A Multi-configuration Part-based Person Detector

Alvaro Garcia-Martin[1], Ruben Heras Evangelio[2] and Thomas Sikora[2]
[1]*Video Processing and Understanding Lab, Universidad Autonoma de Madrid, Spain*
[2]*Communication Systems Group, Technische Universitat Berlin, Germany*
alvaro.garcia@uam.es,{heras, sikora}@nue.tu-berlin.de

Keywords: People Detection, Part-based Detector, Multi-configuration Body Parts.

Abstract: People detection is a task that has generated a great interest in the computer vision and specially in the surveillance community. One of the main problems of this task in crowded scenarios is the high number of occlusions deriving from persons appearing in groups. In this paper, we address this problem by combining individual body part detectors in a statistical driven way in order to be able to detect persons even in case of failure of any detection of the body parts, i.e., we propose a generic scheme to deal with partial occlusions. We demonstrate the validity of our approach and compare it with other state of the art approaches on several public datasets. In our experiments we consider sequences with different complexities in terms of occupation and therefore with different number of people present in the scene, in order to highlight the benefits and difficulties of the approaches considered for evaluation. The results show that our approach improves the results provided by state of the art approaches specially in the case of crowded scenes.

1 INTRODUCTION

Within the computer vision field, particularly in the research area of digital image and video processing, there exists a rich variety of algorithms for segmentation, object detection, event recognition, etc, which are being used in security systems. The ability to detect people in video and in particular detecting people in crowded scenarios is the key to a number of multiple applications including video surveillance, group behavior modeling, crowd disaster prevention, etc. Due to the rise in popularity of these applications over the last years, people detection has gradually experienced a great development. In parallel, interest on reliable strategies to assess the quality of people detection has also grown.

Currently, many different systems exist which try to solve the problem posed by the task of detecting people. The state of the art includes several successful solutions working in specific and constrained scenarios. However, the detection of people in real world scenarios such as airports, malls, etc, is still a highly challenging task due to the multiple appearances that different persons may have, heavy occlusions, specially in crowded scenarios, view variations and background variability.

The work presented in this paper has been focused on the improvement of people detection in crowded scenarios. To that aim, we use a part-based person model and propose a statistical driven way of combining the individual body part detectors in order to detect persons even in case of failure on the detection of any of the body parts. Thus, we are able to detect people with nearly the same reliability whether they are completely visible (people in front of the group) or only partially visible (people behind). We validate our approach and compare with the state of the art on challenging crowded scenes from multiple public video datasets.

The remainder of this paper is structured as follows: Section 2 describes the related state of the art; Sections 3 describes the proposed approach; Section 4 discusses the experimental results. Finally, Section 5 summarizes the main conclusions and future work.

2 STATE OF THE ART

People detection methods from the state of the art perform well in scenes with relatively small number of pedestrians (Dollár et al., 2012b; Enzweiler and Gavrila, 2009; Gerónimo et al., 2010), but these methods usually fail or significantly reduce their performance in scenes with many subjects that partially occlude each other. Various solutions or improvements have been proposed in order to deal with the occlusion

limitations, including directly intrinsic modifications of the person model or using information from additional sources.

Most of the methods in the existing literature are based on the appearance of the object of interest, i.e., a person. Appearance based approaches can be classified attending to the complexity of the model. Simple person models define the person as a region or shape, which can be described by means of a holistic model (Dalal and Triggs, 2005; Dollár et al., 2012a). Complex models define the person as collection of multiple regions or shapes, i.e., part-based models, which can be combined in order to be more flexible regarding different poses and to support partial occlusions (Felzenszwalb et al., 2010; Leibe et al., 2005). However, even such approaches have difficulties in dense environments. The Implicit Shape Model (ISM) of (Leibe et al., 2005) is improved in (Seemann et al., 2007) by using a probabilistic formulation in order to generate a model that is scalable from a general object–class detector into a specific object–instance detector, thus making the detection more reliable. The detector in (Felzenszwalb et al., 2010) is improved in (Girshick et al., 2011) by using a grammar model which includes an additional "body part" simulating possible occlusions. Also based on (Felzenszwalb et al., 2010), in (Tang et al., 2014) a joint model is proposed, which is trained to detect single people as well as pairs of people under varying degrees of occlusion.

