Pattern Recognition Letters People silhouette extraction from people detection bounding boxes in images

Christophe Coniglio, Cyril Meurie, Olivier Lézoray, Marion Berbineau

Pattern Recognition Letters
journal homepage: www.elsevier.com

People silhouette extraction from people detection bounding boxes in images

Christophe Coniglio[a], Cyril Meurie[a], Olivier Lézoray[b,**], Marion Berbineau[a]

[a]*Univ Lille Nord de France, F-59000 Lille, IFSTTAR, COSYS, LEOST, F-59650, Villeneuve d'Ascq*
[b]*Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France*

ABSTRACT

In many applications such as video surveillance or autonomous vehicles, people detection is a key element, often based on feature extraction and combined with supervised classification. Usually, output of these methods is in the form of a bounding-box containing an extracted people along with the background. But in specific application contexts, this bounding box information is not sufficient and a precise segmentation of people silhouette is needed inside the bounding box. For videos, this is actually solved by using background subtraction strategies. However, this cannot be considered for the case of still images that also occur in many video surveillance applications. To that aim, we propose to consider that issue in this paper. The principle is to devise a complete scheme for people segmentation inside people detection bounding boxes. Such a scheme relies on several steps: pre-processing, feature extraction and probability map computation to approximately locate people silhouette, and graph cut clustering to refine the silhouette from the map prior. Since many different methods can be considered, along with their associated parameters, tuning, we use a systematic approach towards determining the best combination scheme to conceive a segmentation scheme. The F-measure is used as a benchmark for evaluation. Experimental results show the benefit of the proposed approach that goes beyond the actual state-of-the-art.

1. Introduction

People detection is a key element of any intelligent video surveillance system since it provides the localization of people in images, mandatory for people tracking and recognition that is the basis of safety systems. In the computer vision literature, people detection is known as a difficult problem due to multiple possible combinations that can occur among people (with variations of angle, pose, *etc.*) and their clothes (habits, outfits, *etc.*). With these challenging issues, people detection has been a very active research area in computer vision in recent years (Gong et al., 2014). Numerous approaches have been proposed. A first major scientific leap has been made with the use of handcrafted multidimensional features (Haar (Mohan et al., 2001), HOG (Dalal and Triggs, 2005a; Zhu et al., 2006)) that were provided to machine learning methods such as Support Vector Machines (Hearst et al., 1998) or boosting (Siala et al., 2009), to cite a few. These methods have been recently superseded by deep learning methods, such as Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012), that automatically learn features and classify them with nonlinear multi-Layer Perceptrons. The success of these methods comes from their excellent ability to learn features from large datasets and this has contributed to their recent massive usage in people detection tasks, see (Redmon and Angelova, 2015) and references therein.

People detection methods usually provide a result in the form of a rectangular bounding box located around the person that has been detected. However, a bounding box is not always sufficient for many evolved applications such as action recognition, people recognition, or people clothes parsing that need more precise information on the person inside the bounding box: the silhouette of the person inside the bounding box is then needed. In the case of video study, the standard way is to consider motion-based background subtraction strategies (Sobral and Vacavant, 2014), that provide partial information on the belonging of pixels to the person silhouette. This initial information can be combined with seeded clustering methods such as graph cuts (Boykov and Kolmogorov, 2004) to obtain a precise people segmentation by accounting for the contextual informa-

**Corresponding author: Tel.: +33 231452927; fax:+33 231452698;
e-mail: olivier.lezoray@unicaen.fr (Olivier Lézoray)

tion in the image. However, background subtraction methods need temporal information to estimate the belonging of pixels to background or foreground, which is not possible for the case of single shot still images.

Considering the case of still images, few works have addressed the issue of people silhouette segmentation. Recent preliminary works have aimed at combining Convolutional Neural Networks with Conditional Random Fields for the general task of semantic scene labeling (Zheng et al., 2015). For the more specific tasks of people silhouette segmentation in still images, not so many works have been proposed so far. In (Jojic et al., 2009) the authors proposed 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 was also proposed. In (Migniot et al., 2011), the authors propose to use a shape template prior to roughly extract people. The shape template prior is obtained by a training step using a people segmentation ground truth in bounding boxes. This shape template prior is combined with a graph cut clustering method to refine the segmentation. Since this method relies on a very strong assumption on the position of people in bounding boxes obtained from people detection (people are supposed centered inside), the same authors have proposed in (Migniot et al., 2013) to modify the segmentation with edge detection priors.

So far, not so many methods are dedicated to the problem of people silhouette extraction from the output of people detection based methods in still images. One first explanation is probably related to the lack of segmentation ground truths for people segmentation datasets. Indeed, without ground truth labeled data, the developed methods cannot be adapted to the data by fitting a learned model and the results cannot be correctly evaluated. In a recent conference work (Coniglio et al., 2015), we have proposed a preliminary segmentation scheme for the segmentation of people using combined probabilities extracted from appearance, shape template prior, and color distributions. In the light of this, we propose in this paper to design a new complete strategy for precise people segmentation from bounding boxes obtained as the output of detection-based methods. To do so, we consider several possible image pre-processing associated with different people extraction methods based on different features. These people extraction methods provide segmentation probability maps from different priors that are combined with a weighted combination and finally refined with a graph-cut clustering. However, given the large amount of data, making the most of each method to achieve good segmentation results is difficult because all these pattern recognition techniques involve parameters that have to be manually tuned. To avoid this, we propose to automatically determine the best segmentation strategy as a whole, along with the best parameters of each constituting considered method, and that provides in addition a low processing time. This problem being difficult, we consider a genetic algorithm that can determine, the best configuration by computationally exploiting the available labeled data.

