A Graph Based People Silhouette Segmentation Using Combined Probabilities Extracted from Appearance, Shape Template Prior, and Color Distributions

Christophe Coniglio[1,2(✉)], Cyril Meurie[1,2], Olivier Lézoray[3], and Marion Berbineau[1,2]

[1] Univ Lille Nord de France, 59000 Lille, France
coniglio.christophe@ifsttar.fr
[2] IFSTTAR, COSYS, LEOST, 59650 Villeneuve d'Ascq, France
[3] Normandie Univ., UNICAEN, ENSICAEN, GREYC UMR CNRS 6072, Caen, France

Abstract. In this paper, we present an approach for the segmentation of people silhouettes in images. Since in real-world images estimating pixel probabilities to belong to people or background is difficult, we propose to optimally combine several ones. A local window classifier based on SVMs with Histograms of Oriented Gradients features estimates probabilities from pixels' appearance. A shape template prior is also computed over a set of training images. From these two probability maps, color distributions relying on color histograms and Gaussian Mixture Models are estimated and the associated probability maps are derived. All these probability maps are optimally combined into a single one with weighting coefficients determined by a genetic algorithm. This final probability map is used within a graph-cut to extract accurately the silhouette. Experimental results are provided on both the INRIA Static Person Dataset and BOSS European project and show the benefit of the approach.

1 Introduction

In video surveillance applications, the extraction of people silhouette is a well-known bottleneck to efficiently perform tasks such as people recognition [1,2]. If such applications have been very popular in the computer vision community, nowadays other application fields also need such a people silhouette segmentation. A recent example comes with clothes segmentation [3,4] that has received much interest from the fashion industry. Indeed, commercially, it can be used in online cloth retail portals where people can search various clothes from image examples. In a more general way, since people silhouettes describe human appearance, it is an important step towards understanding the human-based content in images. Anyway, to well operate, all these applications need precise people silhouette extraction. Indeed, any object not belonging to people but incorporated in the extracted silhouette can strongly degrade the performance of the intended applications.

S. Battiato et al. (Eds.): ACIVS 2015, LNCS 9386, pp. 1–12, 2015.
DOI: 10.1007/978-3-319-25903-1_26

Traditionally, people silhouette segmentation is performed in videos with standard motion-based background subtraction strategies [5]. With static images, detection-oriented methods have emerged and make extensive use of machine learning together with efficient appearance representations (such as Histograms of Oriented Gradients -HOG) [6,7]. The result of such a detection is a bounding-box (BB) with the detected people inside. However, this does not directly provide the silhouette segmentation that has to be further estimated inside the obtained bounding box. Such an extraction of people silhouette is therefore less easy than with videos [8]. In [9] the authors propose a global approach that consists in analyzing links between structural elements in images formed by regions of pixels. A process based on color and texture features allowing to link these structural elements is also proposed. In [10,11] the authors propose to use several characteristics of bounding boxes (such as the fact that people are centered in BB images). Thus, it is easier to define a template driven to delimit the position of people. In [11] the authors propose to extend the previous method with part-based templates and extract the person silhouette with combined graph-cuts. In [10], a local window SVM classifier from HOG features is combined with edge detection to enhance an iterative graph-cut segmentation. To conclude, in the context of video-surveillance, one can notice that very few studies deal with people silhouette segmentation but solely focus on the re-identification step. This is probably explained by the fact that realizing a ground truth silhouette segmentation dataset is tedious and costly in time. Thus, few ground truths are available in the literature, that is why, it seems important to point out that we have achieved a consequent ground truth of people silhouettes for the BOSS database [12]. In this paper, we propose an approach towards the automatic extraction of people silhouette from bounding box images with a graph based segmentation using combined probabilities extracted from appearance, shape template prior, and color distributions all optimized with a genetic algorithm (GA).

