PEOPLE-BACKGROUND SEGMENTATION WITH UNEQUAL ERROR COST

Álvaro García-Martín[1], Andrea Cavallaro[2] and José M. Martínez[1]

[1]Universidad Autónoma of Madrid (Spain) – [2] Queen Mary University of London (United Kingdom)

ABSTRACT

We address the problem of segmenting a video in two classes of different semantic value, namely background and people, with the goal of guaranteeing that no people (or body parts) are classified as background. Body parts classified as background are given a higher classification error cost (segmentation with bias on background), as opposed to traditional approaches focused on people detection. To generate the people-background segmentation mask, the proposed approach first combines detection confidence maps of body parts and then extends them in order to derive a background mask, which is finally post-processed using morphological operators. Experiments validate the performance of our algorithm in different complex indoor and outdoor scenes with both static and moving cameras.

Index Terms— People detection, detection confidence map, background confidence map, people-background segmentation.

1. INTRODUCTION

A two-class segmentation ensuring that no people or body parts are appearing in the background class is desirable for many computer vision applications, such as robotics and driver assistance systems. This type of segmentation is useful not only as a preprocessing step, but also for other video analysis processes such as tracking and people density estimation. While the focus of person detection approaches is to obtain a high detection performance and to reduce false positive detections, we aim at determining the areas without people in the scene by giving a higher penalty to pixels representing a person, but that have been incorrectly classified as background. This results in a segmentation mask with a bias on the background as opposed to a segmentation with bias on people.

People detection approaches can be classified into two groups, namely holistic and part-based detectors. As *holistic object model* representation, the Implicit Shape Model (ISM) consists of a codebook of local appearances (SIFT features) that are prototypical for the object category and a spatial probability distribution which specifies where each codebook entry may be found on the object [1]. Holistic detectors can use a sliding-window-based approach defined with Haar features and a cascade Adaboost classifier [2] or can be defined using locally normalized Histograms of Gradient (HOG) orientations descriptors [3]. A *part-based model* based on HOG extending [3] is presented in [4], whereas edge feature detectors trained on the whole human body and its parts (head, torso and legs) are used independently in [5]. In this case responses of each detector are combined to form a joint likelihood model that includes cases of multiple and possibly inter-occluded people. An extension of [1] to body parts detection that uses pictorial structures to represent parts configurations is presented in [6]. Another extension of [1] combines appearance and motion information, and introduces the Implicit Motion Model (IMM), a motion person model inspired by the ISM [7].

The person detection *confidence* map generated during the classification process at each point in an image can be used not only to recover missing detections [8], but also for estimating the global density of people in a scene [9] or for updating a tracker [10] and for extracting a set of potential tracks using the local temporal context of dense detection scores [11]. A cascaded filtering of the detector confidence map can be used to refine the detection and to reduce false positives [12]. The confidence map of an object can also be treated as image descriptor to learn history of confidences and improve the final detection [13].

In this work, we use people detection confidence maps with a different objective, namely to localize the areas in an image frame where *there are no people*. To this end, we present a people-background segmentation approach with unequal error cost between classes in order to ensure that no body parts are classified as background. The proposed method is based on [4] for detecting body parts, and extends this representation by appropriately grouping them. Then we fuse detection confidence maps according to regions that are expected to be covered by the body parts. The corresponding background segmentation mask is finally generated after binarization and post-processing.

The remainder of the paper is organized as follows: Section 2 describes the overall approach; Section 3 discusses the experimental results. Finally, Section 4 summarizes the main conclusions.

2. PEOPLE-BACKGROUND SEGMENTATION

Starting from the body-part representation introduced in [4], in this section we define five methods: an independent body parts approach, IBP; a dependent body parts approach, DBP; their extended versions, IEBP and DEBP, respectively; and the post-processed version of DEBP, which we will refer to as DEBP-P.

Let us consider the part-based multi-scale detector (Figure 1(a)), where $P_n\,(x, y, s)$ represents the confidence at pixel position (x, y) for body part n ($n = 1, ..., N$) associated to scale s ($s = 1, ..., S$). Let also each body part be modeled by a 3-tuple (F_n, $v_{n,0}$, d_n) [4], where F_n is the HOG (Histogram of Oriented Gradients) filter response (detection confidence) [3] 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 whole body; 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 confidence score for part n at scale s is given as

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

Work partially supported by the Universidad Autónoma de Madrid ("FPI-UAM") and by the Spanish Goverment ("TEC2011-25995 EventVideo"). This work was done while the first author was visting Queen Mary University of London.