The paper is organized as follows. Section 2 presents the different components that compose the proposed segmentation schemes that have to be optimized. Section 4 describes the considered data sets. Section 5 presents our evaluation methodology, the best settings of each segmentation scheme and the results obtained on each dataset with comparison with state-of-the-art approaches. Last section concludes.

2. Proposed method

2.1. Synopsis of the approach

As mentioned in the introduction, there are few works dedicated to the precise segmentation of people inside bounding boxes, obtained from people detection methods. In addition there are few databases that provide ground-truth segmentation for evaluation and comparison purposes. This paper adresses both these issues. First a new people segmentation scheme inside a bounding box is developed. Second, exhaustive evaluation is performed with the help of new ground truth datasets. The segmentation scheme exploits some established methods from the literature and efficiently combines them for this specific problem. Figure 1 sums up the whole approach. Our scheme is designed to extract precisely people silhouettes in images in the form of a bounding box, and obtained using basic techniques of person detection. 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 a graph-cut segmentation that can operate at the pixel or the superpixel level. The proposed method has been designed in a modular way so as to be able to: assess the benefit of each step, control the complexity and the computational speed of the whole strategy. To do so, we consider eight different schemes that combine in different ways state-of-the-art methods for pre-processing, probability map estimation and segmentation refinement. For comparison purposes we will also consider other state-of-the-art methods (Migniot et al., 2011; Yang et al., 2013).

As shown in Figure 1, to segment precisely the people silhouette from a bounding box, three major steps are considered that we resume in the sequel:

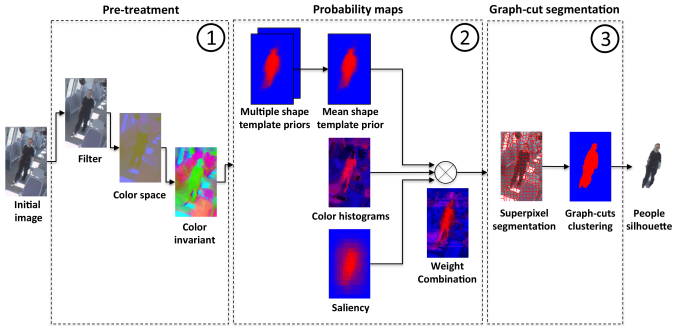

Fig. 1. Synopsis of our approach for segmenting people silhouette in bounding boxes.

① Image pre-treatment: the image of the bounding box can benefit from several pre-processings that can enhance the performance of the next steps, as we have shown in

(Coniglio et al., 2014). These processing steps are filtering, color space change and color invariant transformation. Different methods can be considered for each preprocessing and their parameters have to be tuned.

② Probability map estimation: this step aims at estimating the conditional probability of a pixel to belong to either background or foreground (the people silhouette). Different cues can be considered to perform this and operate with features from appearance, shape template, and color distributions priors, as we have considered in (Coniglio et al., 2015). A single probability estimator being not enough robust, several of them are combined.

③ Segmentation refinement: given the estimated probability map, a graph-cut classifies into two classes (foreground and background) the pixels from the combined estimation of conditional probabilities. This can be performed at the pixel or the superpixel level.

When we choose a specific configuration for each one of the three above steps, we obtain specific, but different, segmentation schemes. These different configurations are resumed in Table 1. The two first schemes correspond to the state-of-the-art approaches of (Migniot et al., 2011) and (Yang et al., 2013). The first one is very simple, yet effective, and combines a mean shape template prior with a graph cut clustering. The second one considers salient object extraction and segmentation for extracting the people silhouette. The next schemes 1-3 have a similar architecture but consider additional color space transformation and feature-based probability maps that are combined altogether. Schemes 4-6 are similar to scheme 3 but consider each a new information: scheme 4 considers the interest of saliency maps, scheme 5 adds some pretreatments (color space change, color invariant transformation and filter) and scheme 6 considers the use of a superpixel segmentation (based SLIC or ERGC methods). Finally, schemes 7 and 8 are similar to scheme 3 but consider two different multiple shape template priors (based on Histogram Difference or HOG + SVM classifier). As we will see with the experiments, each method is built on top of the previous by keeping its most important steps or enhancing some of them. We detail now each possible scheme that can be considered in these three steps. We do not provide too many details on the considered methods, and refer the reader to our previous works and reference therein (Coniglio et al., 2014, 2015).

2.2. Pre-treatment

The first step applied on bounding box images is performing several pre-processing methods listed in Table 2. We can apply three different consecutive steps of pre-processing : a color space change, a filtering and a color invariant transformation. These pre-processings steps can have several benefits: changing color space allows to better differentiate some colors, filtering reduces small artefact and noise, whereas color invariant is used to reduce light effects such as high brightness. These methods can be of essential importance for color histogram and graph-cut clustering steps. Since all these pre-processing steps involve parameters (choice of best pre-processing method and their settings), this will be done at once by the genetic optimization detailed in section 5.1.

Table 2. Different pre-treatment methods.

Color space	Filter	Colorimetric invariant
RGB	Gaussian	Greyworld
HSV	Median	Reduced coordinates
HLS	Bilateral	$l_1l_2l_3$
$L^*a^*b^*$		$m_1m_2m_3$
$L^*u^*v^*$		affine normalization
YUV		RGB rank

2.3. Probability maps

The second step allows determining a conditional probability of belonging to foreground (people silhouette) or background for each pixel. This information is extracted from three different kinds of methods: shape template, color histograms and saliency map priors. Three different possible shape template priors and two different color histogram priors are considered. When several probability maps are extracted by several methods, they are combined altogether with a weighted combination.

2.3.1. Shape template prior

Given a bounding box image, a first cue that can be considered is a shape template prior to estimate the people silhouette position.