2 Proposed Method

Our method is designed to extract precisely people silhouettes in images in the form of a bounding box and obtained using basic techniques of person detection (such as Histograms of Oriented Gradients -HOG) [6]. To perform the segmentation, we need to estimate a probability map that provides the class memberships of each pixel for the two classes people/background to discriminate. This probability map is used to initialize the capacities in a graph-cut segmentation. As we already mentioned it, estimating an accurate probability map is difficult. Therefore, on the one hand, our contribution consists in developing a new method to estimate probability maps based on local small windows probability appearance (as illustrated in Figure 2). Each local small window is described by its own learned classification characterized by optimized HOG features. On the other hand, we propose to combine our probability maps with those obtained by the following standard techniques: template shape prior probabilities that are estimated from a learning database in the form of a mean shape and two color

distribution probabilities obtained by color histograms and Gaussian Mixture Models (initialized by our probability map and the probability obtained by the template shape prior). The resulting six probability maps are combined altogether with weighting factors optimally determined with a genetic algorithm (GA). Figure 1 sums up the whole approach. Contrary to the works presented in [11], our method does not only use one information (part based templates) but the previous six cues to initialize the graph cut segmentation. In opposition to [10] , we prefer to detect a probability to belong to people or background, instead of detecting possible edges of people in local small windows.

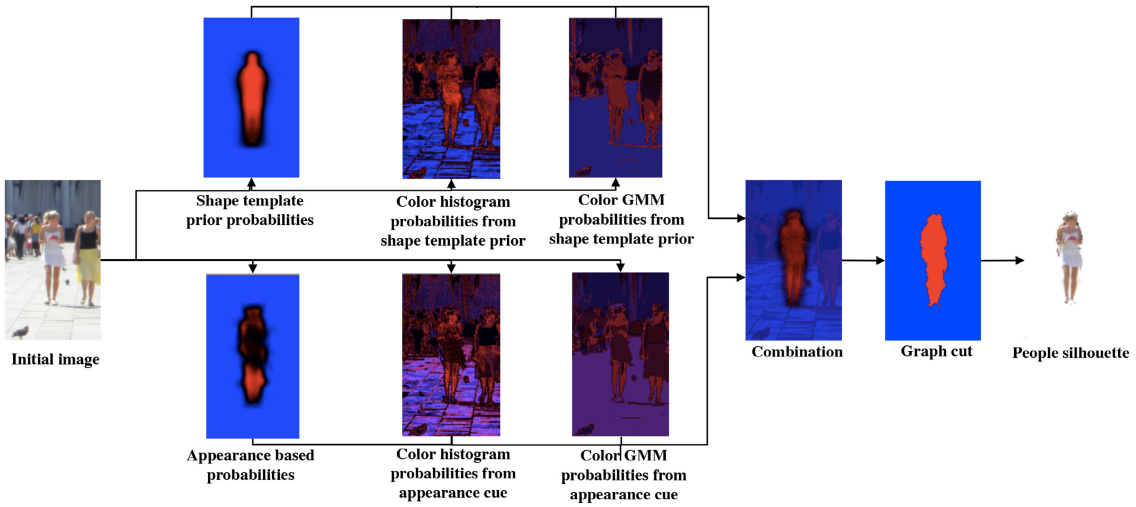

Fig. 1. Synopsis of the proposed approach.

2.1 Appearance Based Probabilities

In [6], the authors have proposed to classify local windows inside a bounding-box into two classes (people or not people) using HOG descriptor combined with a single SVM classifier. We will use a similar but different method and we will classify local windows inside the bounding-box using several SVM classifiers (one per considered local window). With such an approach, the classification process is not anymore performed globally but at the local window level in order to classify any piece of people as being or not a piece of the persons silhouette (see Figure 3). Given the estimations from the local windows' classifications, we can obtain an appearance based probability map by averaging the predictions of overlapping local windows for each image pixel. In order to have the finest probability results per pixel, we consider all the possible local windows with an offset of one pixel. To build the HOG descriptor, each local window is divided into blocks and each block is divided into cells (see Figure 2). HOG features are extracted from cells and a HOG descriptor is the vector of the cells' HOGs features. To obtain better HOG descriptors, we consider an overlapping between cells, so that each cell contributes to several blocks' descriptors. A SVM classifier classifies each HOG descriptor extracted from local windows, therefore it is necessary to have an individual training for each local window. During these training

steps, a particular attention is necessary for the background label. Indeed in the case where the person to extract is surrounded by other people (for example in a crowed environment), it is preferable to ignore this example during the training since this can lower the generalization capabilities of the SVM classifier. Given all the obtained classifications from the overlapping local windows, an average of all pixel class memberships is performed to determine the final labeling. In this paper we will show such results with colors ranging from blue (background class) to red (people class). In order to reduce the changes of HOG descriptors due to brightness changes between images, a pre-processing step normalizes the HOG descriptor per block. Finally, since we do not have any preconceived

