Enhancing Human Detection using Crowd Density Measures and an adaptive Correction Filter

Volker Eiselein [1*], Hajer Fradi [2#], Ivo Keller [*], Thomas Sikora [*], Jean-Luc Dugelay [#]

[*] *Communication Systems Group, Technische Universität Berlin, Germany*
[1] `E-mail: eiselein@nue.tu-berlin.de`

[#] *Multimedia Communications Dept., EURECOM, Sophia Antipolis, France*
[2] `E-mail: fradi@eurecom.fr`

Abstract

In this paper we improve a human detector by means of crowd density information. Human detection is especially challenging in crowded scenes which makes it important to introduce additional knowledge into the detection process. We compute crowd density maps in order to estimate the spatial distribution of people in the scene and show how it is possible to enhance the detection results of a state-of-the-art human detector using this information. The proposed method applies a self-adaptive, dynamic parametrization and as an additional contribution uses scene-adaptive learning of the human aspect ratio in order to reduce false positive detections in crowded areas. We evaluate our method on videos from different datasets and demonstrate how our system achieves better results than the baseline algorithm.

1. Introduction

Automatic detection and tracking of people in video data is a common task in the research area of Video Surveillance and its results lay the foundations for a number of more complex applications such as mugging detection, sports analysis or traffic safety. 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 and include e.g. graph-based approaches ([11], [15]), particle filtering frameworks ([3]) and methods using Random Finite Sets ([6]).

Although there are different approaches to the tracking problem, all of them rely on efficient detectors which have to identify the position of persons in the scene while preventing to generate too many false detections (clutter) in areas without people. Techniques based on background subtraction such as [10] are widely applied thanks to their simplicity and effectiveness but fail in crowded scenes where individual people cannot be distinguished from each other.

While significant recent improvements have been made in the field of object detection and human recognition, crowded scenes remain particularly challenging because of the partial occlusions between individuals. Recent works therefore typically focus on exploiting global level constraints to improve detection or tracking results in crowded scenes [16, 2, 17]. For example, in [2], information about the crowd flow and the scene layout are used to impose constraints for the tracking algorithm.

Similarly, crowd density measures can provide valuable and additional information to enhance person detection in crowded scenes. For instance, in [12], the number of persons is introduced as prior information to the detection step which is formulated as a clustering problem with a known number of clusters. But counting people is by itself a complex task in presence of crowds and occlusions. Besides, using the number of people as a crowd measure has the limitation of giving only global information for the entire image and discarding local information about the crowd.

We therefore resort to crowd information at a local level by computing crowd density maps. This solution 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 [17] has investigated this idea using an energy formulation. However, the authors use the confidence scores from person detection as input to the density estimation which does not introduce complimentary information into the process. In addition, a learning step with a given set of human-annotated ground truth detections is required, which makes the system not fully automatic.

In contrast to the previous work, we intend to demonstrate in this paper the effectiveness of an automatic crowd description provided by crowd density maps in order to enhance human detection results. The proposed crowd density

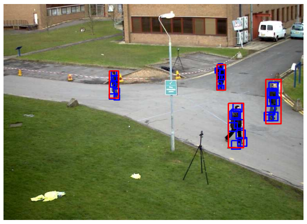


Figure 1. Exemplary human detections using the part-based model from [7]: Blue boxes describe object parts which also contribute to the overall detection (red).

map is typically based on using local features as an observation of a probabilistic crowd function. A feature tracking step is involved in the crowd density process to alleviate the effects of feature components irrelevant to the crowd. For person detection, we apply a part-based model that has been proposed in [7]. Since in crowded scenes, occlusions can occur which hamper the detector's accuracy, we show how the estimated crowd density can overcome this problem. As another contribution of this paper, we design a correction filter based on the aspect ratio of a person in order to deal with false positive detections of wrong size.

The remainder of the paper is organized as follows: In the next section, we introduce the human detector we use. Details on our crowd density framework are given in Section 3 while Section 4 explains how we use this information together with an aspect ratio-based correction filter in order to improve detection results. A detailed evaluation of our work follows in Section5. Finally, we briefly conclude and give an outlook on possible future works.

2. Human Detection

Human detection is a common problem in Computer Vision as it is a key technology to provide a semantic understanding of video data. Accordingly, it has been studied intensively and different approaches have been proposed ([5] [7]) which are often gradient-based.