Other approaches make use of additional external information to the person model in order to increase the detection performance in crowded scenarios. The most typical ones include tracking (Garcia-Martin and Martinez, 2012), motion (Patzold et al., 2010), depth or 3D information, etc. The use of person density estimation to improve person localization and tracking performance in crowded scenes is proposed in (Rodriguez et al., 2011). In (Milan et al., 2014) a continuous energy minimization framework for multi-target tracking, which includes explicit occlusion reasoning and appearance modeling, is presented. Nevertheless, in the work presented in this paper, we will disregard any possible additional improvement which could be achieved by using external information to the person model. Instead of that, we concentrate on the person model itself.

Most closely related to our work is the approach in (Girshick et al., 2011), which demonstrates the advantages of taking into account in the person model the possibility of failure or occlusion of some body parts. In our case, we do not specifically train the model to capture specific occlusion patterns. We define a more generic scheme in which the absence of any particular body part can be modelled by defining multiple configurations of the part-based models learned during the training phase. Therefore, we are able to deal with occlusions by automatically selecting which of all the possible person model configurations adjust better to any kind of occlusion. In particular, we solve the problem posed to the approach in (Girshick et al., 2011) by crowded scenarios, where the range of possible different occlusions is much bigger and, therefore, the complexity of the grammar model and its training increases exponentially.

There are also other approaches that make use of person models based only on some parts of the body as the head (Ali and Dailey, 2012) or head and shoulders (Zeng and Ma, 2010) since these are the most visible parts in crowded scenarios. Our solution can be considered as a generalization where any possible body part configuration is evaluated in order to take advantage not only on this specific simplified models but also any possible useful configuration.

3 APPROACH

Our proposed approach is based on the detector presented in (Felzenszwalb et al., 2010) but, instead of using the confidence provided by each of the individual body-part detectors for every person candidate, we define several body-part detectors configurations in order to robustly cope with partial occlusions, which profusely appear in crowded scenarios.

3.1 Base algorithm

The detector in (Felzenszwalb et al., 2010) is a part-based person model. It consists of mixtures of multi-scale deformable part models in a star-structure defined by a root model, where the root and each of the deformable body parts are modeled by a HOG as firstly proposed in (Dalal and Triggs, 2005).

The detector proposed in (Felzenszwalb et al., 2010) defines N body parts positioned around the root filter ($n = 0$), which models the appearance of the whole body. The N body parts are computed at twice the resolution in relation to the root filter in order to refine the detection based only on the root information. Each of the n detectors, included the root ($n = 0, ..., N$), is modeled by a 3-tuple (F_n, $v_{n,0}$, d_n), where F_n is the HOG filter response (detection confidence) for part n; $v_{n,0}$ is a two-dimensional vector defining the relative position of part n with respect to the anchor position (x_0, y_0) of the root; and d_n is a four-dimensional vector specifying coefficients of a quadratic function defining the cost for each possible

placement of the part relative to the anchor position. The $BP_n(x,y,s)$ represents the confidence at pixel position (x,y) for body part n ($n = 0,...,N$) associated to scale s ($s = 1,...,S$). Thus, the confidence score for part n at scale s is given as

$$BP_n(x,y,s) = F_n(x,y,s) - \langle d_n, \phi(dx_n, dy_n) \rangle \quad (1)$$

with

$$(dx_n, dy_n) = (x_n, y_n) - (2(x_0, y_0) + v_{n,0}) \quad (2)$$

giving the displacement of part n relative to the anchor and

$$\phi(dx,dy) = \left(dx, dy, dx^2, dy^2\right) \quad (3)$$

defining the potential spatial deformation distributions. Figure 1-(c) shows one example of a multi-part person model with $N = 9$.