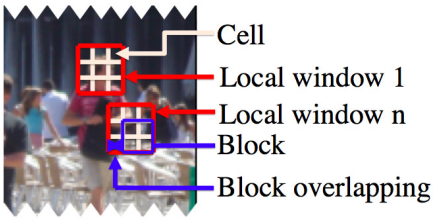

Fig. 2. Example of cut into local windows, cells, blocks, and block overlapping.

idea on the ideal setting of both HOG descriptor and SVM classifier, it is difficult to predict the best configuration setting. First, the HOG descriptor is described by sizes of local windows, blocks, and cells, by the overlap between blocks and by the number of bins of the HOG feature and the type of normalization (L_1 or L_2 Norms). Second, a SVM classifier can use different kernels (linear, polynomial, Gaussian) that have several parameters. We could arbitrarily fix all these parameters but this does not ensure that we will have the best possible configuration. In addition, if we use large HOG descriptors this amounts to use larger local windows and because of the final averaging, the classification is much less efficient on edges. A compromise has to be found. To optimally determine the latter, we have determined the best parameters by the use of a genetic algorithm. For the SVM, each training is performed with cross-validation and grid-search to determine the best parameters. To compare the different possible configurations of the HOG descriptor, we perform a segmentation with graph cuts (see section 2.5 for details) from the estimated probability map to obtain a final segmentation in two classes (people or background). Figure 3 shows the approach. The obtained segmentations are compared with the $F_{measure}$ score and are shown in Table 1. Table 1 resumes the five best settings for the local windows HOG based classifier. The best size of cell turns to be of 5 pixels with 9-bins histograms. Block overlapping is not of interest and mid-size local windows and blocks are sufficient. This is in concordance with the remark we made before on the blurring effect obtained with too large HOG descriptors that are averaged. We have retained finally the best parameter configuration shown in Table 1. An example of result is shown in Figure 4.

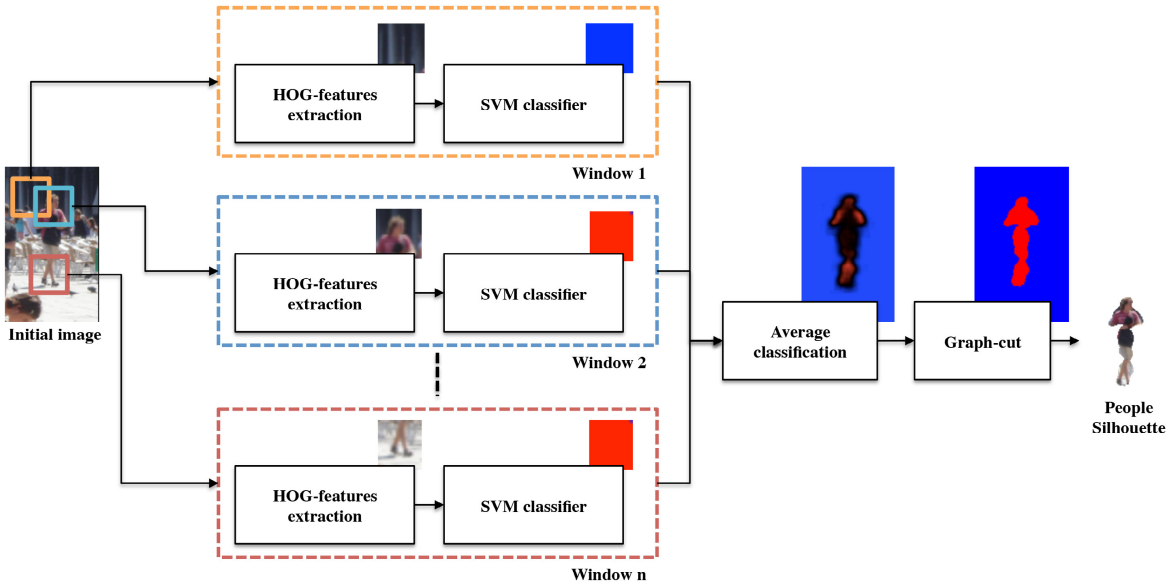

Fig. 3. Synopsis of the strategy employed to evaluate the best parameter settings of local windows HOG based classifier.

