Towards View Invariant Person Counting and Crowd Density Estimation for Remote Vision-Based Services

Roland Perko, Thomas Schnabel, Alexander Almer and Lucas Paletta

JOANNEUM RESEARCH Forschungsgesellschaft mbH, Graz, Austria
roland.perko@joanneum.at

Abstract

Crowd monitoring in mass events is a highly important technology to support the safety of event attending persons. Proposed methods are often limited to one specific viewing condition and have to be retrained or even redesigned if the viewing angle is changing which is particularly mandatory in airborne based applications. We present a novel framework for highly view invariant person counting and crowd density estimation from single airborne or terrestrial images based on a generalized human head detector and a regression based density estimate. Employing manually labeled reference data, we present detailed accuracy analyses for object detection and for density based person counting. The resulting human counter demonstrates a mean error of 5% over three different data sets. At the same time it thus provides a highly efficient quality indicator for benchmarking security critical decision support services.

1 Introduction

Counting the number of people in single images or in videos taken on large scale events, like music festivals or sport events, is very important to prevent escalations and human casualties [4, 8]. Moreover crowd counting is used for various surveillance purposes, see e.g. [1]. Our envisioned system gathers images from an airborne platform, like an UAV, or an airborne digital camera (e.g. UltraCam [5]). In general such cameras are capturing individual images rather than videos such that our workflow does not rely on motion estimates between adjacent images. In contrast, we propose an object detector custom tailored for people counting that can handle different viewing conditions. The human density is then extracted from the object detection scores which also yields the number of people in the given image. The density estimate itself gives a lot of information as regions of highly crowded people can become most critical when we think of disasters like at Roskilde Festival in 2000, or in Duisburg at Love Parade in 2010. Such information can then be used to alert security staff who then triggers appropriate actions, like opening or closing a gate or restricting the access for following people in a festival scenario. Since recent approaches, like [9, 8], can only handle one fixed viewing angle we want to design a method that gets rid of this constraint and yields appropriate results independent of the viewing direction. Thus, the main point in this paper is that we train one object detection model and one density estimation model using data from different data sets captured under varying view angles. Therefore, our methodology can handle images from different views without the need of constant retraining, with an average relative counting error below 5% w.r.t. the true people count.

Our Contribution. Our main contribution in this work is twofold: First, we introduce an object detector for person detection from remote bird's-eye views that is rather simple and thus fast, nonetheless highly viewpoint invariant. Second, we show that this detector can estimate crowd densities in a highly accurate manner as well as the people count from above. Both estimators are jointly trained on three different data sets showing the ability of generalization, i.e. we learn one model that works for all data sets. The accuracy of both steps is evaluated using appropriate statistical techniques.

2 State of the art

Some principles for crowd monitoring and person counting have been published. For example, [1] count people in an outdoor scenario based on a fixed mounted static video camera using a motion segmentation followed by a feature extraction, that serves as input for a Gaussian regression model. The main drawback w.r.t. our application is the prior motion segmentation. Such a system can only work on video data and can only identify moving people, therefore all standing people are not counted. In addition, other moving objects like cars or pets will also appear in the motion segmentation. The work of [9] deals with airborne nadir looking images. This very interesting approach is similar to our methodology in terms that it extracts local features and uses them to estimate the crowd density. The authors also include a feature selection step to reject local features, which potentially are not corresponding to persons. The density itself is extracted using a kernel density estimate based on the feature occurrence. The number of individu-

als is spatially aggregated also using the feature responses. In the following we discuss related work in particular for object detection, object counting and density estimation.

Object Detection. A standard approach for detecting an object of a known category from single images is to exhaustively analyze the content of image patches at all image positions, and at multiple scales (see e.g. [2, 7, 3, 10]). Each extracted patch is classified according to its local appearance and associated with a detection score. Some of these frameworks yield very good detections on the cost of heavy computational load (especially [2] and [3]).

Object Counting and Density Estimation. One basic idea is to detect each individual object instance in the image and count their number. However, in computer vision object detection is far from being solved and the detection is a harder problem than counting alone. Huge problems arise when objects are overlapping and occlude each other. Thus, the counting by density estimation principle was introduced. The main concept is to estimate an object density function, whose integral over any image region gives the count of objects within this region [6]. In the learning step the proposed methods employ the ground truth location of objects and the learning can be posed as a convex linear or quadratic program. An additional benefit of the method is that after learning the density function can be estimated by simple multiplication of the individual features with learned weights and is therefore very efficient.

