq m_k$
     end for
   - $Index_2^k \leftarrow$ Remaining index after removing all overlapped detections more than a certain threshold $\Delta_o = 0.5$
   - $C_k \leftarrow$ Normalize Density Map C_k to [0...1]
   - For all pixels $x \in I_k$, compute detection thresholds using a predefined range of detection thresholds $[\tau_{min}...\tau_{max}]$ and the normalized C_k
   - $Index_F^k = \{\}$
   - **for** $c = 1$ **to** length $(Index_2^k)$ **do**
     $\tau_{dyn}(d'^k_{Index_2^k(c)}) \leftarrow$ average of detection threshold values of all pixels belonging to the RoI
     if $(s'^k_{Index_2^k(c)} \geq \tau_{dyn})$, then $Index_F^k \leftarrow \{Index_F^k, c\}$
     end for
   - $\mathcal{D}''_k \leftarrow \mathcal{D}'k\{Index_F^k\}$
3. $(\alpha_k, \beta_k, \gamma_k) \leftarrow$ Update Filtering Parameters $(\{\mathcal{D}''_l\}_{1 \leq l \leq k})$

end for

detections. It proceeds by sorting the detections $\mathcal{D}'_k$ according to their corresponding scores and greedily selecting the highest scoring ones while skipping overlapped detections that are covered by more than 50% by a bounding box of a previously selected detection. The following step consists of thresholding the remaining detections using the computed dynamic threshold according to the crowd density. Finally, the filtering parameters α_k, β_k, and γ_k are updated according to the new selected detections $\mathcal{D}''_k$. In the following, $\mathcal{D}_k$ denotes the selected detections.

5. Tracking-by-detection using Probability Hypothesis Density

To demonstrate the impact of improving detection results on tracking, we use PHD filter [35] in a tracking-by-detection framework. Other tracking methods such as Multiple Hypotheses tracking (MHT) [36] or Joint Probabilistic Data Association Filter (JPDAF) [37] could also be applied but schematically this will not result in a fundamentally different approach because all these methods rely on a previous accurate detection step before combining the given detections to tracklets. The choice of a PHD tracker instead of other methods is driven by two main reasons: a) its known sensitivity towards missed detections (in order to show improvements by the enhanced detection step) and b) its provable Bayes optimality [38] which makes it superior to MHT.

For implementation, we use a Gaussian-Mixture Probability Hypothesis Density (GM-PHD) filter [39] which assumes a linear motion model and expresses the PHD function $\Theta(\mathbf{x})$ at time step k as a mixture of Gaussians with their respective mean and covariance values $\mu_k^{(i)}, \Sigma_k^{(i)}$:

$$\Theta_k(\mathbf{x}) = \sum_{i=1}^{J_k} w_k^{(i)} N(\mathbf{x}; \mu_k^{(i)}, \Sigma_k^{(i)}) \tag{8}$$

This filter models the PHD function $\Theta(\mathbf{x})$ at time step k as a mixture of Gaussians and propagates them in an estimation step from the previous state **x'** according to the object motion model $f(\mathbf{x}|\mathbf{x'})$. A survival probability $p_S(\mathbf{x'})$

can account for exit points in a scene. Additionally, birth distributions $N_b(\mathbf{x})$ are added in the estimation step for all detections in order to account for new objects:

$$\Theta_{k|k-1}(\mathbf{x}) = N_b(\mathbf{x}) + \sum_{i=1}^{J_k} p_S(\mathbf{x'}) \cdot f(\mathbf{x}|\mathbf{x'}) \cdot \Theta_{k-1|k-1}(\mathbf{x'}). \tag{9}$$

In the following correction step, the PHD function is adapted according to the currently received measurement set $\mathcal{D}_k$:

$$\begin{aligned} \Theta_{k|k}(\mathbf{x}) \quad &= (1 - p_{det}(\mathbf{x})) \cdot \Theta_{k|k-1}(\mathbf{x}) + \\ &\int \frac{p_{det}(\mathbf{x}) \cdot L_{d_i^k}(\mathbf{x}) \cdot \Theta_{k|k-1}(\mathbf{x})}{\mathcal{C} + \int p_{det}(\mathbf{x}) \cdot L_{d_i^k}(\mathbf{x}) \cdot \Theta_{k|k-1}(\mathbf{x}) d\mathbf{x}} \mathbf{d} d_i^k \end{aligned} \tag{10}$$

where the detection probability p_{det} and the clutter rate C characterize the used human detector, and $L_{d_i^k}(\mathbf{x})$ is the likelihood for a given measurement d_i^k and a state $\mathbf{x}$.