Table 1. The five best parameter settings for the estimation of appearance based probabilities with local windows classifier. Training realized on training set and scores obtained on the test set of the *INRIA Static Person Dataset*.

SVM kernel	local-window size	block size	block overlap size	Cell size	HOG #Bins	Type of Normalization	Recall	Precision	$F_{measure}$
RBF	8	8	Ø	4	9	Ø	0.765	0.865	0.806
RBF	10	10	Ø	5	9	L_2 Norm	0.76	0.865	0.804
RBF	5	Ø	Ø	5	9	L_2 Norm	0.757	0.862	0.802
RBF	15	10	5	5	9	L_2 Norm	0.792	0.823	0.799
RBF	12	12	Ø	6	9	L_2 Norm	0.756	0.858	0.798

2.2 Shape Template Prior Based Probabilities

The use of shape template priors is common in literature [13,14]. Indeed, we can notice that images of people, in the form of bounding-boxes, are generally centered on the person. This comes from the fact that bounding-boxes are mostly results of a people detection process based on machine learning trained with positives images for people in the center of the image. We propose to use a shape template prior based probability template in contrast to a binary shape template. In the case of binary shape template, one applies directly the template on the image as a mask. The use of a probability shape template is more appropriate for our choice of segmentation method using graph cuts that needs such a membership information. Our shape template prior based probabilities is obtained from an averaging of all the ground-truth shapes in the training set of the *INRIA Static Person Dataset*, see Figure 4. This prior is therefore computed only once.

2.3 Color Distributions Based Probabilities

As it can be seen in Figure 4, the obtained appearance and shape prior probability maps do not provide very fine results. To enhance these results, we propose to rely on color distributions given the initial results from appearance and shape prior probability maps.These two latter probability maps are used to make two lists of pixels that describe two classes (people and background). To be added to a list, a pixel needs to have a probability of belonging to a class upper than a thresholding defined by the final genetic optimization. For each given list, we estimate their color distributions using both color histograms (one per color channel with 256 bins) or Gaussian Mixture Models (using 5 Gaussians to describe each class). Given these two distributions for both appearance and shape prior probability maps, we estimate the class memberships using the color distributions. This provides four additional probability maps: two with color distributions estimated from color histograms for both appearance and shape prior probability maps and two with GMMs estimated for both appearance and shape prior probability maps. Figure 4 shows an example of the obtained color distributions based probabilities. One can see that the obtained results are much more fine than with the appearance and shape prior probability maps. However since only color is used, the colors that describe a person can also be found in the background and identified as people. This shows that not a single probability map is sufficient and we have to make the most of all of them.

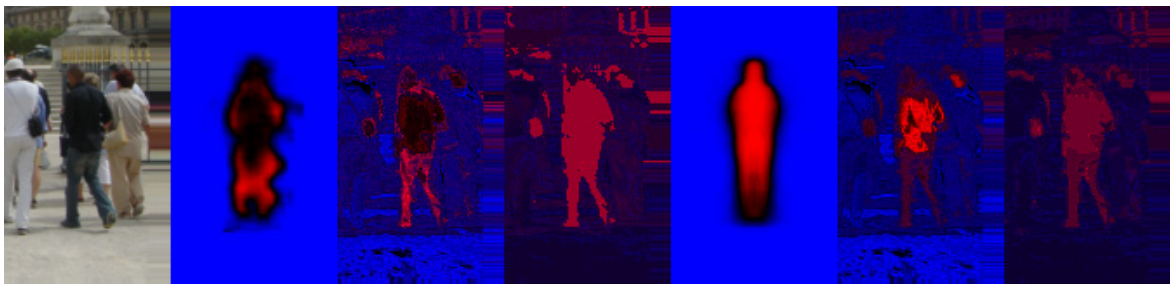

Fig. 4. Example of probability map estimation. From left to right: initial image, appearance based probabilities, color histogram based probabilities from appearance based probabilities, Color GMM based probabilities from appearance based probabilities, shape template prior based probabilities, color histogram based probabilities from shape template prior based probabilities and GMM based probabilities from shape template prior based probabilities.

