CROWD ANALYSIS IN NON-STATIC CAMERAS USING FEATURE TRACKING AND MULTI-PERSON DENSITY

Tobias Senst, Volker Eiselein, Ivo Keller and Thomas Sikora

Technische Universität Berlin
Communication Systems Group
EN 1, Einsteinufer 17, 10587 Berlin, Germany

ABSTRACT

We propose a new methodology for crowd analysis by introducing the concept of Multi-Person Density. Using a state-of-the-art feature tracking algorithm, representative low-level features and their long-term motion information are extracted and combined into a human detection model. In contrast to previously proposed techniques, the proposed method takes small camera motion into account and is not affected by camera shaking. This increases the robustness of separating crowd features from background and thus opens a whole new field for application of these techniques in non-static CCTV cameras. We show the effectiveness of our approach on various test videos and compare it to state-of-the-art people counting methods.

***Index Terms*—** Crowd Density, Multi Person Density, Feature Tracking, Crowd Analysis, Video Surveillance

1. INTRODUCTION

In recent time, detection and description of dense crowds has become a major field in video surveillance research. Law enforcement agencies and emergency teams expect for the future to get important scene information about people behavior and motion patterns in crowds - if possible, directly on-site and in real-time. Crowd density estimation is also an important aspect for context extraction in a video and can e.g. be used to protect the privacy of people under surveillance by adapting a visual privacy filter according to the level at which people in a crowd can be recognized [1].

While the motivation of crowd description algorithms for security reasons is undisputed, the research field still poses a number of critical challenges: Firstly, although high definition cameras are becoming cheaper and cheaper, many existing CCTV cameras record in low resolution such as 352×240 or 640××480. The need to work in a 24/7 continuous operation usually makes it hard to guarantee good image quality (contrast, brightness), and due to a larger number of cameras being monitored at the same time, the overall processing time should be kept low (i.e. ideally real-time capabilities). As an additional issue, the higher the crowd density, the lower is usually the number of pixels describing a single individual which makes it impossible to apply standard person detectors such as histograms of oriented gradients [2].

In response to these problems, many approaches based on regression techniques have been presented in the past. Chan *et al.* proposed in [3] a system which exploits local crowd features such as segment and texture features acquired from background subtraction in order to determine the number of pedestrians in a crowd. Using perspective normalization and a Gaussian process (GP) regression, the number of people can be estimated and a segmentation within the groups can be performed. In a similar manner, Fradi and Dugelay [4] use GMM-based foreground segmentation and a GP regression in order to infer the relation to the number of pedestrians from it. The result is then weighted using perspective normalization and a density estimate derived from the distribution of FAST [5] features. However, background subtraction-based approaches can cause problems in environments with camera shaking (e.g. pole-mounted outdoor camera under windy weather conditions).

Albiol et al. [6] used corner features of which the motion is computed in order to cluster them into foreground and background features. The authors use the ratio between the number of moving and static points and deduce thus a number of moving persons. A disadvantage is that due to this one-frame motion information only currently moving persons are considered and no camera motion is allowed. In another approach, Fradi and Dugelay [7] compute crowd density derived from a distribution of moving FAST features using kernel density estimation. The resulting density maps give a first impression of highly-crowded locations but do not represent the spatio-geometrical properties of groups and are not suitable for segmentation.

In our approach we introduce a new methodology of extracting crowd features based on long-term motion information in order to achieve independence from the static-camera constraint while still obtaining sufficient information for counting and segmentation of groups. In contrast to previous methods, the number of people is not derived directly by an image-feature regression. Instead, we deduce a Multi-Person