In our work we use the well-known part-based model from [7]. It is based on histograms of oriented gradients (HoG) [5] and marks the current state-of-the-art. The detector uses a feature vector over multiple scales and a number of smaller parts within the region of interest (RoI) to get additional cues about an object (see Figure 1). A pyramidal extension identifies in a early stage which regions of an image are likely to contain a person and which areas can be discarded in order to speed-up the detection process. In this work we use the implementation from [8] 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 are processed by an additional non-maxima suppression (NMS) step. In the baseline method, regions with high detection scores are kept while detections overlapping with these more than a certain degree are removed.

While this detector works generally well, as other methods it has weaknesses in crowded scenes with occlusions. 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 approach on crowd density estimation.

3. Crowd Density Map Estimation

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. Local features extraction

One of the key aspects of crowd density measurements is crowd feature extraction. Under the assumption that regions of low density crowd tend to present less dense local features compared to a high-density crowd, we propose to extract local feature points as a description measure of the crowd. In our work, we test different types of features for their performance: Features from Accelerated Segment Test (FAST) [19], Scale-Invariant Feature Transform (SIFT) [14], and Good Features to Track (GFT) [21].

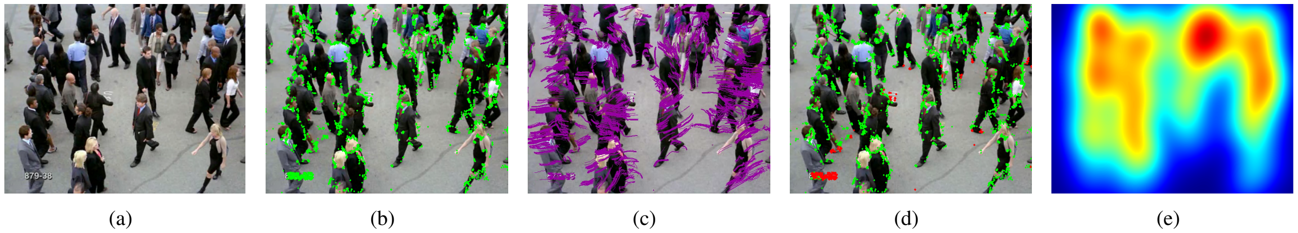
FAST was proposed for corner detection in a reliable way. It has the advantage of being able to find small regions which are outstandingly different from their surrounding pixels. Besides, the use of this feature is motivated by the derived results in [4], showing reliable detection of crowded regions from aerial images using FAST. SIFT is another well-known texture descriptor which defines interest point locations as maxima/minima of the difference of Gaussians in scale-space. Under this respect, SIFT is rather independent on the perceived scale of the considered object which makes it interesting for crowd measurements. These two aforementioned features are compared to the classic "Good Features to Track" feature detector [21], which is based on the detection of corners containing high frequency information in two dimensions.

3.2. Local features tracking

Using the extracted features directly to estimate the crowd density map without a feature selection process might incur at least two problems: firstly the high number of local features increases the computation time of the crowd density. Secondly and more important, the local features contain components irrelevant for crowd density (e.g. background). To reduce this effect, motion information is used.

Motion estimation is performed using the Robust Local Optical Flow (RLOF) [20], which computes very accurate sparse motion fields by means of a robust norm[1].

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



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

Figure 2. Illustration of the proposed crowd density map estimation using local features extraction: (a) Exemplary frame, (b) Local feature points (in this case by FAST algorithm), (c) Feature tracks, (d) Distinction of 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

However, 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 occlusion. While RLOF handles these noise effects better than the standard KLT feature tracker from [23], it still is not prone against all errors. This is why motion information is aggregated to form longterm trajectories.

In every time step, the overall mean motion m_t of a trajectory t is compared to a certain threshold β which is set according to image resolution and camera perspective. Moving features are then identified by the relation $m_t > \beta$ while the others are considered as part of the static background. The advantage of using trajectories instead of computing the motion vectors only between two consecutive frames is that the estimate is more robust to noise and the overall motion information is more accurate. As a result, the number and position of the tracked features undergo an implicit temporal filtering step which improves consistency.

3.3. Kernel density estimation

After generating feature tracks to filter out static points, we define the crowd density map as a kernel density estimate based on the positions of local features. Starting from the assumption of a similar distribution of feature points on the objects, the observation can be made that the closer 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. If we consider a set of K local features extracted from a given image at their respective locations $\{(x_i, y_i), i \in \{1..K\}\}$ the density $C(x, y)$ is defined as follows:

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

where σ is the bandwidth of the 2D Gaussian kernel. The resulting density function is then the crowd density map we will use for further processing.

4. Improving human detection performance

In this section we propose two extensions for the used human detector in order to improve its performance in crowds. Apart from introducing crowd density information, in Section 4.2 we present an adaptive filtering step which additionally enhances the results.