In the used GM-PHD filter, this correction step is performed by generating $(J_{k-1} + |\mathcal{D}_k|) \cdot (1 + |\mathcal{D}_k|)$ new Gaussian distributions. While their mean and covariance values are chosen according to the position of the respective state and detection, the weights of the corrected curves are computed as follows:

$$w_k^{[j]}(d_i^k) = \begin{cases} (1 - p_{det}) \cdot w_{k|k-1}^{[j]}, & \text{no detection} \\ \frac{p_{det}(\mathbf{x}) \cdot L_{d_i^k}(\mathbf{x}) \cdot w_{k|k-1}^{[j]}}{\mathcal{C} + \int p_{det}(\mathbf{x}) \cdot L_{d_i^k}(\mathbf{x}) \cdot w_{k|k-1}^{[j]} d\mathbf{x}}, & \text{for } d_i^k \in \mathcal{D}_k \end{cases} \tag{11}$$

In order to keep the overall number of Gaussians at a suitable level, merging and pruning procedures as proposed in [40] are carried out.

The standard PHD filter does not apply any image information in order to distinguish between objects. In our framework we use a feature-based label tree extension as proposed in a previous work [4]. This extension uses image color information to distinguish objects and performs especially well in case of near objects and occlusions which are present in our scenarios. After this step, object extraction is done by reporting hypotheses with a weight of $\Theta(\mathbf{x}) > T_{extract}$ (usually set to $T_{extract} = 0.5$). From (11), it can be seen that the PHD filter

is sensitive to missed detections. In case no current detection confirms a state estimate, its weight is reduced by the constant factor $(1 - p_{det})$. Should it fall below $T_{extract}$, it will not be reported and the corresponding track will not be continued in this frame. In the following, we will demonstrate how the idea of using crowd density information to complement detection subsequently improves tracking performance.

6. Experimental Results

6.1. Datasets and Experiments

The proposed approach is evaluated within challenging crowded scenes from multiple datasets. In particular, we select some videos from PETS [33], UCF dataset [41], and Data-Driven Crowd Analysis dataset [42]. These videos are annotated for all frames using Viper [43] (except for UCF-879 where the annotation comprises only the first 200 frames).

To demonstrate the effectiveness of the proposed detection algorithm, we compare our results to the baseline algorithm [19]. In particular, two detection thresholds (as τ_{min} and τ_{max}) are tested for the baseline algorithm, whereas the proposed method uses a dynamically chosen threshold between these values according to the crowd density. Additional tests were conducted to assess the impact of the correction filters. For quantitative evaluations, we use the CLEAR metrics [44]: the Multi-Object Detection Accuracy (MODA) and the Multi-Object Detection Precision (MODP).

For the evaluation of the tracking performance, we use the OSPA-T metric proposed in [45]. To demonstrate the impact of improving detection results on tracking, we compare the tracking results in terms of this metric using the baseline detector to the results using our improved detector.