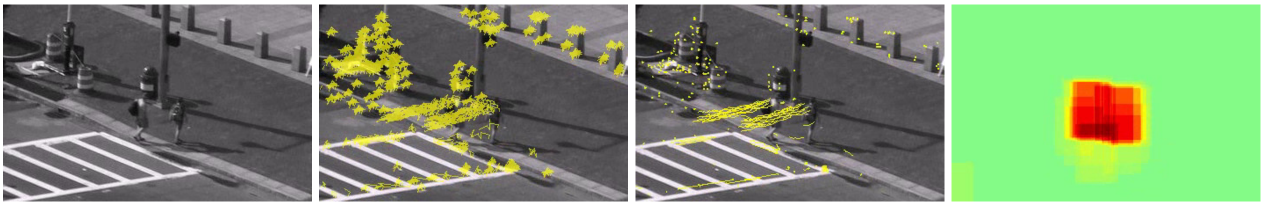

Fig. 1. Processing steps of estimating the proposed Multi-Person density. From left to right: Original video frame, extracted feature points and path lines influenced by camera shaking, stabilized path lines, color-coded Multi-Person density estimate.

Density (MPD) motivated by the Probability Hypothesis Density [8] and compute the number of people by integrating over it. The MPD is estimated using a new image feature-based human model in which the likelihood of a person detection depends on the number of FAST features within a region of interest. This likelihood has previously been trained scene-independently on the well-established CAVIAR[1] dataset.

We will show that the resulting MPD can then be used for further crowd analysis, such as group segmentation or density estimation by integration.

2. MULTI-PERSON DENSITY USING IMAGE FEATURES

We follow the paradigm of Fradi and Dugelay [7] where a relation between the density of local features in the foreground and the underlying crowd density is assumed. Accordingly, this model implies a general link between the number of foreground features and the number of persons in the image. We use this approach in order to motivate our single-person model based on local image features, where the person is assumed to produce a certain number of foreground features. This could be seen as a very simple person detector and is used to build the Multi-Person Density (MPD) from which the crowd properties will be estimated in two steps. An overview of this process is shown in Fig. 1.

Firstly image features will be extracted and classified as foreground by means of their long-term motion information. Trajectories or path lines of the features are therefor estimated using the RLOF[2] sparse optical flow tracker [10]. The result is defined by the path line set T and the unclassified image features $\dot{\mathbf{x}}_0^t, \ldots, \dot{\mathbf{x}}_{N-1}^t$ as the endpoints of the path lines for a time t. Optical flow based trackers can be easily affected by occlusion. As a consequence, features can pile up at objects such as the light post in the PETS 2009 scene and cause high feature densities in these areas. To reject these erroneous points, a forward-backward verification as described in [10] has been applied. This doubles the computational effort but since the number of features to track is low, the absolute runtime is still reasonable.

To classify image features as foreground, their motion can be considered e.g. by thresholding for a minimal average motion as proposed in [7]. This has been shown to be very suitable but is only valid for static cameras. For a more general approach we propose to apply global motion compensation. We assume a homography-based background motion model H^t estimated by the last motion vectors $(\dot{\mathbf{x}}^t - \dot{\mathbf{x}}^{t-1})$. Using H^t, a stabilized set of trajectories $\bar{T}$ is computed for which the background motion component has been removed.

Finally we consider the mean motion over the stabilized path lines and apply Otsu's [11] adaptive thresholding in order to obtain the foreground feature set $S = \{\bar{\mathbf{x}}_0^t, \ldots, \bar{\mathbf{x}}_{N-1}^t\}$. In contrast to methods regarding only the motion between consecutive frames such as [6], long-term motion has a better signal-to-noise ratio. In cases of little overall motion in the image, Otsu's method gives very small threshold values which are not reliable for an overall foreground-background separation. Therefore we use $t_{thresh} = max(t_{Otsu}, t_{min})$ with $t_{min} = 1$ pixel as final threshold yielding the input to the person detector.

The MPD $\mathcal{M}(x, y)$ is now computed in a windowing approach over the image and can be done very efficiently using integral images. In a preliminary step, the expected number of features in a region of interest $\Omega_{x,y}$ is obtained as a Gaussian-distribution $P_1 \sim \mathcal{N}(\mu_1(\Omega), \sigma_1^2(\Omega))$ which has been previously trained with person samples from the public CAVIAR dataset. Note that the distribution P_1 in our approach is related to the area of Ω and can thus be used for different person sizes. For the training step, FAST [5] features have shown to have a lower residual error and lower computational complexity than other point features and are thus used throughout our approach.