The final detection confidence or score $C(x,y,s)$ is computed as the sum of the root and N body parts at each pixel position and scale.

$$C(x,y,s) = \sum_{n=0}^{N} BP_n(x,y,s) \quad (4)$$

The final multi-scale detection hypotheses are extracted after a thresholding followed by a non-maximum suppression process, used to eliminate possible repeated detections. The chosen threshold or minimum score required in order to consider the detected object as a person depends directly on the total number of body parts detections.

3.2 Multiple person model configurations

The previously described approach (Felzenszwalb et al., 2010) is based on the detection of several parts and the combination of all of them. Since the total score depends tightly on the number of parts detected, this approach is not able to reliably cope with occlusions. Therefore, it fails to detect people in groups, where most of the persons are only partially visible.

In order to cope with any kind of body part occlusion, we propose to use multiple person model configurations t ($t = 1,...,T$) with $1 \leq T \leq 2^N$, where each person model configuration t consist of a subset of M body parts ($m = 1,...,M$), with $m \subset n$ of the original detector (Felzenszwalb et al., 2010) and $1 \leq M \leq N$. Thus, the confidence for each configuration is defined as

$$C_t(x,y,s) = \sum_{n=0}^{N} \alpha_n^t \cdot BP_n(x,y,s) \quad (5)$$

where α^t is a binary selector vector for each configuration t

$$\alpha_n^t = \begin{cases} 1 & , n \subset t \\ 0 & , otherwise \end{cases} \quad (6)$$

As in the base algorithm, the final multi-scale detection hypotheses are extracted after a thresholding followed by a non-maximum suppression process in order to eliminate possible repeated detections. However, there are two main differences in our approach with respect to the base algorithm. In first place, there is not only one detection threshold, but there is one for each configuration. Each minimum score required is chosen to be coherent with the number and kind of body parts taken into consideration (see Section 3.3). In second place, we apply the non-maximum suppression process to the results provided by all the person model configurations together.

3.3 Body parts contributions

Once defined the different person model configurations, it is necessary to determine the decision threshold or minimum score required for each configuration in relation to the threshold used if considering the whole set of body parts. To that aim, let consider the confidence or score of each body part n as a continuous random variable BP_n and its associated probability density function $f_{BP_n}(bp_n)$, the final detection confidence as a continuous random variable C and its associated probability density function $f_C(c)$. The minimum confidence k required to consider a detected object as a person corresponds to the probability of $F_C(k) = P(C \leq k)$.

Analogously, each configuration confidence can be considered as a continuous random variable C_t with an associated probability density function $f_{C_t}(c_t)$. In order to estimate the minimum confidence k_t required for each configuration, it is necessary to determine a correction factor R_t that takes into account the number of body parts included in each configuration and their respective contribution or information relevance in relation to the original configuration with N parts. For example, assuming that all the body parts had the same contribution the correction factor $R_t = \frac{1}{N}$ could be used for each configuration t. Nevertheless, since the individual part detectors are not equally discriminative, their contribution to the overall model can not be considered the same.

Therefore, in order to estimate the contribution of each body part n, we first estimate the similarity of the distribution of the scores obtained by using the configuration with the whole set of body parts (F_C), with the distribution of the scores obtained by using

the configuration with all except the considered body part n. To that aim, we use the Kullback-Leibler Divergence (D_{KL}) (Kullback and Leibler, 1951) and define the similarity KL_n between each body part BP_n to the complete model C as the Kullback-Leibler Divergence between the distribution F_C and the distribution without that body part n, $F_{C'}$:

$$KL_n = D_{KL}(F_C||F_{C'}), \text{ being } C' = \sum_{i=0, i\neq n}^{N} BP_i \quad (7)$$

This measure is normalized $\bar{KL}_n$ so that $\sum_{n=1}^{N} \bar{KL}_n = 1$. Finally, the correction factor R_t is computed as the accumulative body parts contributions:

$$R_t = \sum_{n=0}^{N} \alpha_n^t \cdot \bar{KL}_n \quad (8)$$

A factor of $R = 1$ means that there is not necessary any correction on the decision threshold because the considered configuration corresponds to the use of all the body parts.