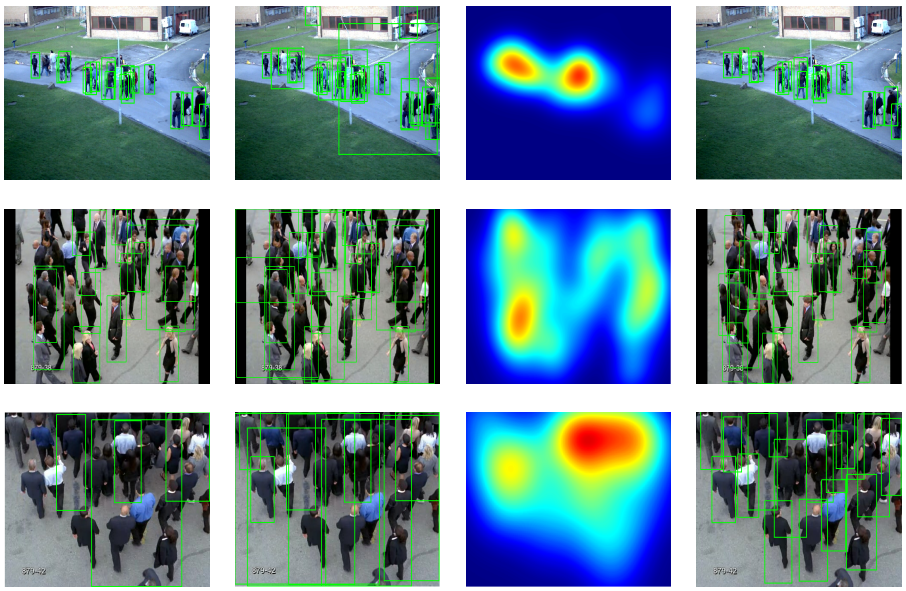
6.2. Results and Analysis

For the detection part, the results using static detection thresholds τ_{min}, τ_{max} (baseline method) are compared to the proposed dynamic threshold $\tau_{dyn} \in$

sequence name	τ_{max}	τ_{min}	τ_{dyn}	Filtering	τ_{dyn}+ Filtering
S1.L1.13-57 (FAST):			0.59/0.59		**0.63/0.63**
S1.L1.13-57 (SIFT):	0.48/0.65$^{(*)}$	0.36/0.57$^{(*)}$	0.59/0.60	0.48/0.66	0.61/0.63
S1.L1.13-57 (GFT):			**0.60/0.60**		0.62/0.63
S1.L1.13-59 (FAST):			**0.60/0.67**		0.60/0.68
S1.L1.13-59 (SIFT):	0.56/0.68$^{(*)}$	0.25/0.61$^{(*)}$	**0.60/0.67**	0.56/0.69	0.60/0.68
S1.L1.13-59 (GFT):			0.59/0.67		**0.61/0.68**
S1.L2.14-31 (FAST):			**0.40/0.59**		**0.47/0.63**
S1.L2.14-31 (SIFT):	0.33/0.63$^{(*)}$	0.09/0.57$^{(*)}$	**0.40/0.59**	0.32/0.65	**0.47/0.63**
S1.L2.14-31 (GFT):			**0.40/0.59**		**0.47/0.63**
S2.L3.14-41 (FAST):			**0.34/0.56**		**0.35/0.57**
S2.L3.14-41 (SIFT):	0.29/0.54$^{(*)}$	0.04/0.56$^{(*)}$	0.34/0.54	0.29/0.54	0.35/0.55
S2.L3.14-41 (GFT):			0.34/0.54		0.36/0.55
UCF-879 (FAST):			0.41/0.55		**0.59/0.58**
UCF-879 (SIFT):	0.44/0.58$^{(*)}$	0.34/0.54$^{(*)}$	0.42/0.55	0.41/0.62	0.57/0.58
UCF-879 (GFT):			**0.43/0.55**		0.58/0.58
INRIA879-42 (FAST):			**0.35/0.55**		**0.42/0.47**
INRIA879-42 (SIFT):	0.27/0.54$^{(*)}$	0.06/0.55$^{(*)}$	**0.35/0.55**	0.20/0.42	0.38/0.45
INRIA879-42 (GFT):			**0.35/0.55**		0.41/0.44

Table 1: MODA / MODP results for three different feature types used in the crowd density estimation (FAST / SIFT / GFT) and for different test videos.