Mean shape template prior. The use of shape template priors is common in literature (Migniot et al., 2010; Lin and Davis, 2010). 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 mean shape template prior based probability map is obtained from an averaging of all the ground-truth shapes in a given training set. This prior is therefore computed only once.

Multiple shape template priors. It is obvious that a simple mean shape template prior cannot well account for the different poses and angles of view that can occur. To cope with this, we propose to construct multiple shape template priors and to automatically pick up the most appropriate one, similarly to what was proposed in (Migniot et al., 2011). The idea is to dispose of a shape template adapted to the person posture. To dispose of multiple shape template priors, we proceed in the following way. The ground truth training dataset is clustered in k clusters with a k-means algorithm (Kanungo et al., 2002). The optimal number of k clusters that determines the number of multiple shape priors will be determined at once by a genetic algorithm. This phase is done offline. Once the multiple shape priors are available, two different methods are considered

Table 1. Table showing, for each people silhouette segmentation method, the considered processing steps.

	State-of-the-art approaches		Proposed methods (with number of genes of each block)							
Step / Scheme	**(Migniot et al., 2011)**	**(Yang et al., 2013)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
① **Pre-treatment**										
Color space			✓(1)	✓(1)	✓(1)	✓(1)	✓(1)	✓(1)	✓(1)	✓(1)
Filter							✓(4)			
Color invariant							✓(1)			
② **Probability maps**										
Multiple Shape template prior										
Histogram Difference									✓(3)	
HOG + SVM										✓(1)
Mean Shape template prior	✓		✓(0)	✓(0)	✓(0)	✓(0)	✓(0)	✓(0)		
Color histograms				✓(2)						
Color histograms strips					✓(3)	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)
Saliency		✓				✓(3)				
Weight combination				✓(2)	✓(2)	✓(3)	✓(2)	✓(2)	✓(2)	✓(2)
③ **Graph-cut segmentation**										
Superpixels segmentation								✓(3)		
Graph-cuts clustering	✓		✓(3)	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)	✓(3)
Total number of genes			(4)	(8)	(9)	(13)	(14)	(12)	(12)	(10)

to determine, given a bounding box, the most plausible shape template prior. A first method is based on color appearance of foreground and background, modeled by color histograms. Given a shape template prior, we initialize these histograms with the colors of the original image, each being weighted by the considered template shape prior. The absolute difference between the two histograms provides a score and the shape template prior with the highest score is retained. The choice of the best color histogram representation (color space and number of bins) will be performed at once with the genetic algorithm. A second method considers histograms of oriented gradients as features combined with Support Vector Machines to determine the most appropriate shape template prior. This method needs to be trained with positive and negative example. The training is performed from the result of the ground truth clustering step: original images associated to a cluster of ground truths are used as positive image to train one class of the SVM and other original images are used as negative examples. Each trained HOG + SVM classifier is then used to determine the best shape template prior (in a one-versus-all approach). The choice of the best kernel for the SVM classifier will be performed at once with the genetic algorithm.

2.3.2. Color Histograms priors

The second cue that we consider uses an appearance-based prior with the color of pixels. We have chosen to define the appearance with histograms methods. Histograms-based methods have the advantage to be much faster than other methods such as Gaussian Mixture Models and can be enough precise to represent the color appearance within a bounding box. We use two color histograms, one for background and one for foreground (people silhouette). Two different types of color histograms are built. The first considers an histogram computed on the whole bounding box, and is weighted with the shape template prior previously defined. The second considers the concatenation of histograms computed on strips in the bounding box. The two types of histograms have several parameters: number of bins and the number of strips for the second method. These parameters will be optimized at once with the genetic algorithm. Given these color histograms, we estimate the class memberships using the color distributions.

2.3.3. Saliency prior

Saliency detection has recently received a lot of attention in image processing (Yang et al., 2013; Cheng et al., 2015). Since the most salient object inside a bounding box is supposed to be the person of interest inside it, saliency detection and segmentation methods can be considered as good candidates for background or foreground probability estimation. Indeed, most of these methods provide a membership to these two classes that we convert in the form of a probability map. We have considered the approach of (Yang et al., 2013) that provides good saliency estimation on reference benchmarks. This method has three parameters to tune: a superpixel compactness value, an error rate and a filtering. These three parameters values will be determined at once by the genetic optimization step.

2.3.4. Probability map weighted combination

The three cues presented previously provide possibly 6 different information on the possible position of people inside the bounding box. This information is represented by a probability of belonging to foreground $P^{foreground}(p_i)$ or background $P^{background}(p_i)$ for each pixel. When several cues are considered, a combination is necessary, and a weighted combination is performed to obtain the global final probability estimation for each pixel:

$$P^{foreground}(p_i) = \sum_k \theta_k P_k^{foreground}(p_i) \text{ with } \sum_k \theta_k = 1 \qquad (1)$$

where $P_k^{foreground}(p_i)$ denotes the conditional estimated probability from the k-th map for the foreground (people) class. The same formula applies respectively for $P^{background}(p_i)$. A genetic optimization will be used at once to determine the optimal weighting coefficients θ_k of the k considered cues.

3. Graph-cut segmentation from probability maps

The final step consists in classifying the image given the estimation of probabilities obtained from the combination of probability maps into two classes (foreground: people and background). For that we use a Graph-cut (Boykov and Jolly, 2001) method. Graph-cut techniques are among the most powerful methods that extract foreground from background. Graph-cut 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 $G = (V, E)$ composed of $|V|$ nodes, where each node p_i is assigned a label $l_i \in L$ and $|E|$ edges. Two different types of graphs are considered (Lézoray and Grady, 2012): 8-grid graphs where nodes correspond to pixels and superpixel graphs where vertices correspond to superpixel regions. The set of edges are both inferred by direct spatial neighborhood relationship. For superpixel generation, we consider two methods: SLIC (Achanta et al., 2012) and Eikonal-based Region Growing Clustering (ERGC) (Buyssens et al., 2014). Both are used only in the scheme 6 for segmentation.