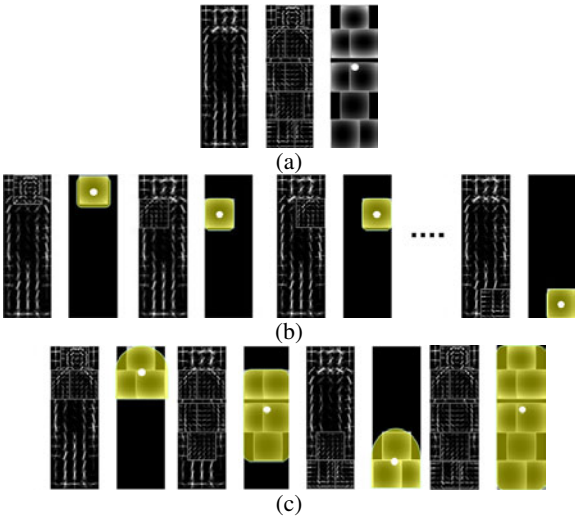


Fig. 1. Body parts representations. (a) Multi-part person model from [4]; (b) IBP model; (c) DBP model. The kernel used in the extensions is shown in yellow.

with

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

giving the displacement of part n relative to the anchor and

$$\phi\,(dx, dy) = \left(dx, dy, dx^2, dy^2\right) \tag{3}$$

defining the potential spatial deformation distributions [4].

We define IBP by using eight ($N = 8$) *independent* body parts I_n, with $n = 1, ..., N$ and specified the anchor position $v_{n,n}$ relative to the body part n instead of the root position (Figure 1(b)). To improve the detection robustness, we then define DBP using M *dependent* body part models D_m, with $m = 1, ..., M$ as combination of independent parts (Figure 1(c)). Each D_m is defined by L_m parts, $I_1, ..., I_{L_m}$, where I_{l_m} is one of the independent parts with its anchor position $v_{l,m}$ relative to the *corresponding* dependent body part D_m. In order to exploit the correlation between body parts, we have chosen $M = 4$ dependent body parts: head and shoulders, trunk, legs and full body. Moreover, in order to recover undetected dependent body parts or normalize the detection confidence between dependent body parts already detected, we propose to extend the dependent body parts definition and reuse the information from other dependent body parts. Each dependent body part D'_m is given by the maximum between the original dependent body part D_m and the average of the other dependent body parts, all of them relative to the same D_m.

If we assume that there are at least two visible dependent body parts for each person, we are able to recover or normalize body parts by averaging the remaining parts and, in turn, we avoid the reproduction of those isolated dependent body parts incorrectly detected:

$$D'_m\,(x, y, s) = max\left(D_m\,(x, y, s)\,, \frac{1}{M-1}\sum_{i\neq m}^{M} D_{i,m}\,(x, y, s)\right), \tag{4}$$

where $D_{i,m}\,(x, y, s)$ is the body part i with anchor position $v_{i,m}$.

Once we have obtained the dependent or independent body parts responses at each pixel position and scale, the confidence of each body part response is extended to define the methods IEBP and DEBP, respectively (Figure 2). IEBP extends each independent body part, whilst DEBP extends each independent body parts combination. Both IEBP and DEBP cover the detected part in the chosen body parts representation as represented by the kernel extensions (yellow shapes) in Figure 1(b) and (c) according to the area that it is expected to cover in a frame.

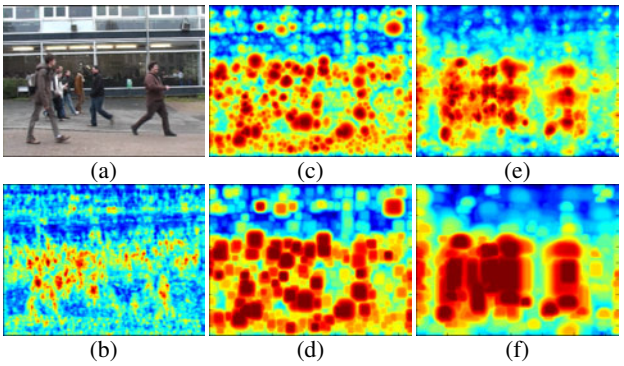


Fig. 2. Confidence maps for a sample frame (a) generated with: (b) the original method [4]; (c) IBP; (d) IEBP; (e) DBP; and (f) DEBP.