The minimum confidence k_t required for each configuration t with associated probability $F_{C_t}(k_t) = P(C_t \leq k_t)$ is modified according to the original person model confidence k and the corresponding correction factor R_t:

$$F_{C_t}(k_t) = 1 - R_t(1 - F_C(k)) \quad (9)$$

Therefore, the final probability $F_{C_t}(k_t)$ required for each configuration is defined between the original $F_C(k)$ and 1 ($F_C(k) \leq F_{C_t}(k_t) \leq 1$). The simpler the person model (less body parts), the higher the probability (i.e., the confidence) required to detect a person and vice versa. Figure 1 shows examples of person model configurations, distributions, corresponding correction factors and minimum confidence required k_t.

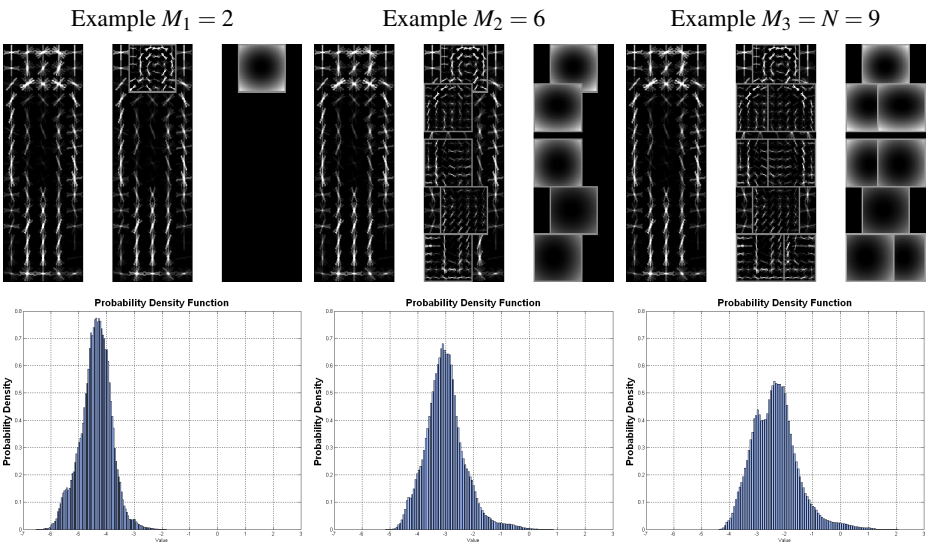
4 EXPERIMENTAL RESULTS

In order to evaluate our people detection approach, we have tested it across several publicly available datasets and compare its results with those provided by the base algorithm DTDP (Discriminatively Trained Deformable Parts (Felzenszwalb et al., 2010)), with those provided by the ISM (Implicit Shape Model (Leibe et al., 2005)) and those provided by one of the most recent and cited people detection approaches from the state of the art ACF (Aggregate Channel Features (Dollár et al., 2012a)). While the DTDP and ISM are part-based detectors, the ACF is a holistic approach.

As presented in this paper, our proposed approach consists of multiple person model configurations. In particular, according to the person model of nine body parts (see Figure 1), we define twenty different configurations ($T = 20$). Every configuration includes at least the root and head body parts, from this basic configuration ($t = 1$), we add progressively configurations with consecutive body parts, i.e., root-head-left shoulder, root-head-left and right shoulder, root-head-left shoulder and left trunk, etc. The last configuration ($t = 20$) includes all the body parts and corresponds to the original configuration of the base algorithm DTDP (Felzenszwalb et al., 2010). The body part contributions (see section 3.3) have been trained using the code provided in (Girshick et al.,) and the INRIA dataset (Dalal and Triggs, 2005). The DTDP results have been obtained using the available code (Girshick et al.,) and the ACF results have been obtained using the available code (Dollár et al., 2012a).