Using Ω as the size of a single person, $P_1(n)$ yields the probability of existence for a person given n as the number of features in $\Omega_{x,y}$, but does not scale to multiple persons. In order to obtain the Multi-Person Density in the same area, we thus assume the following relation:

$$\mathcal{M}(x, y) = \sum_{i=1}^{k} i \cdot P_i(n), \;\; n = |S \in \Omega_{x,y}| \qquad (1)$$

with $P_i \sim \mathcal{N}(i \cdot \mu_1(\Omega), i \cdot \sigma_1^2(\Omega))$.

The approximation in Eq. 1 extrapolates the trained re-

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

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

Method	PETS S1.L1 13-57		PETS S1.L1 13-59		PETS S1.L2 14-06		SideWalk	
	MAE	MRE (%)	MAE	MRE (%)	MAE	MRE (%)	MAE	MRE (%)
Albiol et al. [6]	2.80	12.6	3.86	24.9	5.14	26.1	-	-
Conte et al. [9]	1.92	8.7	2.24	17.3	4.66	20.5	-	-
Fradi and Dugelay [4]	1.78	8.6	3.16	19.2	2.89	37.2	-	-
proposed (unstabilized)	1.84 (1.84)	11.4 (11.4)	2.61 (2.95)	16.6 (16.9)	7.88 (7.91)	27.0 (27.0)	0.38 (0.70)	19.2 (45.8)

Table 1. Mean Average Error (MAE) and Mean Relative Error (MRE) for crowd counting application of our method compared to state-of-the-art algorithms.

lation in a way that an increasing number of persons corresponds with an increasing number of features in the given region. Theoretically, k should go to infinite values but in practice, not more than $k = 5$ persons are to be expected in Ω and higher values for k are thus ignored in our approach. The increasing variance for greater i can be justified by the increasing occlusion if multiple persons are present in the region. As a result, the estimated Multi-Person Density $\mathcal{M}$ can be computed for every pixel in the image and in contrast to previous methods includes a-priori knowledge of the shape of a person.

3. CROWD ANALYSIS USING THE MULTI-PERSON DENSITY

In the last section we introduced our concept of a Multi-Person Density (MPD). In the following we will show how this can be used for crowd analysis:

3.1. People Counting

The MPD concept allows to determine the number of persons in a pre-defined region R by integrating over it:

$$N_{people} = f\left(\sum_R \mathcal{M}\right). \tag{2}$$

However, in practice it turns out that the relation between the integral over $\mathcal{M}$ and the number of persons is more complex than a direct proportionality. In order to account for normalization issues, we use a linear function f that includes a normalization component for $\mathcal{M}$ and an offset. The parametrization of f has been learned by a robust regression as described by Conte et al. [9]. Additionally, we follow [9] and in the post-processing of the people count employ a low-pass filter which smoothes the data and helps reducing jumps which might occur due to different trajectory lengths.

3.2. Crowd Density and Segmentation

The MPD allows not only counting but can also directly be used to segment groups in the video sequence. Its important advantage over other methods is that it implicitly integrates information about the human shape and thus reflects the shape of a group better than previous approaches.

A segmentation of the MPD can be done by simple thresholding and should incorporate knowledge of where at least one person has been detected. In relation to the measured single person probability, the threshold is thus chosen to be

$$t_{group} = \sum_{i=1}^{k} i \cdot P_i(\mu_1(\Omega) - \sigma_1(\Omega)) \tag{3}$$

which allows segmentation of a single person with a suitable margin related to the standard deviation of detection. Groups of people can then be identified by a connected component analysis.

The crowd density is defined as the proportion of number of persons for each pixel and is approximated by integrating over a cell C of size (C_N, C_M):

$$density(x, y) = f\left(\frac{1}{C_N \cdot C_M}\sum_C \mathcal{M}(x, y)\right), \tag{4}$$

where the normalization function f is the same as in Eq. 2.