3 Methods

Workflow. The main idea is to calculate object detection scores from the given images and relate them to the human density by machine learning techniques. As object detector we propose a customized version of the histogram of oriented gradients (HoGs) detector. The resulting scores are discretized such that the density estimation method is able to learn a weight for each of the scores. Thus, after learning the density function can be calculated by simple multiplications. In addition, the density estimate is a real density function, meaning that the integral over the density yields the object count (therefore, the integral over a subregion holds the number of objects in this particular region). Example images, the object detection scores and the density estimates are visualized in Figure 2.

Object Detection. To enable a view invariant person detection we stick to detecting human heads in images, since those are visible in nadir views as well as in side views. Our proposed object detector is based on the construction of a useful descriptor for an image patch. Those descriptors are then used to train a support vector machine (SVM) that is later employed to calculate a confidence score for each location in the image. As basic descriptor we use the well known HoG descriptor [2] which describes an image patch by the occurrence and magnitude of local gradients. We use the HoG variant reported in [3], since it yields slightly better object detection results while simultaneously having a lower dimensional descriptor compared to the original variant in [2]. For each image patch this implementation results in a vector of dimension $4 + 3 * o$ with o being the number of orientations within the gradient histogram. After initial tests we use 9 orientations which results in a 31-dimensional vector for one HoG cell. The size of a HoG cell is set to 15×15 pixels. As one cell would result in a weak descriptor we use 2×2 HoG cells centered on our object and stack those 4 descriptors which finally yield a 124-dimensional feature vector. It can be considered as a rather low-dimensional description especially when compared to the original HoG-based pedestrian descriptor of [2] with 3780 dimensions. Figure 1 sketches the main concept of our descriptor. Shown are a patch holding a person, the gradient magnitude with the spatial arrangement of the 4 HoG cells and one gradient descriptor.

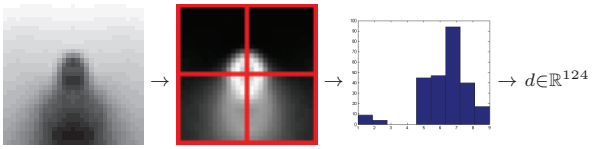

Figure 1: Sketch of the proposed object descriptor. Four HoGs cells are stacked to gather a 124-dimensional feature vector.

For learning we need positive examples extracted from manually labeled objects and negative examples (not holding a person). The positive descriptors are calculated for our manually labeled objects, where we are also incorporating a vertically flipped image to double the number of training data. For each image the same number of negative samples are gathered randomly from the image. To avoid that a negative sample also holds a person, a distance transform is calculated from the positive locations. Then a negative sample must have a distance larger than 1% of the image diagonal (i.e. 18 pixel for an image of size 1440×1080). The descriptors of positive and negative samples are employed to train an SVM, where the resulting model is later used for object detection.

Object Counting and Density Estimation. For counting objects and estimating their density we employ the method in [6]. This method takes densely extracted confidences from our detector and learns the density estimate via a regression to a ground truth density. Thus, each pixel has to be described by a feature vector of the following form $f = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$, which is 1 at the dimension of the corresponding discretized feature and otherwise 0. For density learning our confidences have to be discretized, which is done by setting the minimal value to -2 and the maximal value to $+4$. These bounds are used to scale the confidences to $[0, 255] \in \mathbb{N}$. Now, each of the possible 256 values define a feature vector, as discussed above, which is 1 at the position of the confidence value. Therefore, it yields 256 individual features. The training itself min-

imizes the regularized *Maximum Excess over SubArrays (MESA)* distance (cf. [6]) where we use two distinct approaches to solve the resulting linear or quadratic equation system, namely the L_1 and the Tikhonov regularization (i.e. $\min_x ||Ax - b||$ or $\min_x ||Ax - b|| + ||(x'\Gamma x)/2||$ with $||x \geq 0||$ and Tikhonov matrix Γ being the identity matrix in our case). All details of this methodology are given in [6]. The result is a weight for each of the discretized features and the resulting human density is calculated by multiplying the according weight with the extracted feature value. Thus, for each pixel the density function is given and the sum over all pixels represents the number of objects in the image, i.e. our person count. Therefore, in the testing phase the discretized features, i.e. our object detection scores, are extracted for each image and multiplied by the learned weight vector, directly resulting in the density estimation per pixel and corresponding person count. It should be noted that this approach introduces virtually no overhead over feature extraction [6].