We evaluate all three methods on eleven challenging, publicly available video sequences with ground truth (Milan et al., 2014) (TUD-Stadtmitte (Andriluka et al., 2010), TUD-Campus and TUD-Crossing (Andriluka et al., 2008), S1L1 (1 and 2), S1L2 (1 and 2), S2L1, S2L2, S2L3 and S3L1). The first three sequences are recorded in real-world busy streets, the complexity in terms of crowd or occlusions is medium or low (less than 10 pedestrians are present simultaneously). The last eight sequences are part of the PETS 2009/2010 benchmark (PETS,). We only use the first view of each sequence in all our experiments. They are recorded outdoors from an elevated viewpoint, corresponding to a typical surveillance setup. The sequences are classified originally according to three scenarios (S1, S2 and S3) and three progressive difficulty levels (L1, L2 and L3) for each scenario. These scenarios include higher complexity in terms of crowds and occlusions than the previous ones (generally more than 10 pedestrians are present simultaneously).

We classify the whole set of sequences independently of the original scenario purpose (TUD sequences for people detection, S1 for person count and density estimation, S2 for people tracking and S3 for flow analysis and event recognition). In our experiments, we classify the sequences according to the number of people present simultaneously and, therefore, the degree of occupation of the scene (low, medium or high). Table 1 includes a description of each sequence in terms of occupation (number of pedestrian present simultaneously) and complexity classification. Figure 2 shows sample images of

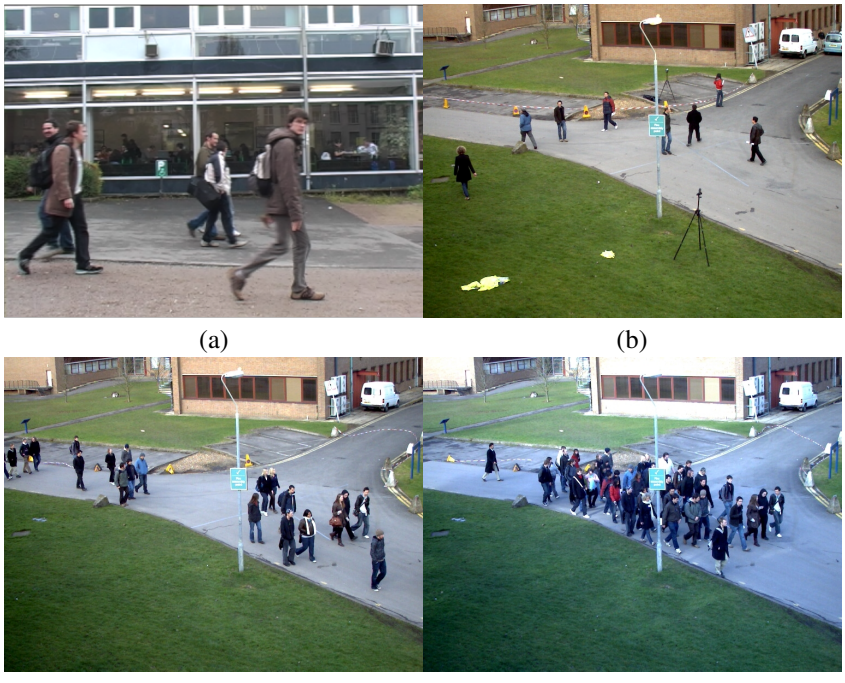


$0 \leq R_1 \leq R_2$ $\quad R_1 \leq R_2 \leq R_3$ $\quad R_3 = 1$

$F_{C_2}(k_2) \leq F_{C_1}(k_1) \leq 1$ $\quad F_{C_3}(k_3) \leq F_{C_2}(k_2) \leq F_{C_1}(k_1)$ $\quad F_{C_3}(k_3) = F_C(k)$

(a) (b) (c)

Figure 1: Examples of person model configurations, distributions, corresponding correction factors and minimum confidence required. (a) Example with root and head. (b) Example with root and 5 body parts. (c) Example with root and 8 body parts (original model with $N = 9$ (Felzenszwalb et al., 2010)).



(a) (b)

(c) (d)

Figure 2: Experimental sequences examples: (a) TUD-Crossing, (b) PETS2009-S2-L1, (c) PETS2009-S1-L1-2 and (d) PETS2009-S1-L2-1.

the used sequences.