4. EXPERIMENTAL RESULTS

We evaluated our method on a number of video sequences from the PETS 2009 dataset[3] for common, static-camera setups in order to compare it with state-of-the-art approaches. An important advantage of our approach is that it allows for non-static cameras. Unfortunately, almost no public video footage for crowd analysis using shaking-camera setups is available. We therefore use the SideWalk sequence from the Change Detection 2014 dataset[4] although it includes only small groups of persons.

Table 1 shows the people counting results of our method in comparison with state-of-the art algorithms. For PETS we count all persons in the whole image as in [9, 4]. Regarding the static camera sequences, it can be seen that our method overall yields comparable results and achieves a state-of-the-art performance.

Unfortunately, results of other methods for the shaking-camera case are not available. For our evaluation of the

[3] http://www.cvg.rdg.ac.uk/PETS2009/

[4] http://www.changedetection.net

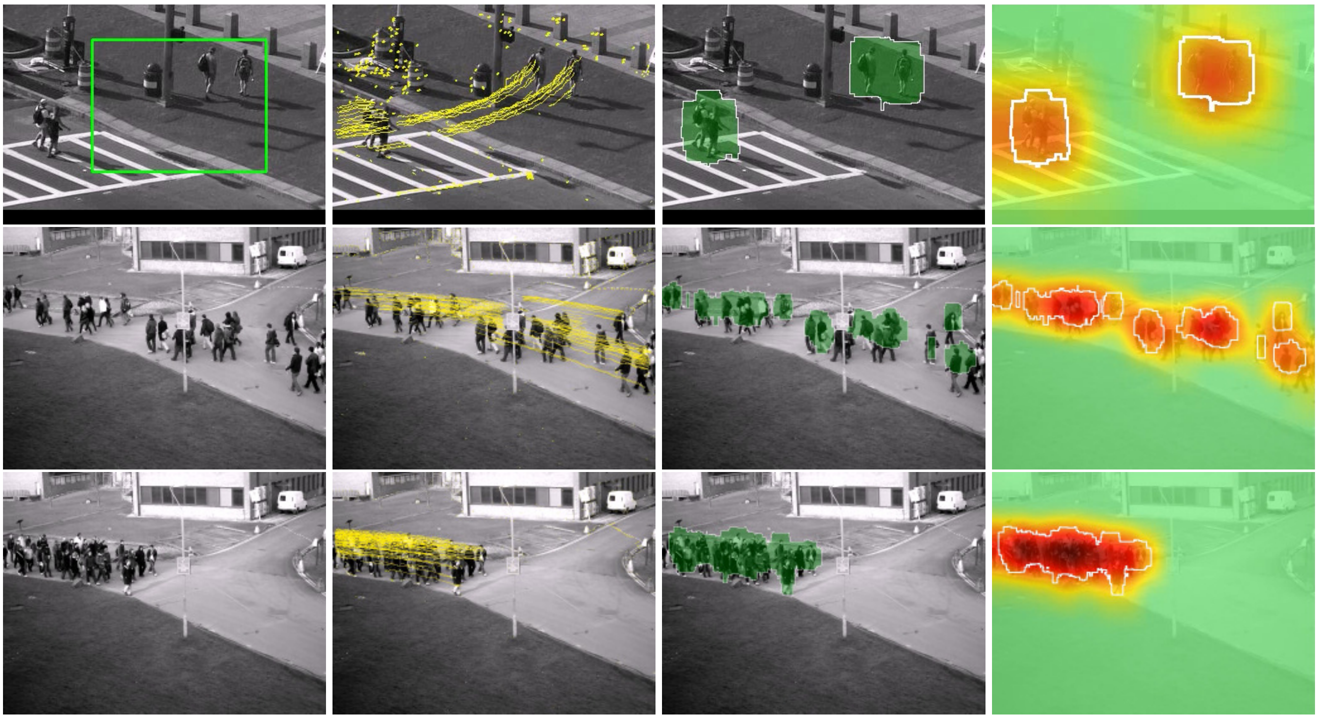

Fig. 2. Visual results of crowd density estimation and group detection using the proposed method. From top to bottom: SideWalk, PETS S1.L1 13-57, PETS S1.L2 14-06. From left to right: Original image with region of interest for SideWalk, stabilized path lines, crowd segmentation, crowd density maps with segmentation.