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 foreground: 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 called capacities and is defined as :

$$W^{l_i}(p_i) = -\gamma * \log(P^{l_i}(p_i)) \quad (3)$$

It uses the probabilities of each vertex to belong to the l_i class (people or background), and is obtained from the weighted combination probability map. When superpixels are considered instead of pixels, the average of probability values is computed over the entire region. The second term is called similarities and 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 edges of p_i with others vertices of graph. The term $S(p_i, p_j)$ expresses a similarity measure between both vertices p_i and p_j and is given by:

$$S(p_i, p_j) = \alpha * exp\left(-\frac{d(p_i, p_j)}{2 * \beta^2} \right) . \frac{1}{dist(p_i, p_j)} \quad (4)$$

with

$$d(s_i, s_j) = \sqrt{\sum_{c=1}^{3} (s_i^c - s_j^c)^2} \quad (5)$$

Where $dist(p_i, p_j)$ is the Euclidean distance between the vertices (the center of superpixels for the case of super pixel graphs). The quantity $d(p_i, p_j)$ denotes the sum of distances between the color channels p_i^c and p_j^c associated to vertices p_i and p_j (average colors for the case of super pixel graphs). The optimization is done with the min-cut/max-flow implementation of (Boykov and Kolmogorov, 2004). The result of graph cut labeling is a binary image where each vertex has been assigned to one class among background and foreground: people. Therefore, we obtain the final people silhouette.

The three parameters α, θ, γ that appear in the capacities and similarities formula are coefficients of great importance in the final segmentation results. These parameters will be determined at once by the genetic optimization.

4. Datasets

To test the ability of our proposed segmentation schemes we consider several datasets. We have picked datasets from people detection and people recognition challenges, where supplied images are from a people-detection-based method. We have also selected datasets in order to have various poses, angles of view and illumination problems. In total we have tested 4 datasets divided into 6 sets, with a total of 1,797 images. Images are mainly recorded in outdoors and transportation environments. People may wear different clothes and may carry various objects (bag, suitcase *etc.*). Moreover people may be in crowd and be surrounded by other people. In these cases we have considered objects carried by people as foreground classes and the people surrounding as background classes. Indeed we consider that the people detection method has been trained to detect one person. All the ground truths used for evaluation have been handmade in order to have a precise evaluation.

4.1. VIPeR dataset

The ViPeR dataset (Gong et al., 2014) is very popular in the evaluation of people-recognition methods. The dataset is composed of bounding boxes of 128x48 pixels of people who are walking and was recorded in streets during the day. People can have several poses (front, side and back). Several colorimetric problems are present with strong illumination and gloomy images. People wear different clothes and can carry several objects (suitcase, bag, clothes, *etc.*). We have noticed that the usual reference ground truth of people segmentation used in the state-of-the-art approach provided by the STEL approach (Jojic et al., 2009) contains several mistakes. So we have handmade by ourselves 250 more precise ground truth images.

4.2. PRID 2011 dataset

The people re-ID 2011 (Hirzer et al., 2011) (PRID 2011) is also a dataset used in the people-recognition challenges. The provided images consist of a set of bounding boxes describing the motion of people. Bounding boxes are focused on people with an angle, that changes between side and perspective. People are recorded in front or back pose. The background is homogeneous in bounding boxes (composed of flagstone). Images have a colorimetric problem in that they tend to be green. Images sizes are 128x64 pixels. People wear different clothes, can carry objects (newsletters, bag) and can push a stroller. We have handmade 250 precise ground truth images with diverse people.

4.3. INRIA Person dataset

The INRIA Person dataset (Dalal and Triggs, 2005b) is also used in people-detection challenges. It has been used to train or to test most people-detection-based methods. It is composed of homogenous images recorded in streets, mountains, and forests where people are in various poses (front, side, back) with different activities (posing for photo, sport, dance). Many images contain several people in the same bounding box representative of crowed environments. Two sizes of images are available: 128x64 pixels and 160x96 pixels. The difference between the two sizes corresponds to the addition of background around the bounding box. Evaluation is done with 390 precise handmade ground truth images provided by (Migniot et al., 2010).

4.4. BOSS European Dataset

The BOSS European Dataset (Boss, 2009) is focused on action recognition and people recognition challenges for transportation environments. The dataset is composed of several videos recorded in a train in motion during a sunny afternoon with different scenarios (people walking, fighting). This dataset does not contain bounding boxes of people detection. To generate these, we have used the well-known people-detection method of (Dalal and Triggs, 2005a) based on Histogram Of Gradients (HOG) combined with Support Vector Machines (SVM) to extract bounding boxes of 160x96 pixels in two sequences. The first set was made from one sequence, and is composed of 453 images divided into 12 people walking in front of the camera. The second set was made from an another sequence, and is composed of 64 images divided into 11 people walking with a perspective angle of view. Each set of images was mixed in order not to find temporal constancy between images. A precise handmade ground truth was completed for each image.

5. Evaluation

To perform the evaluation of the investigated segmentation schemes we propose, a precise methodology is necessary. Indeed, in many papers, the separation between training and test sets is not clear and the provided results are often optimistic since they were obtained for a given (arbitrary) partitioning of the ground truth dataset. A good methodology to evaluate classification results is to use a K-fold cross validation to separate test and training sets, and this enable to obtain scores that are closer to the empirical risk. To assess the performance of the proposed methods, we consider two scores: classification accuracy and processing time. This enables to compare the different methods on different levels (see Table 3). Since all the segmentation schemes are genetically optimized to choose the better tuning of methods and their associated parameters, Table 4 shows the one obtained for all segmentation schemes. Finally, several visual results are provided in Figures 4, 5, 6 and 7. Before entering into the analysis of the result, we detail how genetic optimization is performed and how classification accuracy is computed by K-fold cross validation.