$\{\tau_{min}...\tau_{max}\}$ in Table 1. We set τ_{min} to (-1.2) and τ_{max} to (-0.5), these values have been found empirically suitable for highly-resp. lowly crowded scenes. As shown, the results using (-0.5) as detection threshold are not satisfactory, also by decreasing the threshold to (-1.2), the results are even worse. That is why, we consider that using adjustable detection threshold between these two limits based on local density is a more appropriate method. As shown in the table, the automatic choice of the detection threshold already gives better results than both configurations of the baseline method. Regarding the final results (in the last column), the proposed system using a dynamically chosen detection threshold together with a filtering step based on geometrical characteristics gives the best results for all test videos. These results demonstrate that integrating both proposed steps (filtering and dynamic threshold) into human detector 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. The choice of the feature detector in general does not seem critical to the



(a) (b) (c) (d)

Figure 4: Exemplary visual results comparing the performance of crowd-sensitive threshold to the baseline method: (a) baseline algorithm at τ_{max}, (b) baseline algorithm at τ_{min}, (c) estimated crowd density map, (d) proposed method using dynamically chosen τ_{dyn} computed from the density values and correction filter according to aspect ratio and perceived height. From top to bottom: Frames from PETS, UCF 879, and INRIA 879-38.

performance, expect slight improvement using FAST compared to others.

Figure 4 shows exemplary visual results which also indicate that the performance increases by the proposed method. Although the PETS sequences provide all the same view (View 1), they still pose different problems to the detector. Changing lighting conditions, shadows and different crowd densities between the test sequences are challenging and in all cases the proposed method improves the detection results over the baseline method. Due to the higher crowd density and the tilted camera view, the UCF-879 sequence is even more challenging. However, the proposed method considerably enhances the detection compared to the baseline method. For the INRIA 879-38 sequence,

the camera view is almost completely downward and people are walking very near to the camera which changes their aspect ratio considerably at 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 a prior-knowledge about the shape of a person but is only estimated on previous detections, it is misled in this situation. Accordingly, in this specific case its contribution is smaller.

Since the part-based model represents the current state-of-the-art detector, we consider extending it to operate in crowded scenes and improving its performance is a substantial contribution of this paper. As advantages of our method, the proposed extensions do not need any preliminary learning phase and can be applied on-line. However, it is important to mention that our proposed method is applied best for medium density crowded scenes. It is difficult to perform well for extremely crowded scenes, because the continuous application of a detection algorithm in individual frames will face some difficulties with the small visual information. For such videos, learning scene-specific motion patterns as a global entity will be preferred. Also, only head detection could be more appropriate in such videos, because our method relies on body detection and also the geometrical filters are based on perceived height, which could not work well in these specific cases. To illustrate that, we show our results in one frame from a high crowded video where only heads are visible, see Figure 5.

Figure 5: Detection results from one frame of a video in Data-Driven Crowd Analysis dataset [42]

For tracking, the generated results using the same tracker configuration for all videos to ensure comparability are shown in Table 2. Generally, the results of the proposed method using a dynamical detection threshold and correction filtering are better compared to the baseline method. The gain is especially high for the sequences PETS S1.L2.14-31 and INRIA-879-42 but an overall major improvement can be observed in all videos.

sequence name	original ($\tau = -0.5$)	proposed method
S1.L1.13-57 (FAST):		63.64
S1.L1.13-57 (SIFT):	$65.26^{(*)}$	62.69
S1.L1.13-57 (GFT):		**61.06**
S1.L1.13-59 (FAST):		**62.36**
S1.L1.13-59 (SIFT):	$64.81^{(*)}$	64.61
S1.L1.13-59 (GFT):		64.05
S1.L2.14-31 (FAST):		**66.39**
S1.L2.14-31 (SIFT):	$75.27^{(*)}$	70.82
S1.L2.14-31 (GFT):		71.00
S2.L3.14-41 (FAST):		87.65
S2.L3.14-41 (SIFT):	$88.19^{(*)}$	88.44
S2.L3.14-41 (GFT):		**87.36**
UCF-879 (FAST):		86.89
UCF-879 (SIFT):	89.92	86.95
UCF-879 (GFT):		**86.46**
INRIA-879-42 (FAST):		**73.22**
INRIA-879-42 (SIFT):	$81.15^{(*)}$	75.55
INRIA-879-42 (GFT):		73.56

Table 2: Averaged OSPA-T values for test sequences and different feature types (FAST / SIFT / GFT). We use a cut-off parameter $c = 100$, $\alpha = 30$ and a distance order of $d = 2$.