Once we have obtained all the final body part detection confidence maps $P_n\,(x, y, s)\,, \forall n = 1, ..., N$, we select for each position in the frame the maximum confidence level across scales and across parts to generate the fused confidence map $C\,(x, y)$:

$$C\,(x, y) = \max_{n=1,...,N}\max_{s=1,...,S} P_n\,(x, y, s)\,. \tag{5}$$

Figure 2 shows examples of confidence maps generated on the same frame using the original method [4], IBP, IEBP, DBP, and DEBP.

The final people-background mask is obtained by binarizing $C\,(x, y)$. Assuming that each person in the scene is visible (i.e. at least two dependent body parts are captured in the frame) or is partially occluded by another person, regions that are smaller than the minimum size of a person are eliminated. The minimum size is defined by the person model scale in [4]. The resulting mask undergoes an erosion with a disc the size of the smallest body part to detect in the minimum size of a person, followed by connected components analysis to remove regions that are smaller than the minimum size of a person. Finally a dilation operation with a disc the size of the smallest body part to detect in the maximum size of a person is performed to generate the final mask. We will refer to this overall method as DEBP-P. Sample results are shown in Figure 3.

3. EXPERIMENTAL RESULTS

In order to evaluate our unequal-error-cost people-background segmentation approach, we compare in this section the performance of the original algorithm [4], the independent and dependent body parts approaches (IBP and DBP, respectively), their extended versions (IEBP and DEBP, respectively), and the proposed method DEBP-P[1].

We use a set of publicly available sequences with different complexities, including occlusions, scale variations, different point of views and moving cameras: tree outdoors sequences (TUD-campus

[1]Video results, ground truth and additional data can be found at http://www-vpu.eps.uam.es/publications/PeopleBackgroundSegmentation

(a) (b) (c)

Fig. 3. Examples of results: (a) sample image; (b) DEBP result; (c) DEBP-P result.

Sequence	GTF	ANP	PPP
TUD-campus	7/71	6.1	14.13
TUD-crossing	21/201	6.2	9.55
TRECVID	6/103	9.1	9.38
PETS2006	6/1010	2.3	2.59
PETS2009	6/443	6.5	2.15
AVSS	6/907	2	3.68

Table 1. Description of the experimental dataset (Key. GTF: number of ground-truth frames per sequence; ANP: average number of people per ground-truth frame; PPP: percentage of pixels belonging to a person in the ground truth).

and TUD-crossing from [14] and PETS2009[2]), three indoor sequences (TRECVID2008[3], PETS2006[4] and AVSS2007[5]) and three sequences with moving cameras from [15].

In order to quantify the error, we manually generated a segmentation ground truth for selected frames of the first six sequences (see Table 1). Note that the image border (whose width is half the size of a person on both sides of the image, i.e. 20 or 40 pixels according to the model scale in [4]) is not considered in the quantitative evaluation. The visual results of these annotated first six sequences have been generated with the maximum binarization threshold for which there are no pixels of people misclassified as background, whilst the visual results of the three non-static camera sequences have been generated with the empirical binarization threshold of 0.8.

Table 2 shows the results in terms of AUC (area under the curve) with different false positives penalty factors: 1, 2, 4 and 10. A penalty factor of 1 corresponds to traditional segmentation approaches, whilst higher factors give higher penalties to segmentations with pixels that correspond to a person and are incorrectly classified as background, i.e. a penalty factor of 2 corresponds to a twice penalty and so on.