5.1. Genetic algorithm

As we have mentioned it, the whole strategy we propose involves a lot of different possibilities e.g., pre-filtering methods and associated parameters. That is why a genetic optimization is used at once to automatically determine the best configurations. It is performed with a population of 120 chromosomes that encode possible solutions. Each chromosome corresponds to a complete setting of our proposed people silhouette extraction method. A chromosome is divided into several blocks which are composed of one or several genes (illustrated in parentheses in Table 1). Each gene encodes either the use of a method (binary gene) or the value of a parameter (quantized possible values). As an example, let us consider the filter block of Table 1 used in the 5^{th} scheme. It is composed of 4 genes: one gene is used for the choice of the filter and the other three are used for the filter parameters. The genetic algorithm uses a standard configuration and is composed of four steps (initialization, selection, crossover and mutation). The algorithm begins with an initialization step and iteratively processes steps of selection, crossover and mutation until the population is stable:

- The initialization step constructs a list of candidate solutions (called population). The initialization of the chromosome is made by block. Thus, the genes corresponding to a choice of method are first randomly initialized. Then, the genes corresponding to methods parameters are in turn randomly initialized. Finally, if certain genes are not used, they are set to zero. Figure 2 illustrates the genes encoding of the filter block. The gene corresponding to the method is marked in pink color, those corresponding to their associated parameters are marked in green color and the genes that are not used are marked in yellow color. As an example, let us consider the gaussian blur filter, the first gene corresponds to the filter type, the next two indicate the size of this filter, and the last gene that is not used is set to zero. Let's now consider the example of the bilateral filter, as mentioned above, the first gene indicates the filter type, while the other three correspond to their parameters (neighborhood, sigma space and sigma color).

- The crossover step aims at generating new candidate solutions from existing ones in the population. A child is produced from mixing two chromosomes randomly chosen. Figure 3 illustrates an example of this crossover step in our genetical algorithm. The child is obtained by successively copying the block of one of the two parents (A or B). A random sampling is used to determine the block to copy.

- The mutation step consists in slightly modifying a part of the new chromosome generated in the crossover step. This choice has been done in order to obtain new chromosomes different from parents and also to cover a wide range of solutions. For that, we distinguish two cases illustrated in Figure 3: 1/ In the case of a mutation of a gene corresponding to the choice of a method (illustrated in orange color in Figure 3): the selected gene is randomly modified with

a mutation rate of 25%. The other genes of the considered block are re-initialized (as described in the initialization step); 2/ In the case of a mutation of a gene corresponding to a parameter of a method (illustrated in purple color in Figure 3): the selected gene is randomly modified with a mutation rate of 25% or a small value is added within an interval that is plus or minus 10% of the actual value with a mutation rate of 50%. These parameters values are selected in order to let the genetic algorithm converge quickly.

- For the selection step, the candidates are sorted according to the fitness score detailed below. The selected candidates (the first half) are kept in order to obtain a stable size of population for each generation. The other ones (the second half) are rejected. The genetic algorithm is stopped when the best candidate of the population has not changed during 10 generations.

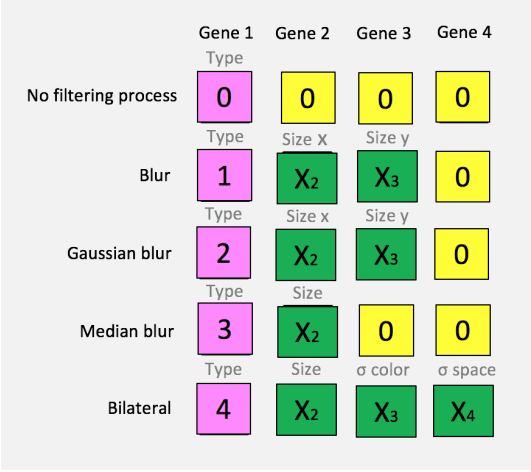

Fig. 2. Genes encoding of the filter block.

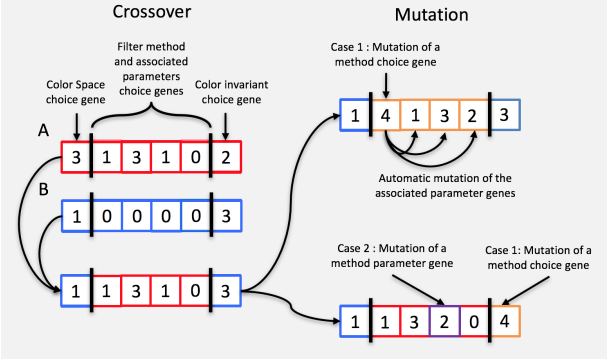

Fig. 3. Example of crossover and mutation steps in our genetic algorithm.

5.2. K-fold Cross Validation

For each segmentation scheme, the training set is used for the learning of methods that compose the scheme and to determine the best parameter settings of the considered methods. This task being impossible to do by hand, each segmentation scheme is optimized as a whole using a genetic algorithm that automatically determines the best configuration according to a fitness measure defined as the classification accuracy. These learning and tuning are performed on the training dataset, the test set is used only for the evaluation of the segmentation scheme. To have an estimated accuracy close to the empirical risk, a K-fold cross validation is performed whenever an evaluation is needed, and each obtained model is estimated according to the F-measure score (higher score is better):

$$F\text{-measure} = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{7}$$

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{8}$$

For the optimization performed by the genetic algorithm, a 8-fold cross validation is performed on the training set. This enables to use the F-measure on the training set as a fitness score. A F-measure score is computed for each image of the training set and averaged on all the images to define the classification accuracy of that fold. The average of this accuracy on all folds provides the score of the scheme on the training dataset under study. A similar procedure is performed to estimate the processing time.