2.4 Combination of the Probability Maps

Thanks to previous steps, we have six probability maps about the position of people (appearance and shape prior probability maps and four derived color distributions based probabilities). In this step, we propose to combine these six informations into a single one. We affect a coefficient $C(k)$ for each probability map and we calculate the final pixels probabilities of the two classes (people and

background) by:

$$P^{class}(p_i) = \sum_{k=1}^{6} C(k) \cdot P_k^{class}(p_i) \text{ with } \sum_{k=1}^{6} C(k) = 1 \quad (1)$$

where $P_k^{class}(p_i)$ denotes the conditional estimated probability from the k-th map for a given *class* among background and people. As our method needs a training step to define the shape template and train all the SVM classifiers with local windows, we also use this training step to set the weighting coefficients on the same training set. To ensure that the coefficients we obtain are optimally chosen, we use a genetic algorithm. With such an algorithm, we represent the coefficient setting problem as an optimization problem. The genetic algorithm is based on three steps (crossover, mutation and selection) that execute in loops with random initialization for the start. These steps are performed on a population of chromosomes. A chromosome correspond to the coding of the six coefficients $C(k)$ of our proposed method. We have used the $F_{measure}$ score as a fitness measure to evaluate each chromosome in the selection step. Our method is fast, the chromosome being relatively small, and the optimization loop quickly converges with a population of almost 100 chromosomes. Results of best selected chromosomes will be presented in Section 3.

2.5 Graph Cut Segmentation

The final step consists to classify into two classes (people and background) the image given the estimation of probabilities obtained from the combination of probability maps. To do so, we use graph cuts [15]. Graph-cut techniques are among the most powerful methods that extract foreground from background. Graph-cuts enable object segmentation with the optimization of a discrete energy function defined on a binary label set $L = \{0, 1\}$ by computing a minimum cut on the graph associated to the image. The key task is the proper definition of this energy in order to capture the properties of object regions and those of boundaries between them. We consider a graph $G = (V, E)$ witch is composed of $|V|$ nodes (the pixels of the image), where each node p_i is assigned a label $l_i \in L$ and $|E|$ edges (inferred from 8-connectivity between pixels). To classify each node of the graph into two classes, we consider the following energy:

$$\hat{\imath} = \underset{l \in F}{\operatorname{argmin}} \left(\sum_{p_i \in V} W^{l_i}(p_i) + t \sum_{p_i \in V} \sum_{p_j \in N_{p_i}} S(p_i, p_j) \cdot \delta_{l_i \neq l_j} \right) \quad (2)$$

The best segmentation (clustering into the two classes people and background) corresponds to the minimum of the energy $\hat{\imath}$, in the set F of all possible labeling solutions. The first term of the energy is defined as $W^{l_i}(p_i) = -\log(P^{l_i}(p_i))$. It uses the probabilities of each pixel to belong to the l_i class (people or background), and is obtained from the weighted combination of six different probability maps. The second term is obtained from the product of two terms.

The term $\delta_{l_i \neq l_j}$ is the Potts prior that encourages piecewise-constant labeling, and N_{p_i} is the set of 8-connected neighbors of p_i in the grid graph associated to the image. The term $S(p_i, p_j)$ expresses a similarity measure between both pixels p_i and p_j and is given by:

$$S(p_i, p_j) = exp\left(-\frac{d(p_i, p_j)}{2\theta^2}\right) \cdot \frac{1}{dist(p_i, p_j)} \text{ and } d(p_i, p_j) = \sqrt{\sum_{c=1}^{3} (p_i^c - p_j^c)^2} \quad (3)$$

where the $dist(pi, pj)$ is the Euclidean distance between the pixels p_i and p_j. The coefficient θ is a parameter genetically optimized to adjusting the sensitivity to intensity difference between neighboring pixels and $d(p_i, p_j)$ denotes the sum of distance between the color channels p_i^c and p_j^c associated to pixels p_i and p_j. The optimization is done with the min-cut/max-flow implementation of [16]. The result of graph cut labeling is a binary image where each pixel has been assigned to one class among background and people. Therefore, we obtain the final people silhouette.