These results are consistent with our expectations as the tracker relies on improved detections and lower clutter. OSPA-T values change more between different features than the MODA/MODP values due to the filtering effect of the PHD tracker. As the tracker can deal with clutter and also missed detections to some extent, detection improvements enhance the tracking performance but not with the same impact. So it is possible that the tracking results may vary over different feature types, although these may generate similar MODA/MODP results.

The OSPA-T metric for different configurations over one complete scene (PETS S1.L2.14-31) is shown in Fig. 6 (a). For this scene with challenging lighting conditions and medium crowd density, the detection performance is increased considerably 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)-(e) where it can be seen that our method is visibly able to track objects for a longer time and also maintains more tracks than the baseline method.

7. Conclusion

In this paper, we proposed an extension of the part-based human detector by incorporating local crowd density and geometrical correction filters in the non-maximum suppression step and used the resulting detections for tracking. The crowd density information is represented as a new statistical model of spatio-temporal local features that varies temporally over the video and spatially across the frame. By means of automatically estimating crowd density maps, the detection threshold is adjusted according to this contextual information. In order to cope with false positive detections of inappropriate size, dynamically-learning correction filters exploiting the aspect ratio and the perceived height of detections are proposed. None of the proposed extensions need a training phase and both can be applied on-line. An extensive evaluation on several datasets shows the effectiveness of incorporating crowd density into the detection process. Also, tracking performance based on the improved detections is tested.

There are several possible extensions of this work: First, including more contextual information in addition to the crowd density to improve human detection and tracking in crowded scenes might be investigated. Second, 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. Finaly,

the improved detections can be employed for high level analysis such as crowd change detection. This can be achieved using some crowd descriptors defining the topological structure of the detected bounging box over the time.

Acknowledgement

This work was conducted in the framework of the EC funded Network of Excellence VideoSense.

References

[1] M. Hofmann, M. Haag, G. Rigoll, Unified hierarchical multi-object tracking using global data association, in: PETS, 2013.

[2] M. Pätzold, T. Sikora, Real-time person counting by propagating networks flows, in: AVSS, 2011, pp. 66–70. doi:http://doi.ieeecomputersociety.org/10.1109/AVSS.2011.6027296.

[3] M. D. Breitenstein, F. Reichlin, B. Leibe, E. K. Meier, L. V. Gool, Robust tracking-by-detection using a detector confidence particle filter, in: ICCV, 2009.

[4] V. Eiselein, D. Arp, M. Pätzold, T. Sikora, Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors, in: AVSS, 2012.

[5] R. Heras Evangelio, T. Sikora, Complementary background models for the detection of static and moving objects in crowded environments, in: AVSS, 2011.

[6] N. Ihaddadene, C. Djeraba, Real-time crowd motion analysis., in: ICPR, IEEE, 2008, pp. 1–4.
URL http://dblp.uni-trier.de/db/conf/icpr/icpr2008.html#IhaddadeneD08

[7] A. Albiol, M. J. Silla, A. Albiol, J. M. Mossi, Video analysis using corner motion statistics, in: IEEE International Workshop on PETS, 2009, pp. 31–37.

[8] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model., in: CVPR, IEEE, 2009, pp. 935–942.
URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2009.html#MehranOS09

[9] Y. Zhang, L. Qiny, H. Yao, P. Xu, Q. Huang, Beyond particle flow: Bag of trajectory graphs for dense crowd event recognition, in: ICIP, 2013.

[10] F. Raudies, H. Neumann, A bio-inspired, motion-based analysis of crowd behavior attributes relevance to motion transparency, velocity gradients, and motion patterns, PLoS ONE 7 (12).

[11] S. Chiappino, P. Morerio, L. Marcenaro, C. S. Regazzoni, Bio-inspired relevant interaction modelling in cognitive crowd management, Journal of Ambient Intelligence and Humanized Computingdoi:10.1007/s12652-014-0224-0.
URL http://link.springer.com/article/10.1007/s12652-014-0224-0

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

[13] X. Zhao, D. Gong, G. Medioni, Tracking using motion patterns for very crowded scenes, in: ECCV, 2012, pp. 315–328.