If K-fold cross validation provides an accurate estimation of the segmentation scheme performance on the training set, this does not directly provide a final scheme, since K schemes are generated. We have chosen to retain the one with the highest classification accuracy on the training set for the generation of the results in Figures 4, 5, 6 and 7. To assess the performed of that retained scheme, an evaluation by 8-fold cross validation is performed but only the test set this time, and an average F-measure is retained.

5.3. Experimental results

Table 3 shows the results obtained for each dataset with each segmentation scheme whereas Table 4 shows the parameters tuned by the genetic optimization. Table 3 shows results divided into 2 categories: F-measure and processing time. We have chosen K = 8-fold-cross validation for all datasets and schemes. Each scheme has been performed on one CPU thread cadenced at 3.4Ghz. To have a fair evaluation with state-of-the-art approaches, the schemes of (Migniot et al., 2011) and (Yang et al., 2013) have also been genetically optimized. We consider the segmentation scheme of (Migniot et al., 2011) as the reference scheme and this constitutes our baseline that we would like to overcome. In table 3, the results appearing in green color are equivalent to or higher than the best method of the-state-of-the-art approaches and those appear in red color correspond to the best of our overall score.

We first analyze the results presented in Table 3. The baseline results of (Migniot et al., 2011) and the best scores are bold faced. Scheme of (Migniot et al., 2011) that we consider as our baseline is a simple method (mean shape template prior + graph

Table 3. F-measure results and time processing obtained for the considered segmentation schemes with respect to the state-of-the-art approaches of (Migniot et al., 2011), (Yang et al., 2013)

	State-of-the-art approaches		Proposed methods							
Dataset / Scheme #	**(Migniot et al., 2011)**	**(Yang et al., 2013)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
BOSS S-1 (160x96)	**0.899**	0.550	0.897	0.905	**0.913**	0.912	0.916	0.840	0.900	0.895
(Boss, 2009)	**16ms**	90ms	16ms	22ms	**23ms**	118ms	25ms	110ms	39ms	28ms
BOSS S-2 (160x96)	**0.857**	0.660	0.862	0.871	**0.883**	0.881	0.881	0.810	0.860	0.851
(Boss, 2009)	**16ms**	100ms	16ms	25ms	**25ms**	121ms	203ms	108ms	33ms	27ms
INRIA (128x64)	**0.839**	0.730	0.837	0.854	**0.859**	0.860	0.859	0.790	0.845	0.832
(Dalal and Triggs, 2005b)	**9ms**	59ms	11ms	14ms	**15ms**	73ms	60ms	82ms	18ms	15ms
INRIA (160x96)	**0.839**	0.590	0.839	0.852	**0.855**	0.850	0.856	0.800	0.848	0.835
(Dalal and Triggs, 2005b)	**15ms**	92ms	15ms	24ms	**25ms**	114ms	110ms	110ms	32ms	29ms
VIPeR (128x48)	**0.887**	0.750	0.887	0.887	**0.894**	0.900	0.894	0.830	0.883	0.864
(Gong et al., 2014)	**10ms**	60ms	10ms	12ms	**13ms**	86ms	82ms	74ms	17ms	15ms
PRID2011 (128x64)	**0.818**	0.780	0.832	0.894	**0.900**	0.900	0.876	0.813	0.882	0.822
(Hirzer et al., 2011)	**12ms**	59ms	12ms	18ms	**20ms**	84ms	38ms	81ms	33ms	29ms
Average	**0.856**	0.677	0.859	0.877	**0.884**	0.883	0.880	0.814	0.870	0.850
	13ms	77ms	13ms	19ms	**20ms**	99ms	86ms	94ms	26ms	24ms

cut clustering) but provides good results on almost all datasets (an average score of 0.856). This scheme is interesting since its processing time is very low (an average processing time of 13 ms). Indeed, the shape prior being computed offline on the whole dataset, only the graph-cut has to be run on the bounding box.

If we now have a look to variations of this segmentation schemes (schemes 1 to 3), we can see some enhancements. Scheme 1 adds color space change, scheme 2 adds an appearance-based prior (with global color histograms) associated with a weighted combination and scheme 3 replaces the global color histograms with concatenated stripped color histograms. Each modification from schemes 1 to 3 enables to gradually enhance the results. This shows the interest to change the color space and to combine the shape prior with an appearance-based prior. Scheme 3 always provides results that are much better than the baseline results of (Migniot et al., 2011) and shows all these benefits (respectively average scores of 0.859, 0.877 and the best with 0.884). However, this comes with additional costs in processing time of approximately 50% (extended from 13ms to 20ms). This additional time is mainly due to the addition of the color histogram step. Fortunately, this processing time is still compatible with real time processing, and can benefit from code enhancements.

Results obtained with the scheme of (Yang et al., 2013) show that saliency cue considered alone is not sufficient to obtain state-of-the-art results. Indeed, for all tested datasets, the segmentation results are lower than those obtained with (Migniot et al., 2011) and the average processing time is high (77ms). However, once combined with the scheme 3, this enables to somehow attain or enhance the results obtained with some datasets (INRIA (128x64), VIPeR and PRID2011). As this can be seen from the results, this is not however very concluding since this is at the cost of high processing time, so saliency is not retained as an interesting cue for people extraction.