3 Experimental Results

Our method has been evaluated on two databases. Each database was separated into two sets of images with 2/3 of positive image for the training set and 1/3 positive image for the test set. The first database is composed of 390 positive images (bounding-boxes of 96×160 pixels with a people in the center with a wide range of different background such as crowded environments) and 150 negative images (bounding-boxes of 96×160 pixels without people inside) from the INRIA Static Person Dataset [17] (this database contains only static images). We have used the people ground-truth segmentation provided by [13] to define the reference people segmentations. The second database is from the BOSS European project database [12] and contains video sequences. We have selected a video sequence that contains several difficulties due to real-world transportation systems such as strong brightness changes or many shadows. On the chosen sequence, there are twelve persons that move in front of the camera inside a train. We have used the Dalal's algorithm [6] (with parameters tuned by a genetic algorithm) to extract people bounding-boxes from the video sequence. Our database contain 453 positive images (bounding-boxes of 96×160 pixels) and 200 negative images (bounding-boxes of 96×160 pixels without people, taken in the same video sequence).

The proposed method contains three separate training (shape template prior, local windows classifier and genetic algorithm for probabilities combination). We begin by training separately on the training set, the shape template and the local windows classifier. For the local windows classifier, we have already defined the best parameters settings for the HOG descriptor, and the training step consists only to train each SVM classifier on local windows. Then we use the genetic algorithm on the obtained (and derived from color distributions) probability maps with the training set.

Table 2 shows parameters and weighting coefficients defined by the genetic algorithm. We can see several differences on the results between the two databases (INRIA and BOSS). In case of INRIA database, probability maps are close altogether. Concerning the BOSS database, the most important probability maps are those obtained from HOG local windows and the two GMM. Then the derived color histogram probabilities are not as much essential. These differences are certainly due to the facility to detect the background in the BOSS database.

Table 2. Parameters and weighting coefficients given by genetic algorithm for combination step

	Database	
	INRIA	BOSS
Appearance based probabilities	17.6%	27.5%
Color histograms probabilities from Appearance based probabilities	10.5%	2.3%
GMM probabilities from Appearance based probabilities	20.7%	24.8%
Shape template prior Probability	24.7%	14.3%
Color histograms probabilities from Shape template prior Probability	15.3%	3.5 %
Color GMM probabilities from Shape template prior Probability	11.2%	27.5%
Thresholding for color distribution (% of belonging probabilities)	18%	25%
Coefficient θ in graph-cut segmentation	56.6	22.6

The processing time obtained with a non-optimized C++ program is a full process of 180ms by image on a 1.8 GHz Intel Core i5. Table 3 shows the $F_{measure}$ results with our method compared to recent literature [10,11,18]. The fitness measure used to perform the genetic optimization is the $F_{measure}$ score:

$$F_{measure} = 2 \; . \; \frac{\text{precision} \; . \; \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ and precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Where TP is true positive pixels, FN is false negative pixels and FP is false positive pixels of the initial image.

First of all, on INRIA database, our method obtains a good score ($F_{measure}$ of 0.860) upper than [10] ($F_{measure}$ of 0.841) which uses only one appearance based feature. This confirms our interest to combine different cues to estimate the position of people and initialize the graph-cut. The method proposed in [11] uses a part-based template combined with a graph cut results for each part. This method allows to increase the precision of the people template. Even if our proposed method obtains a close score ($F_{measure}$ of 0.860) compared to [11] ($F_{measure}$ of 0.885), our future works will consist to use a similar technique to improve our template-driven method and further increase our results. We also think it is important to mention that there is hardly very few works on segmentation of people silhouette with quantitative results on static images. In [19], we

Fig. 5. Four first rows present results on INRIA Static Dataset and last rows on BOSS Dataset. Each row presents three silhouette extraction results. Each result shows the initial image (left), the combined probability map (people in red and background in blue) in the middle and the segmentation obtained with the proposed method (right).

have proposed another approach exploiting the video information, for which the $F_{measure}$ was 0.89. Our proposed approach performs better without any temporal information use. The gap between the $F_{measure}$ score on the INRIA and the BOSS databases can be explained by the uniformity and the complexity of the background. Figure 5 shows segmentation results on the two databases. We see that people silhouettes are well segmented even in crowded environments. Nevertheless, the border delineation between people and background is not always perfect and there is still some room for improvement.