4 Results

Test Data. For evaluation of the presented concept videos from three different scenarios were acquired in HD quality. Only individual video frames were used to simulate our envisioned airborne acquisition. Data from other tests showed that the images are analogous to images taken by an aerial platform. Exemplary images are shown in Figure 2. This Figure shows input images, the object detection score and the density estimate. The first scenario, referred as *Hubsteiger*, originates from a fire drill where we positioned a AXIS P3364 camera on a picker at approximate 25 meters above ground. The images of this camera contain fish eye distortion and persons are observed under a slightly oblique look angle. The second one, referred as *Lakeside*, originates from a music festival in Styria, Austria. A Canon HV30 video camera was mounted on a tower (approximately 30 meters above ground). Here the crowd is sensed under a flat look angle of about $14°$, such that the whole silhouettes of persons are visible. The third one, referred as *Towercam*, originates from the same fire drill as *Hubsteiger* but here a NOKIA Lumia 710 mobile phone was mounted on the top of a building at about 40 meters to capture the crowd in nadir direction. Finally, as we want to show the ability of generalization we constructed a combined data set that contains all images from data set 1 to 3. Even though the presented sequences are not taken from an airborne platform, the images have very similar properties as expected from UAVs or other sensing devices. Therefore, the presented workflow will also work on airborne imagery. We manually labeled 170 images to get the ground truth person counts for training and later for the testing phase (overall more than 43000 persons were annotated, cf. Table 1). From the standard deviation of the people count in Table 1 it can be seen that DS1 *Hubsteiger* is the most difficult data set, as the number of people changes most dramatically.

ID	Number of images	Persons				
		total	min	max	mean	std
DS1: Hubsteiger	45	11508	15	317	255.7	84.7
DS2: Lakeside	80	22300	249	319	278.8	13.4
DS3: Towercam	45	9468	144	263	210.4	33.4
ALL	170	43276	15	319	254.6	55.1

Table 1: Manually labeled persons in the three data sets together with their statistics. This information serves as ground truth for training and for testing.

Object Detection. To evaluate the object detection accuracy we extracted descriptors from positive and negative samples, for each data set and for the combined set (note, that the learning of the combined set involves huge amounts of data, i.e. more than 173000 124-dimensional vectors holding positive and negative samples). Then, we learned SVM models and calculated the average accuracy by a 5-fold cross validation. For each run 4-folds were used to train the model and 1-fold served for testing. We also compared a linear SVM to a SVM with a radial basis function (RBF) kernel. For the RBF case we also varied two parameters $\gamma \in [0.5; 1; 2]$ and $c \in [2, 4, 8]$ (with γ being a parameter of the RBF kernel function for two samples x_i and x_j with $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$ and c being a regularization parameter). While the linear SVM yields accuracies from 93% to 97%, the RBF SVM performs better with 98.5% to 99.6% (best results were achieved with $\gamma = 2$ and $c = 4$). Since the RBF SVM always achieves higher accuracies, this kernel was used for the density estimation later on. The detector for the combined data set gives nice results, with an accuracy close to 99% using the RBF SVM. Therefore, one detecor will be enough to process all given data sets. For training the final object detector we randomly selected 20% of positive and negative samples from all data sets and trained a RBF SVM with the parameters stated above.

Object Counting and Density Estimation. The accuracy for counting by density estimation of the training and testing process is listed in Table 2. We first used all available image to train the density estimation model. Then we took every second image, then every fourth and so on, while the testing of the model was performed on the remaining images. It can be observed that the accuracy of training increases with a lower number of training samples. This makes sense, as the model adapts more and more to the specific samples but looses its ability for generalization (i.e. the well known over-fitting problem). That is why the accuracy of testing is decreasing with a lower number of training samples. Thus, we can learn that about 20 images are sufficient for training our system. We can also observe that the regularization (L_1 or Tikhonov) only has a small effect on the testing results. Overall, we can state that an average error of human counts of about 12 can be reached, which correspond to a relative error below 5% (to be precise 4.7%). Figure 2 visualizes some density esti-

mates and Figure 3 shows the results when using every 4^{th} image for learning. Shown are the estimated human count for the two regularizations given in blue and green color, together with the manually measured counts shown as red dots. The dashed black lines show the separation between the three data sets. Overall, it can be seen that the estimation is quite close to the ground truth data. Especially for data set 1 *Hubsteiger* our framework is also able to get good estimates when a lower number of people populates the scene (e.g. when people are entering the area in the first few images and when they leave from image number 40 to 45; cf. Figure 3).