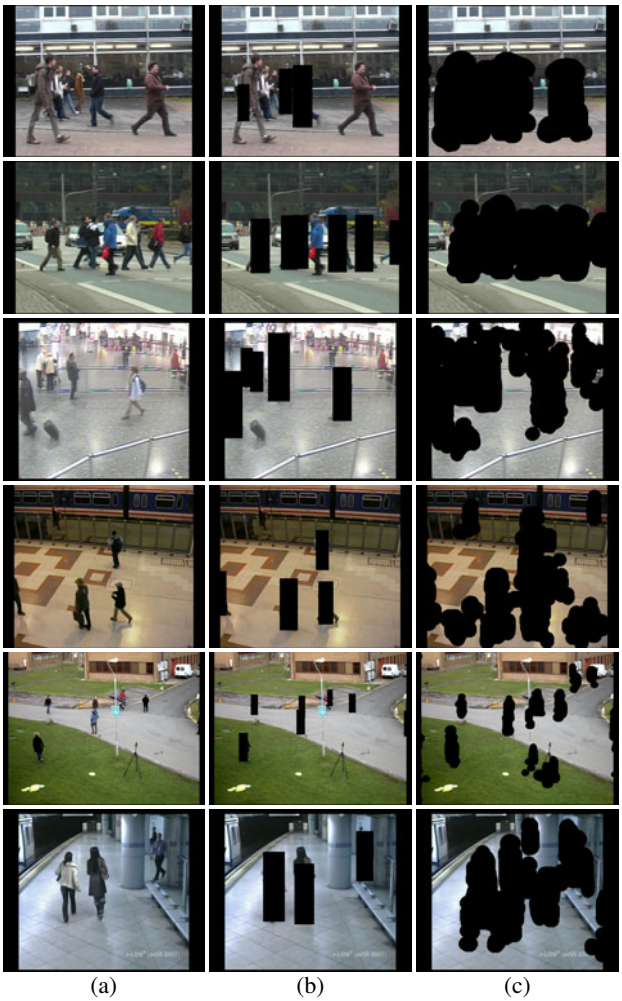
The results show that *dependent-part* approaches (DBP and DEBP) outperform *independent-part* approaches (IBP and IEBP), due to the greater robustness provided by the combined body parts detections. The extended versions (IEBP and DEBP) are significantly better than their non-extended counterparts (IBP and DBP), due to the reduction of the number of false positives (pixels that belong to a person incorrectly classified as background) without

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

[3]http://www.itl.nist.gov/iad/mig//tests/trecvid/2008/

[4]http://www.cvg.rdg.ac.uk/PETS2006/

[5]http://www.avss2007.org



(a) (b) (c)

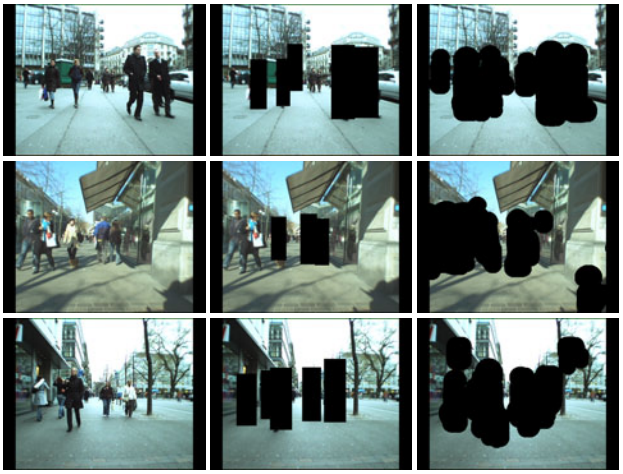
Fig. 4. Sample results: (a) original frame; (b) person detector result; and (c) DEBP-P result.

a substantial increase of false negatives (pixels that belong to the background incorrectly classified as people). Despite the fact that IBP and IEBP were initially designed to reduce false positives, the lack of dependency among parts generates many false negatives leading to worse performance compared to the corresponding original algorithm While the other approaches decrease drastically their performance with the increase of the penalty factor, the combination of dependent and extended body part approach DEBP, has the lowest decrease and the best system performance (0.98 ~ 0.74). Its post-processed version, the proposed approach DEBP-P, practically maintains the same performance and improves slightly the results for higher penalty factors (0.98 ~ 0.77).