For evaluating people detection performance based on ground-truth, we aim to compare the overall performance of different detection systems, so we have chosen the Precision-Recall (PR) evaluation method. In order to evaluate not only the (binary) yes/no detection but also the precise pedestrians locations and extents, we use also the three criteria, defined by (Leibe et al., 2005), that allow comparing hypotheses at different scales: the relative distance, cover, and overlap. Only one hypothesis per object is accepted as correct, so any additional hypothesis on the same object is considered as a false positive. In addition, we use the integrated Average Precision (AP) to summarize the overall performance, represented geometrically as the area under the PR curve (AUC-PR).

Figure 3 shows examples of results in terms of PR curves over the sequences TUD-Crossing, PETS2009-S2-L1, PETS2009-S1-L1-2 and PETS2009-S1-L2-1. Table 1 shows results in terms of AUC-PR. In both cases, the results show clearly how the performance decreases for every tested approach from the simplest sequences (TUD sequences) to the medium and high complexity sequences (PETS2009). The ACF detector provides the best results over the simplest scenarios, where there are few occlusions, but it provides worse results than the DTDP over the scenarios with more occlusions. The main reason for this behavior is that the ACF detector is based on a holistic person model and presents difficulties dealing with occlusions. The DTDP detector provides good results over the simplest sequences but worse than the ACF detector; however, the DTDP detector is based on a part-based person model and for this reason provides better results over the complex sequences with more occlusions. The ISM detector provides similar results than the DTDP detector over the simplest sequences but the worse results of all the three detectors over the complex sequences. In this case, the feature-part-based model is not robust enough to deal with partial occlusions.

Our proposal, the "DTDP multi configurations" provides better results than the DTDP detector in all the cases. It is clear that the improvement is more significant in those scenarios with higher complexity or occupation (PETS2009 sequences with high occupation, 12.7-16.7% improvement respect to the original DTDP detector) than in those scenarios with lower complexity (TUD and PETS sequences with low occupation, 2.4-5.3% improvement). This was expected, since the improvement possibilities on those sequences with more occlusion difficulties is higher.

Comparing our approach with the ACF detector, our proposal provides similar or slightly worse results on simple scenarios (TUD sequences with low occupation: -2.4 to 0% improvement with respect to the ACF detector), due to the lower performance of the base detector DTDP. However, over those scenarios with higher complexity or occupation, our approach provides significant improvements (PETS2009 sequences with high occupation, 11.8-31.9% improvement respect to the original DTDP detector).

The tests have been performed on aN AMD Opteron(tm) Processor 4386 with a 4xCPU frequency of 3 GHz and 4GB RAM. During our experiments, the original DTDP computational cost is around 3.6 seconds per frame with 640x480 images (TUD sequences) and around 4.9 seconds per frame with 768x576 images (PETS sequences). The proposed multi-configurations approach includes the same main core of the original approach and the additional computational cost of computing T configuration confidences (see equation 5) instead of only one (see equation 4). In the case of twenty different configurations ($T = 20$), the computational cost is around 4.9 seconds per frame with 640x480 images (TUD sequences) and around 6.2 seconds per frame with 768x576 images (PETS sequences). Assuming that we are able to run every configuration in parallel, the final computational cost will be established by the configuration with the maximum number of body parts, i.e., the original DTDP computational cost.

To sum up, the results show how the proposed combination of multiple configurations is more robust to partial occlusions than the original DTDP detector. In particular, our proposed detector provides better results than all other reference detectors from the state of the art (DTDP, ISM and ACF) over typical sequences with a high degree of occupation (groups of people together), where the presence of occlusions is typical.

5 CONCLUSIONS

People detection methods from the state of the art perform well in scenes with relatively few people, but are severely challenged by scenes with many subjects that partially occlude each other. We observe that typical occlusions are due to overlaps between people and propose a people detector tailored to various occlusion levels. We propose a generic multiple body parts combination framework in order to deal with these specific partial occlusions in crowded scenarios.