In contrast, scheme 5 adds some additional pre-processing to scheme 3 (filtering and color invariants), and this enables to further enhance the results of scheme 3 on BOSS-S1 and INRIA (160x96) (respectively a score of 0.916 and 0.856). This method is interesting but the average computing cost is multiplied by a factor of 4.3 with an average accuracy very close to scheme 3. The difference of the processing time is due to the complexity of the filters, knowing that the genetic algorithm does not take this information into account for optimizing parameters. For example with this scheme, the processing time increases from 25ms (BOSS-S1 (160x96)) to 110ms (INRIA (160x96)). The addition of more pre-processings is therefore not very concluding.

Scheme 6 considers the interest of working at the superpixel level instead of the pixel level for the graph-cut clustering. As it can be seen in Table 4, the best superpixel method is ERGC but this lowers the classification accuracy (an average score of 0.814). On the one hand, disposing of superpixels enables to work on a graph of reduced size for graph-cut clustering, but this is at the cost of computing the superpixels, which is high as it can be seen (an average processing time of 94 ms). In addition if the superpixel is not accurate, this has a very strong influence on the final accuracy and the use of superpixels is therefore not concluding at all.

Finally, schemes 7 and 8 consider replacing the mean shape template prior of scheme 3 (that provided the best results in average) by optimized multiple shape template priors. This processing time remains comparable to scheme 3 (respectively a processing time of 26ms and 24ms) but the accuracy are lower than with a simple mean shape prior (respectively a score of 0.870 and 0.850). This can seem surprising but confirms similar results obtained in (Migniot et al., 2010, 2011, 2013). In light of these observations, we propose to retain the scheme 3 (which is the best in average) with the optimized parameters given in Table 4.

We have mentioned in section 2 that each method considered in a given segmentation scheme is automatically optimized at once with the help of a genetic algorithm. This is done at two levels: choosing the best method when several ones are available (e.g., choosing the right filtering method), choosing the best parameters of a given methods (e.g., the best parameters of the graph-cut clustering). Table 4 resumes the optimal methods and parameters that have been retained for each seg-

mentation scheme. If there are some small changes, one can see that there is constancy between the optimal settings of the different schemes. The color space choice, color histograms and graph-cuts, that are the most important methods of the best schemes, keep similar configurations. In addition, the settings found by the weights of the combination step of different priors enables to see that the appearance-based prior is of essential importance. This shows the difference between the best segmentation scheme we have retained (scheme 3) and the state-of-the-art approach of (Migniot et al., 2011).

Figures 4, 5, 6 and 7 show a comparison between the results obtained with the baseline results of (Migniot et al., 2011) and our best proposed scheme (scheme 3). Whatever the dataset considered, the results for both these schemes are very satisfying. The people are always correctly extracted and background is well detected. Nevertheless with the baseline results of (Migniot et al., 2011) small error appear mainly with bad people extremity segmentation (foot, hand, legs *etc.*). Some of these problems are corrected with the proposed scheme 3.

Fig. 6. People silhouette extraction results on the VIPeR dataset (Gong et al., 2014) (line 1 : 6 original images; line 2: the people extraction results obtained with the state-of-the-art scheme of (Migniot et al., 2011); line 3: the people extraction results obtained with our proposed strategy (scheme 3).

6. Conclusion

In this paper, we have considered a problem that is not so commonly considered in literature with respect to the problem of people detection in images: the problem of people silhouette segmentation inside bounding boxes that are outputs of typical people detection methods. On the basis of the state-of-the-art approach of (Migniot et al., 2011), we have proposed several possible enhancements to this segmentation scheme. The following ones have been considered: image pre-processing,

Fig. 7. People silhouette extraction results on the PRID2011 dataset (Hirzer et al., 2011) (line 1 : 6 original images; line 2: the people extraction results obtained with the state-of-the-art scheme of (Migniot et al., 2011); line 3: the people extraction results obtained with our proposed strategy (scheme 3).

appearance-based priors and superpixel graphs. Since it is very difficult to assess the benefit of adding one typical method inside a given segmentation scheme, we have considered an approach driven by the data to evaluate the benefit of each segmentation scheme. To do so, precise handmade segmentations of people silhouette have been made, and a systematic optimization of the composing methods has been performed at once by a genetic algorithm. This enables to more accurately evaluate the benefit of one scheme versus the others. With such a systematic approach, we have been able to design a segmentation scheme that goes beyond the actual state-of-the-art by incorporating a color space change, a weighted combination of mean shape and appearance-based priors, and graph-cut clustering. The approach is at the end enough fast to be deployed for real time processing, which is essential for industrial applications. In future works, we plan to tackle the problem of people re-identification from their extracted silhouettes.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE T PATTERN ANAL 34, 2274–2282.

Boss, 2009. Boss europeen project (on board wireless secured video surveillance). URL: `http://www.multitel.be/image/research-development/research-projects/boss.php`.

Boykov, Y., Jolly, M., 2001. Interactive graph cuts for optimal boundary region segmentation of objects in n-d images, in: ICCV, pp. 105–112.

Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE T PATTERN ANAL 26, 1124–1137.

Buyssens, P., Gardin, I., Ruan, S., Elmoataz, A., 2014. Eikonal-based region growing for efficient clustering. IMAGE VISION COMPUT 32, 1045–1054.

Cheng, M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S., 2015. Global contrast based salient region detection. IEEE T PATTERN ANAL 37, 569–582.