Table 3. Segmentation results ($F_{measure}$) on INRIA and BOSS databases

	Database	
	INRIA	BOSS
MIGNIOT 2011[11]	0.885	-
MIGNIOT 2013[10]	0.841	-
SHARMA 2007[18]	0.820	-
Proposed method	**0.860**	**0.911**

4 Conclusion

In this paper, we have proposed to use a graph based segmentation to extract the silhouette of people from bounding boxes. To do so we have proposed to combine several probability maps estimated from cues relying on appearance (with HOG descriptors extracted on local windows classified by different SVM classifiers), on a shape template prior and on derived probabilities from color distribution (color histograms or GMMs). The weighting coefficients as well as the parameters of the appearance local window classifiers have been optimally determined with a genetic algorithm. Experimental results have shown good results comparable to the actual state-of-the-art on the standard INRIA Static person dataset, and very good results on the real-world BOSS dataset. Future works will aim at enhancing the people silhouette border delineation by improving our template-driven method and integrating a superpixel segmentation into the graph-cut clustering.

References

1. Mori, G., Ren, X., Efros, A.A., Malik, J.: Recovering human body configurations: combining segmentation and recognition. In: Computer Society Conference on Computer Vision and Pattern Recognition, ser. CVPR 2004, pp. 326–333 (2004)
2. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2010
3. Yang, M., Yu, K.: Real-time clothing recognition in surveillance videos. In: ICIP, pp. 2937–2940. IEEE (2011)

4. Gallagher, A., Chen, T.: Clothing cosegmentation for recognizing people. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2008
5. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models forreal-time tracking. In: Computer Vision and Pattern Recognition, vol. 2, pp. 2246–2252 (1999)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: International Conference on Computer Vision & Pattern Recognition, vol. 2, pp. 886–893 (2005)
7. Zhu, Q., Avidan, S., Yeh, M.-C., Cheng, K.-T.: Fast human detection using a cascade of histograms of oriented gradients. In: CVPR 2006, pp. 1491–1498 (2006)
8. Gong, S., Cristanio, M., Yan, S., Loy, C.: Person re-identification. Springer (2014)
9. Jojic, N., Perina, A., Cristani, M., Murino, V., Frey, B.: Stel component analysis: modeling spatial correlations in image class structure. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2044–2051 (2009)
10. Migniot, C., Bertolino, P., Chassery, J.-M.: Iterative human segmentation from detection windows using contour segment analysis. In: VISAPP 2013 - Proceedings of the International Conference on Computer Vision Theory and Applications, pp. 405–412 (2013)
11. Migniot, C., Bertolino, P., Chassery, J.-M.: Automatic people segmentation with a template-driven graph cut. In: International Conference on Image Processing (2011)
12. Boss europeen project (on bord wireless secured video surveillance). http://www.multitel.be/image/research-development/research-projects/boss.php
13. Migniot, C., Bertolino, P., Chassery, J.-M.: Contour segment analysis for human silhouette pre-segmentation. In: 5th International Conference on Computer Vision Theory and Applications (VISAPP 2010), Angers, France, May 2010
14. Lin, Z., Davis, L.S.: Shape-based human detection and segmentation via hierarchical part-template matching. IEEE Trans. Pattern Anal. Mach. Intell., pp. 604–618 (2010)
15. Boykov, Y., Jolly, M.: Interactive graph cuts for optimal boundary region segmentation of objects in n-d images. In: IEEE International Conference on Computer Vision, vol. 1, pp. 105–112 (2001)
16. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9), 1124–1137 (2004)
17. Inria person dataset. http://pascal.inrialpes.fr/data/human/
18. Sharma, V., Davis, J.: Integrating appearance and motion cues for simultaneous detection and segmentation of pedestrians. In: Computer Vision ICCV IEEE, pp. 1–8. IEEE (2007)
19. Coniglio, C., Meurie, C., Lzoray, O., Berbineau, M.: A genetically optimized graph-based people extraction method for embedded transportation systems real conditions. In: 17th International Conference on Intelligent Transportation Systems, pp. 1589–1595 (2014)