We have validated our approach and compared it with other state of the art approaches on several pub-

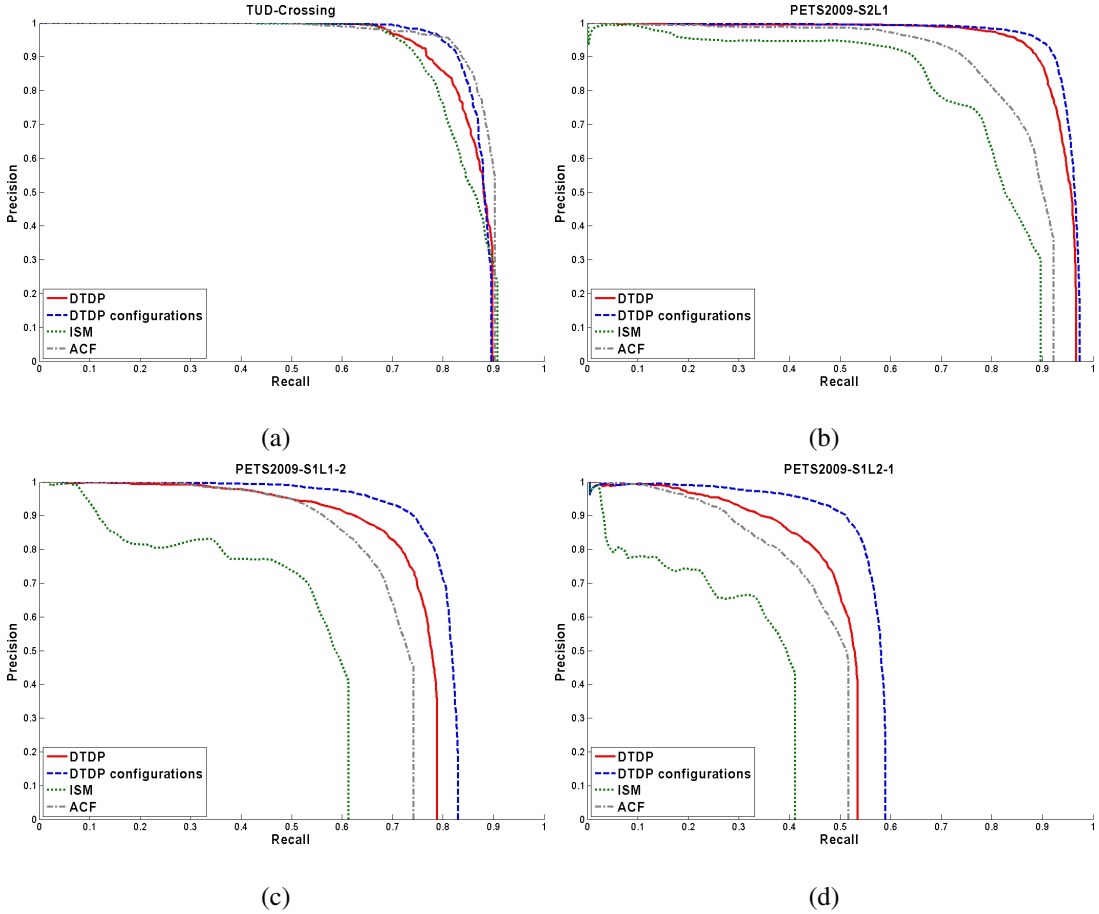


Figure 3: Examples of people detection performance in terms of Precision-Recall curves.

Table 1: Experimental results. Occupation in terms of number of pedestrians present simultaneously. Complexity classification. People detection performance in terms of area under the Precision-Recall curve (AUC-PR). Percentage increase ($\%\Delta^1$ and $\%\Delta^2$) calculated with respect to original performance DTDP and ACF respectively.