4.1. Integration of Crowd Density information

The usage of detection thresholds in many human detectors can cause difficulties 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 will usually increase the number of detections and allow recognizing more persons, they will also increase the number of false positives. On the other hand, higher thresholds will only detect more reliable candidate regions and might cause the detector to miss people in the scene.

This is especially difficult in heterogeneous scenes with crowded areas where lower thresholds would be suitable due to occlusions. However, higher values reduce the number of false positives in less crowd spaces. It is therefore desirable 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 shown in Section 3, crowd density maps provide exactly this information. We propose therefore to use them to adjust the detection threshold according to the local crowd density.

After the detection step, we obtain a set of candidate RoIs for a given threshold τ: $D(\tau) = \{d_1, d_2, ..., d_n\}$, $d_i = \{x, y, w, h\}$, where x, y denotes the position of the RoI and w, h the respective width and height.

Using a pre-defined range of detection thresholds given by an upper / lower boundary τ_{max}/τ_{min}, we apply the following method of computing a suitable value automatically:

$$\tau_{dyn} = \tau_{min} + (\tau_{max} - \tau_{min}) \cdot \hat{C}(D_i), \quad (2)$$

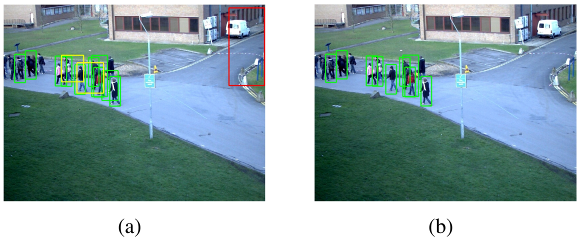


Figure 3. Effects of the proposed correction filter on a frame of the PETS 2009 dataset [9]: (a) detections without filtering, (b) filtering according to aspect ratio. While the unfiltered detections might include too large candidates (red) and also detections comprising several persons at correct height (yellow), the aspect ratio allows removing most of them in a simple and elegant manner.

with

$$\hat{C}(D_i) = \frac{\sum_{j=0}^{h_i-1} \sum_{k=0}^{w_i-1} C(x_i+j, y_i+k)}{w_i \cdot h_i} \tag{3}$$

as the average crowd density value for detection d_i.

The implementation of this procedure can be effectively done as follows: Firstly, a set of candidate RoIs $D_{min}(\tau_{min})$ is computed for the minimal detection threshold. This set contains all possible detections which can be extracted for the given threshold range.

To obtain the dynamic threshold τ_{dyn} for every candidate RoI in D_{min}, the average crowd density $\hat{C}(D_i)$ is computed as in (3) for all regions and inserted into (2). Thresholding using these values then gives a set of result detections which is post-processed by the same consecutive non-maxima suppression step as in the standard method.

4.2. Filtering detections according to aspect ratio

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 the objects (highlighted in yellow in Figure 3 (a)) or that a region is chosen even though it is obviously too large to contain a human (shown in red in Figure 3 (a)). If the score of such a 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 also be recognized. In this case a false positive detection is thus generated while also a number of potentially avoidable missed detections decreases the detection performance. We propose a filtering step in order to cope with such detections of wrong size:

Given a set of candidate RoIs $D = \{d_1, d_2, ..., d_n\}$, we define:

$$r = median(\frac{width(d_i)}{height(d_i)}), i \in 1..n \tag{4}$$

which is computed over all accepted detections. New detection candidates are only accepted if they deviate less than a given threshold Δ_r (in our experiments, $\Delta_r = 0.15$).

As the used NMS step is greedy and overlap-oriented, it is now possible to filter out an unlikely large region and to detect smaller objects in the same area which would have been suppressed otherwise. An example of this correction filter can be seen in Figure 3 (b).

5. Experimental results

For evaluation, we test our method on videos of the following datasets: PETS 2009 [9], UCF dataset[1] and INRIA dataset[18]. As metrics we use the CLEAR metrics proposed in [22]. These are split in two parts: the Multi-Object Detection Accuracy (MODA, N-MODA) and the Multi-Object Detection Precision (MODP, N-MODP).

The first step in computing the metrics for a set of detection RoIs $D = \{d_1, d_2, ..., d_n\}$ and the corresponding ground truth detections $G = \{g_1, g_2, ..., g_n\}$ is to match both sets in order to identify which ground truth detections have been found by the detector. Taking a spatial overlap ratio between all pairs as input, we use the well-known hungarian algorithm [13] for this assignment. As proposed in [22], a threshold of 0.2 for the overlap ratio prevents assignments between badly matching pairs.