Table 4. Best settings of methods obtained for the segmentation schemes with the genetical optimization

	State-of-the-art approaches		Proposed methods							
Step / Scheme #	**Migniot et al.**	**Yang et al.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
① **Pre-treatment**										
.Color space	-	-	Lab	Lab	**YUV**	YUV	Lab	YUV	YUV	Luv
.Filter	-	-	-	-	-	-	Bilateral	-	-	-
.Color invariant	-	-	-	-	-	-	RGB rank	-	-	-
② **Probability maps**										
Multiple Shape template priors										
Histogram difference										
.color space	-	-	-	-	-	-	-	-	HLS	-
.#bins color histogram	-	-	-	-	-	-	-	-	78	-
HOG + SVM										
.SVM kernel	-	-	-	-	-	-	-	-	-	Linear
Mean Shape template prior										
.# class	-	-	-	-	-	-	-	-	2	3
Color histograms										
.FG # bins	-	-	-	36	-	-	-	-	-	-
.BG # bins	-	-	-	96	-	-	-	-	-	-
Color histograms strip										
.FG # bins	-	-	-	-	**25**	68	102	15	57	75
.BG # bins	-	-	-	-	**61**	91	118	106	106	138
.# strips	-	-	-	-	**32**	31	35	28	37	31
Saliency										
.Superpixels compactness	-	76	-	-	-	276	-	-	-	-
.error	-	0.218	-	-	-	0.503	-	-	-	-
.Filter	-	Gaussian	-	-	-	Gaussian	-	-	-	-
Weight combination										
.Shape template prior	-	-	-	51%	**39%**	17%	33%	42%	32%	47%
.Color histograms	-	-	-	49%	**61%**	65%	67%	58%	68%	53%
.Saliency	-	-	-	-		18%	-	-	-	-
③ **Graph-cut segmentation**										
Superpixels segmentation										
.Method	-	-	-	-	-	-	-	ERGC	-	-
.Size	-	-	-	-	-	-	-	$\sim 49px$	-	-
.Compactness	-	-	-	-	-	-	-	110	-	-
Graph cuts clustering										
.α	82	-	84	77	**60**	62	81	69	70	63
.β	6	-	6	10	**13**	11	23	13	29	25
.γ	3	-	3	20	**25**	26	14	24	41	50

Fig. 4. People silhouette extraction results on the BOSS dataset (Boss, 2009) (line 1 : 6 images of the sequence 1 + 6 images on the sequence 2; line 2: the people extraction results obtained with the state-of-the-art scheme of (Migniot et al., 2011); line 3: the people extraction results obtained with our proposed strategy (scheme 3).

Fig. 5. People silhouette extraction results on the INRIA dataset (Dalal and Triggs, 2005b) (line 1 : 6 images with 128x64 resolution + 6 images with 160x96 resolution; line 2: the people extraction results obtained with the state-of-the-art scheme of (Migniot et al., 2011); line 3: the people extraction results obtained with our proposed strategy (scheme 3).

Coniglio, C., Meurie, C., Lézoray, O., Berbineau, M., 2014. A genetically optimized graph-based people extraction method for embedded transportation systems real conditions, in: ITSC, pp. 1589–1595.

Coniglio, C., Meurie, C., Lézoray, O., Berbineau, M., 2015. A graph based people silhouette segmentation using combined probabilities extracted from appearance, shape template prior, and color distributions, in: ACIVS, pp. 299–310.

Dalal, N., Triggs, B., 2005a. Histograms of oriented gradients for human detection, in: CVPR, pp. 886–893.

Dalal, N., Triggs, B., 2005b. INRIA person dataset. Technical Report. INRIA.

Gong, S., Cristanio, M., Yan, S., Loy, C., 2014. Person re-identification. Springer.

Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE INTELL SYST APP 13, 18–28.

Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H., 2011. Person re-identification by descriptive and discriminative classification, in: Image Analysis. Springer, pp. 91–102.

Jojic, N., Perina, A., Cristani, M., Murino, V., Frey, B., 2009. Stel component analysis: Modeling spatial correlations in image class structure, in: CVPR, pp. 2044–2051.

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., Wu, A., 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE T PATTERN ANAL 24, 881–892.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks, in: NIPS, pp. 1097–1105.

Lézoray, O., Grady, L., 2012. Image Processing and Analysis with Graphs: Theory and Practice. Digital Imaging and Computer Vision, CRC Press / Taylor and Francis.

Lin, Z., Davis, L.S., 2010. Shape-based human detection and segmentation via hierarchical part-template matching. IEEE T PATTERN ANAL , 604–618.

Migniot, C., Bertolino, P., Chassery, J.M., 2010. Contour segment analysis for human silhouette pre-segmentation, in: VISAPP, pp. 74–80.

Migniot, C., Bertolino, P., Chassery, J.M., 2013. Iterative human segmentation from detection windows using contour segment analysis, in: VISAPP, pp. 405–412.

Migniot, C., Bertolino, P., J-M.Chassery, 2011. Automatic people segmentation with a template-driven graph cut, in: ICIP, pp. 3149–3152.

Mohan, A., Papageorgiou, C., Poggio, T., 2001. Example-based object detection in images by components. IEEE T PATTERN ANAL 23, 349–361.

Redmon, J., Angelova, A., 2015. Real-time grasp detection using convolutional neural networks, in: ICRA, pp. 1316–1322.

Siala, M., Khlifa, N., Bremond, F., Hamrouni, K., 2009. People detection in complex scene using a cascade of boosted classifiers based on haar-like-features, in: IV, pp. 83–87.

Sobral, A., Vacavant, A., 2014. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. COMPUT VIS IMAGE UND 122, 4 – 21.

Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H., 2013. Saliency detection via graph-based manifold ranking, in: CVPR, pp. 3166–3173.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S., 2015. Conditional random fields as recurrent neural networks, in: ICCV, pp. 1529–1537.

Zhu, Q., Avidan, S., Yeh, M.C., Cheng, K.T., 2006. Fast human detection using a cascade of histograms of oriented gradients, in: CVPR, pp. 1491–1498.