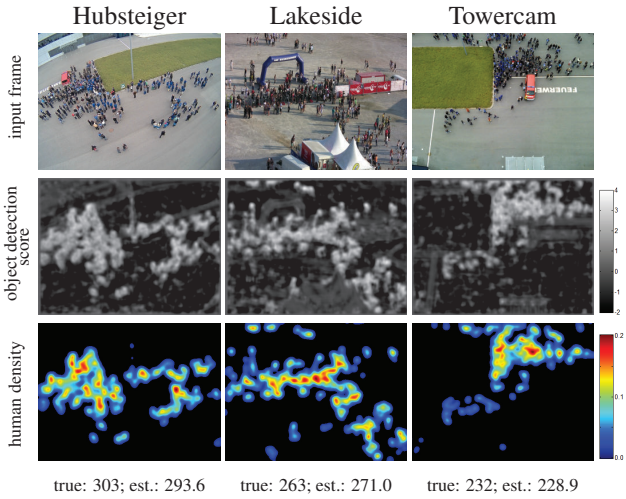

Figure 2: Exemplary results for the three different data sets. The top row shows a representative input image for each data set. Middle row gives the densely extracted object detection scores which are in the range from -2 to +4. Bottom row visualizes the calculated human density functions scaled from 0.0 (blue) to 0.2 (red) in persons/pixel. The true number of persons in the images and the estimated count using the L_1 regularization are stated below.

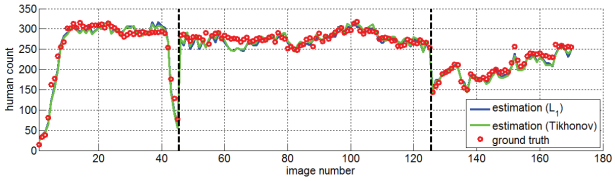

Figure 3: Person counting: Estimated person count using L_1 regularization (blue) and Tikhonov regularization (green). The red dots indicate the manually measured ground truth. The vertical lines show the transition between the data sets 1, 2 and 3.

5 Conclusion

In this work we presented a method for highly view invariant people counting and crowd density estimation from single airborne or terrestrial images. We introduced a generalized human head detector, whose detection scores can successfully be used to derive a crowd density estimate and from this the person count. Overall, the estimated human counts were highly accurate and resulted in 5% average counting error for three test data sets, which were acquired with different cameras under varying viewing angles. The proposed framework is therefore highly important as well as promising for the application in airborne security applications. In future the framework will be enhanced to also allow scale independent processing. The simplest solution would be to rescale any image to a common fixed scale employing the meta-information gathered from the intrinsic camera parameters, GPS and IMU.

step	training			testing		
	#	L_1	Tikhonov	#	L_1	Tikhonov
1	170	10.3	10.7	0	-	-
2	85	10.0	10.9	85	11.3	11.0
4	43	8.7	9.6	127	11.2	11.3
8	22	7.1	7.4	148	12.2	12.0
16	11	6.2	5.3	159	13.1	13.2
32	6	4.7	3.8	164	13.2	12.9
64	3	2.0	2.4	167	14.9	13.2
128	2	1.2	0.5	168	37.2	23.1

Table 2: Accuracy of density learning and testing. Given are the average errors of the total human count over the training and test images, for two regularization options and different training and test set splits. A count error of 10 represents an relative error of 4%.

Acknowledgments. This work has been funded by the Ministry of Austria for Transport, Innovation and Technology within the security research program KIRAS: Project 832353 "EN MASSE: System for multi-sensor crowd monitoring for real time visualization of a common operational picture (COP) and short-time forecast" and Project 845479: "MONITOR: Near real-time multi-sensor monitoring and short-term forecasts to support the safety management at mass events.". The authors would like to acknowledge Domen Tabernik and Matej Kristan for very fruitful discussions on the object detector.

References

[1] A.B. Chan, Z.-S.J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, pages 1–7, 2008.

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

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[4] D. Helbing, A. Johansson, and H.Z. Al-Abideen. Dynamics of crowd disasters: An empirical study. *Physical Review E*, 75:046109, 2007.

[5] F. Leberl, M. Gruber, M. Ponticelli, S. Bernögger, and R. Perko. The UltraCam large format aerial digital camera system. In *ASPRS*, 2003.

[6] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *NIPS*, number 23, pages 1324–1332. 2010.

[7] R. Perko and A. Leonardis. A framework for visual-context-aware object detection in still images. *CVIU*, 114(6):700–711, 2010.

[8] R. Perko, T. Schnabel, G. Fritz, A. Almer, and L. Paletta. Airborne based high performance crowd monitoring for security applications. In *SCIA*, volume 7944, pages 664–674. 2013.

[9] B. Sirmaçek and P. Reinartz. Automatic crowd density and motion analysis in airborne image sequences based on a probabilistic framework. In *ICCV Workshops*, pages 898–905, 2011.

[10] D. Tabernik, M. Kristan, M. Boben, and A. Leonardis. Learning statistically relevant edge structure improves low-level visual descriptors. In *ICPR*, 2012.