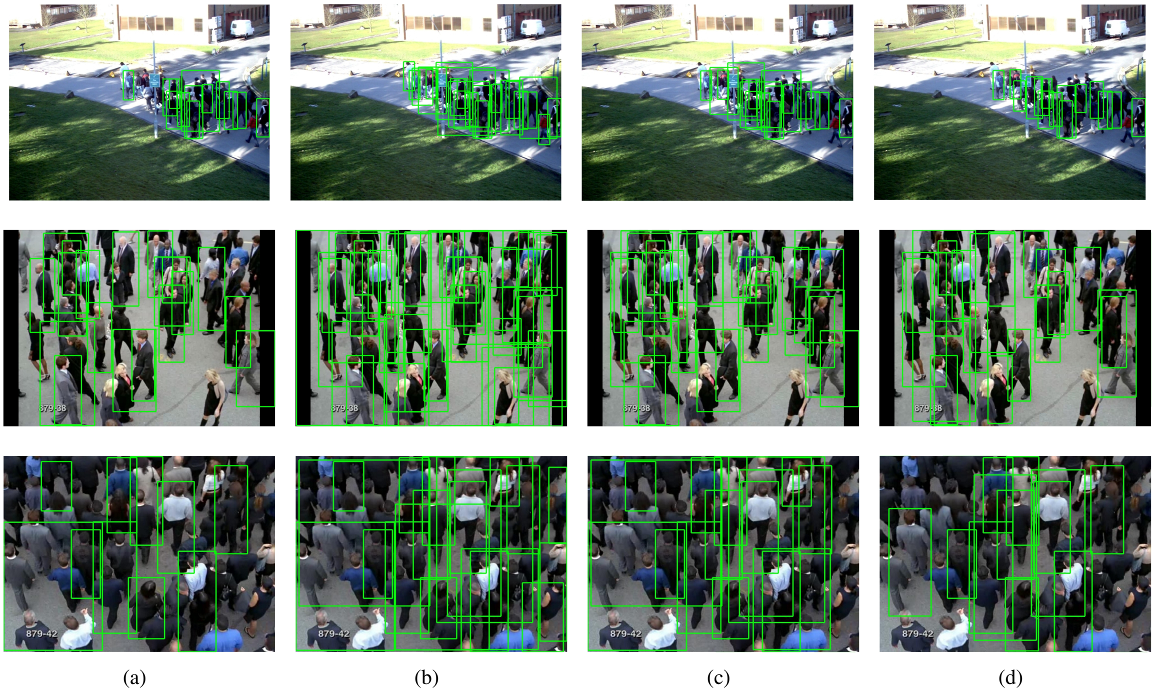
Once the assignment for all frames is done, MODP (t) is computed as the summed and normalized overlap ratio between all assigned pairs in the image while N-MODP gives normalized localization results for the entire sequence.

The N-MODA metric measures the accuracy aspect of the system's performance over the video sequence and is essentially a normalized sum of false positives and missed detections. Both N-MODP and N-MODA illustrate best performance results by a value of one while lower values indicate a worse performance.

For evaluation, the test videos were annotated by hand (UCF 879 annotation comprises the first 200 frames, INRIA 879-38_I annotation the first 300 frames). Results are given in table 1. For the baseline algorithm [7], two detection thresholds (as τ_{min} and τ_{max}) are tested while 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 filter using the aspect ratio of the detections.

In most cases, the automatic choice of the detection threshold already gives better results than both configurations of the baseline method but the proposed system using a dynamically chosen detection threshold and correction by aspect ratio gives best results for almost all test videos. The choice of the feature detector in general does not seem critical with regard to the performance.

Although the PETS2009 sequences provide all the same view (View 1), they still pose different problems to the de-



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

Figure 4. Exemplary visual results show how a crowd-sensitive threshold increases the detection performance compared to the baseline method while the proposed algorithm using an additional correction filter enhances the results further: (a) baseline algorithm at τ_{min}, (b) baseline algorithm at τ_{max}, (c) dynamically chosen τ, (d) Proposed method using dynamically chosen τ and correction filter according to aspect ratio. From Top to bottom: Frames from PETS 2009, UCF 879[1] and INRIA 879-38_I[18]. For PETS and UCF, the proposed method generates more accurate detections and less clutter compared to the baseline method. Results for INRIA are also visibly better but due to the camera view, the effect of the correction filter is small.

tector. 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. Accordingly, the absolute detection results do not reach the values for PETS but the proposed method still enhances the detection considerably compared to the baseline method. For the INRIA 879-38_I sequence, the camera view is almost completely downward and people are walking very close to the camera which makes their aspect ratio change considerably for different positions. Additionally, for this special 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-priori knowledge about the shape of a person but is only trained on previous detections, it is misled in this situation. Accordingly, in this special case its contribution is smaller.