Figure 4 and Figure 5 show examples of static and non-static camera scenarios, respectively. Figure 4 shows the performance of the original algorithm in terms of detection: we can see examples of missing detections or false detections (people only partially detected) in each scenario. The best results (0.98 ~ 0.95) are obtained in the sequence PETS2009, due to the person model [4]. Although the person model supports different body parts configurations (deformable part model), it favors people with arms and legs close to the body. In the case of the PETS2009 sequence, people are better

AUC	TUD-Campus				TUD-Crossing				TRECVID				PETS2006				PETS2009				AVSS			
False positive penalty factor	1	2	4	10	1	2	4	10	1	2	4	10	1	2	4	10	1	2	4	10	1	2	4	10
Original	.83	.75	.65	.51	.87	.81	.73	.63	.83	.74	.65	.50	.77	.68	.59	.46	.88	.82	.75	.64	.89	.83	.75	.63
IBP	.81	.74	.66	.56	.78	.68	.58	.44	.65	.53	.41	.27	.68	.56	.45	.31	.70	.58	.44	.28	.69	.57	.46	.32
IEBP	.84	.79	.72	.63	.84	.76	.68	.55	.74	.63	.52	.37	.74	.63	.52	.37	.80	.70	.58	.42	.77	.67	.56	.41
DBP	.93	.90	.85	.79	.93	.89	.84	.76	.85	.77	.68	.54	.85	.78	.70	.58	.86	.77	.66	.50	.90	.85	.78	.67
DEBP	**.95**	**.93**	**.90**	**.85**	**.95**	**.93**	**.90**	**.85**	**.92**	**.88**	.83	.74	.93	.89	.85	.77	**.98**	.96	.93	.87	.95	.93	.90	.83
DEBP-P	.93	.91	.88	.84	.94	.92	**.90**	**.85**	**.92**	**.88**	**.84**	**.77**	**.94**	**.91**	**.87**	**.80**	**.98**	**.98**	**.96**	**.95**	**.96**	**.95**	**.93**	**.90**

Table 2. Area under the curve (AUC) with different false positive penalty factors.



(a) (b) (c)

Fig. 5. Sample results for moving cameras: (a) original frame; (b) person detector result; and (c) DEBP-P background mask.

suited to the model due to the far field view. However, in the other scenarios, the person model must be adapted to larger pose variations (higher body part deformation costs), getting worse results. The other factors that have influenced the results are the presence of shadows and reflections in TRECVID, PETS2006 and AVSS that makes the detection more difficult; and the greater scales variation in TRECVID and PETS2006 that makes the confidence map more complex and introduces more false body part detections that worsen the results. A separate analysis for each scale, as opposed to the current approach of combining first all the scales and then performing segmentation, could improve the results.

4. CONCLUSIONS

We presented a people-background segmentation approach that aims to ensure that there are no people or body parts assigned to the background class at the cost of potentially increasing the number of background pixels classified as people. This type of biased segmentation is useful not only as preprocessing for applications involving people detection but also for other video analysis processes such as tracking and people density estimation. The experimental results show the performance of our proposal in various complex scenarios and independently of camera motion.

As future work, we will incorporate temporal information in the model and explore the possibility of detecting automatically the range of scales presented in each part of the scene and the binarization threshold. Finally, we will extend the proposed method to other person detector approaches and object classes.

5. REFERENCES

[1] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, 77(1), 259-289, 2008.

[2] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, 57(2), 137-154, 2004.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[4] 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.

[5] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, 75(2), 247-266, 2007.

[6] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*, 2009.

[7] A. Garcia-Martin, A. Hauptmann, and J. M. Martinez, "People detection based on appearance and motion models," in *AVSS*, 2011.

[8] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multi-person tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9), 1820-1833, 2010.

[9] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," in *ICCV*, 2011.

[10] S. Avidan, "Ensemble tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2), 261-271, 2007.

[11] J. Yu, D. Farin, and B. Schiele, "Multi-target tracking in crowded scenes," *Pattern Recognition, LNCS*, 6835, 406-415, 2011.

[12] S. Stalder, H. Grabner, and L. V. Gool, "Cascaded confidence filtering for improved tracking-by-detection," in *ECCV*, 2010.

[13] L. Middleton and J. R. Snowdon, "Histogram of confidences for person detection," in *ICIP*, 2010.

[14] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *CVPR*, 2008.

[15] A. Ess, B. Leibe, K. Schindler, and L. V. Gool, "A mobile vision system for robust multi-person tracking," in *CVPR*, 2008.