	Occupation Up to #	Complexity	AUR-PR					
			DTDP	ISM	ACF	**Ours**	$\%\Delta^1$	$\%\Delta^2$
TUD-Campus	8	Low	0.76	0.76	**0.80**	**0.80**	+5.3	0.0
TUD-Stadmitte	6	Low	0.79	0.71	**0.83**	0.81	+2.5	-2.4
TUD-Crossing	8	Low	0.85	0.84	**0.88**	0.87	+2.4	-1.1
PETS2009-S1L1-1	34	Medium	0.63	0.45	0.63	**0.67**	+6.3	+6.3
PETS2009-S1L1-2	26	Medium	0.73	0.49	0.68	**0.80**	+9.6	+17.6
PETS2009-S1L2-1	42	High	0.48	0.30	0.44	**0.56**	+16.7	+27.3
PETS2009-S1L2-2	40	High	0.50	0.36	0.51	**0.57**	+14.0	+11.8
PETS2009-S2L1	8	Low	0.93	0.78	0.85	**0.95**	+2.2	+11.8
PETS2009-S2L2	35	Medium	0.66	0.55	0.58	**0.75**	+13.6	+29.3
PETS2009-S2L3	42	High	0.55	0.34	0.47	**0.62**	+12.7	+31.9
PETS2009-S3L1	7	Low	0.93	0.82	0.94	**0.95**	+2.2	+1.1

lic datasets. The results, over sequences with different number of people present in the scene simultaneously, demonstrate the achieved improvements in typical crowded scenes where the number and range of possible different occlusions are much higher than in simpler scenarios.

As future work, we will try to extrapolate this scheme to other people detectors or even to any kind of object part-based approach. In addition, we propose to explore other methods in order to estimate both the probability density functions and the similarity between them.

ACKNOWLEDGMENTS

This work has been done while visiting the Communication Systems Group at the Technische Universität Berlin (Germany) under the supervision of Prof. Dr.-Ing. Thomas Sikora. This work has been partially supported by the Universidad Autónoma de Madrid ("Programa propio de ayudas para estancias breves en España y extranjero para Personal Docente e Investigador en Formación de la UAM"), by the Spanish Government (TEC2011-25995 EventVideo) and by the European Community's FP7 under grant agreement number 261776 (MOSAIC).

REFERENCES

Ali, I. and Dailey, M. N. (2012). Multiple human tracking in high-density crowds. *Image and Vision Computing*, 30(12):966 – 977.

Andriluka, M., Roth, S., and Schiele, B. (2008). People-tracking-by-detection and people-detection-by-tracking. In *Proc. of CVPR*, pages 1–8.

Andriluka, M., Roth, S., and Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection. In *Proc. of CVPR*, pages 623–630.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proc. of CVPR*, pages 886–893.

Dollár, P., Appel, R., and Kienzle, W. (2012a). Crosstalk cascades for frame-rate pedestrian detection. In *Proc. of ECCV*, number 645-659.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012b). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761.

Enzweiler, M. and Gavrila, D. M. (2009). Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195.

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

Garcia-Martin, A. and Martinez, J. M. (2012). On collaborative people detection and tracking in complex scenarios. *Image and Vision Computing*, 30(4):345–354.

Gerónimo, D., López, A. M., Sappa, A. D., and Graf, T. (2010). Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1239–1258.

Girshick, R. B., Felzenszwalb, P. F., and McAllester, D. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ rbg/latent-release4/.

Girshick, R. B., Felzenszwalb, P. F., and Mcallester, D. (2011). Object detection with grammar models. In *Proc. of NIPS*.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *Proc. of CVPR*, pages 878–885.

Milan, A., Roth, S., and Schindler, K. (2014). Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):58–72.

Patzold, M., Evangelio, R. H., and Sikora, T. (2010). Counting people in crowded environments by fusion of shape and motion information. In *Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS '10, pages 157–164, Washington, DC, USA. IEEE Computer Society.

PETS. International workshop on performance evaluation of tracking and surveillance, http://www.cvg.rdg.ac.uk/pets2009/index.html.

Rodriguez, M., Laptev, I., Sivic, J., and Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. In *Proc. of ICCV*, pages 2423–2430.

Seemann, E., Fritz, M., and Schiele, B. (2007). Towards robust pedestrian detection in crowded image sequences. In *Proc. of CVPR*, pages 1–8.

Tang, S., Andriluka, M., and Schiele, B. (2014). Detection and tracking of occluded people. *International Journal of Computer Vision*.

Zeng, C. and Ma, H. (2010). Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In *Proc. of ICPR*, pages 2069–2072.