Figure 4 shows exemplary visual results which also indicate the performance increase by the proposed method. As an advantage of our method the proposed extensions do not need a previous learning phase and can be applied on-line.

6. Conclusion

In this paper we present a strategy of exploiting crowd density information to enhance human detection. By means of automatically estimated crowd density maps, the detection threshold of a human detector can be adjusted according to the scene crowd context. In order to cope with false positive detections of inappropriate size, a dynamically-learning correction filter exploiting the aspect ratio of detections is 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 our method. Future works will include enhancements on crowd density estimation in order to get more accurate density values which will in return also increase the effect of the proposed method.

References

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR 07*, pages 1–6, 2007.

[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV (2)*, pages 1–14, 2008.

sequence name	$\tau_{min} = -0.5$	$\tau_{max} = -1.0$	(dyn. $\tau \in \tau_{min}...\tau_{max}$)	proposed (dyn. τ + aspect ratio)
PETS S1.L1 13.57 (FAST):	0.48 / 0.65$^{(*)}$	0.53 / 0.58 $^{(*)}$	0.58 / 0.61	**0.60 / 0.63**
PETS S1.L1 13.57 (SIFT):			0.57 / 0.61	**0.58 / 0.63**
PETS S1.L1 13.57 (GFT):			0.58 / 0.61	**0.59 / 0.63**
PETS S1.L1 13.59 (FAST):	0.56 / 0.69$^{(*)}$	0.56 / 0.63$^{(*)}$	**0.58 / 0.68**	**0.58 / 0.68**
PETS S1.L1 13.59 (SIFT):			**0.59 / 0.68**	**0.59 / 0.68**
PETS S1.L1 13.59 (GFT):			0.59 / 0.67	**0.59 / 0.68**
PETS S1.L2 14.31:	0.33 / 0.63$^{(*)}$	0.34 / 0.58$^{(*)}$	0.40 / 0.60	**0.43 / 0.62**
UCF 879 (FAST):	0.44 / 0.58$^{(*)}$	0.39 / 0.55$^{(*)}$	0.44 / 0.56	**0.47 / 0.58**
UCF 879 (SIFT):			0.44 / 0.56	**0.47 / 0.58**
UCF 879 (GFT):			0.46 / 0.56	**0.48 / 0.58**
INRIA 879-42_I (FAST):	0.27 / 0.54$^{(*)}$	0.30 / 0.53$^{(*)}$	**0.34 / 0.52**	0.33 / 0.52
INRIA 879-42_I (SIFT):			0.34 / 0.52	**0.34 / 0.53**
INRIA 879-42_I (GFT):			0.34 / 0.52	**0.35 / 0.53**

Table 1. N-MODA / N-MODP results for three different feature types used in the crowd density estimation (FAST / SIFT / GFT) and for different test videos. Baseline method [7] using a fixed τ marked by $^{(*)}$. Higher values indicate better performance. The proposed system using dynamical detection thresholds and correction filtering is in all cases among the best results. The performance does not change significantly for different feature types.

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

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

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 2, pages 886–893, 2005.

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

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

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models, 2010.

[9] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *PETS*, pages 1–6, 2009.

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

[11] M. Hofmann, M. Haag, and G. Rigoll. Unified hierarchical multi-object tracking using global data association. In *PETS*, 2013.

[12] Y. L. Hou and G. K. H. Pang. People counting and human detection in a challenging situation. In *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, volume 41, pages 24–33, 2011.

[13] H. Kuhn. The hungarian method for the assigning problem. *Naval Research Logistics Quaterly*, pages 83–87, 1955.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *Int. J. Comput. Vision*, pages 91–110, 2004.

[15] M. Pätzold and T. Sikora. Real-time person counting by propagating networks flows. In *AVSS*, pages 66–70, 2011.

[16] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *ICCV*, pages 1389–1396, 2009.

[17] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, pages 2423–2430, 2011.

[18] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *ICCV*, 2011.

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

[20] T. Senst, V. Eiselein, and T. Sikora. Robust local optical flow for feature tracking. *Transactions on Circuits and Systems for Video Technology*, 09(99), 2012.

[21] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593 –600, 1994.

[22] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan. The clear 2006 evaluation. In *Multimodal Technologies for Perception of Humans*, volume 4122, pages 1–44, 2007.

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