[14] L. Kratz, K. Nishino, Tracking with local spatio-temporal motion patterns in extremely crowded scenes, in: IEEE Trans. Pattern Anal. Mach. Intell., 2010, pp. 693–700.

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

[16] H. Su, H. Yang, S. Zheng, Y. Fan, S. Wei, Crowd event perception based on spatio-temporal viscous fluid field, in: Ninth IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2012, Beijing, China, September 18-21, 2012, 2012, pp. 458–463. doi:10.1109/AVSS.2012.32.
URL http://doi.ieeecomputersociety.org/10.1109/AVSS.2012.32

[17] Y. L. Hou, G. K. H. Pang, People counting and human detection in a challenging situation, in: IEEE Transactions on Systems, Man, and Cybernetics, Part A, Vol. 41, 2011, pp. 24–33.

[18] M. Rodriguez, I. Laptev, J. Sivic, J.-Y. Audibert, Density-aware person detection and tracking in crowds, in: ICCV, 2011, pp. 2423–2430.

[19] 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.

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

[21] R. B. Girshick, P. F. Felzenszwalb, D. McAllester, Discriminatively trained deformable part models, release 5, http://people.cs.uchicago.edu/ rbg/latent-release5/.

[22] B. Benfold, I. Reid, Stable multi-target tracking in real-time surveillance video, in: CVPR, 2011, pp. 3457–3464.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

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

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

[30] 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, ICCV Workshops (2011) 150–157.

[31] 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.

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

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

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

[35] R. Mahler, Multitarget bayes filtering via first-order multitarget moments, Aerospace and Electronic Systems, IEEE Transactions on 39 (4) (2003) 1152 – 1178. doi:10.1109/TAES.2003.1261119.

[36] T. Long, L. Zheng, X. Chen, Y. Li, T. Zeng, Improved probabilistic multi-hypothesis tracker for multiple target tracking with switching attribute states., IEEE Transactions on Signal Processing 59 (12) (2011) 5721–5733.
URL http://dblp.uni-trier.de/db/journals/tsp/tsp59.html#LongZCLZ11

[37] J. Vermaak, S. J. Godsill, P. P. rez, Monte carlo filtering for multi-target tracking and data association, IEEE Transactions on Aerospace and Electronic Systems 41 (2005) 309–332.

[38] R. Maher, Advances in Statistical Multisource-Multitarget Information Fusion, Artech House, 2014.
URL http://books.google.fr/books?id=jGbdoAEACAAJ

[39] 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.

[40] D. Clark, B.-N. Vo, Convergence analysis of the gaussian mixture phd filter, in: IEEE Transactions on Signal Processing, Vol. 55, 2007, pp. 1208–1209.

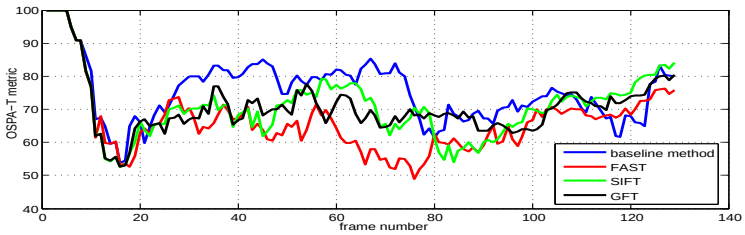
[41] S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: CVPR 07, 2007, pp. 1–6.

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

[43] V. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, D. Doermann, T. Drayer, Performance Evaluation of Object Detection Algorithms, in: ICPR, 2002, pp. 965–969.

[44] 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.

[45] 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.



(a)

(b) (c)

(d) (e)

Figure 6: (a) OSPA-T distance over full sequence PETS S1.L2.14-31 (b)-(e) Exemplary visual tracking results of our proposed method compared to the baseline method from this scene: (b) baseline method, (c) proposed method using FAST features, (d) proposed method using SIFT features, (e) proposed method using GFT