SideWalk sequence, where we count the number of persons within the area A[x=97, y=38, w=190, h=144] (see Fig. 2), Table 1 shows how the proposed stabilization improves the algorithm's performance for shaking cameras. Though the number of persons in the SideWalk scene is lower than in PETS 2009, the relative error still indicates that our method gives comparable results to the fixed-camera setups. However, it would be desirable to have more video data of non-static cameras in order to see the scalability of this approach.

Visual results for crowd segmentation and crowd density estimation are given in Fig. 2. It can be seen that they reflect both the shape of the groups and the density in an accurate way. It is also visible that the separation between groups can be effectively done by the proposed segmentation step. The light post in the middle of the PETS images degrades the overall performance of our method slightly because path lines are interrupted when people are walking past it. However, the introduced error is limited because new path lines are started when a person re-appears after the occlusion.

With 768×576 for the PETS and 352×240 pixels for the SideWalk sequence, the resolution of the video data used is not high which shows that the proposed method works on standard video data and is not affected by low resolution. The run-time of our method implemented in C++ is approx. 60ms/frame on a standard PC (Intel i7 processor, 3.5 GHz) and thus shows a low computational complexity suitable for real-time applications.

5. CONCLUSION

In this paper we proposed a Multi-Person Density and a new feature tracks-based person filter as a novel concept for crowd analysis in video surveillance applications. The presented method allows the identification of crowded regions and their segmentation while also allowing to count the number of persons in that region.

The usage of feature tracking by means of robust local optical flow reduces the static-camera requirement and enhances the robustness of our method against camera shaking. Evaluation of the proposed method was done on representative video sequences from static and non-static cameras and showed comparable results to state-of-the-art methods.

Acknowledgment

The research leading to these results has received funding from the European Community's FP7 under grant agreement number 261743 (VideoSense) and number 261776 (MOSAIC).

6. REFERENCES

[1] Hajer Fradi, Volker Eiselein, Ivo Keller, Jean-Luc Dugelay, and Thomas Sikora, “Crowd context-dependent privacy protection filters,” in *International Conference on Digital Signal Processing (DSP 2013)*, 2013.

[2] Navneet Dalal and Bill Triggs, “Histograms of oriented gradients for human detection,” in *International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 2005, pp. 886–893.

[3] Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” in *Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008, pp. 1–7.

[4] Hajer Fradi and Jean-Luc Dugelay, “Low level crowd analysis using frame-wise normalized feature for people counting,” in *International Workshop on Information Forensics and Security (WIFS 2012)*, 2012, pp. 246–251.

[5] Edward Rosten and Tom Drummond, “Machine learning for high-speed corner detection,” in *European Conference on Computer Vision (ECCV 2006)*, 2006, pp. 430–443.

[6] Antonio Albiol, Maria J. Silla, Alberto Albiol, and Jose Manuel Mossi, “Video analysis using corners motion analysis,” in *International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2009)*, 2009, pp. 31–38.

[7] H. Fradi and J.-L. Dugelay, “Crowd density map estimation based on feature tracks,” in *International Workshop on Multimedia Signal Processing (MMSP 2013)*, 2013, pp. 40–045.

[8] Ronald P. S. Mahler, “Multitarget bayes filtering via first-order multitarget moments,” *Transactions on Aerospace and Electronic Systems (AESS 2003)*, vol. 39, no. 4, pp. 1152 – 1178, 2003.

[9] Donatello Conte, Pasquale Foggia, Gennaro Percannella, Francesco Tufano, and Mario Vento, “A method for counting moving people in video surveillance videos,” *EURASIP Journal in Advances in Signal Processing*, vol. 2010, 2010.

[10] Tobias Senst, Volker Eiselein, and Thomas Sikora, “Robust local optical flow for feature tracking,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1377–1387, 2012.

[11] Nobuyuki Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man and Cybernetics,*, vol. 9, no. 1, pp. 62–66, 1979.