2015 IEEE International Symposium on Multimedia

Shape-based pedestrian segmentation in still images

Julio Cezar Silveira Jacques Junior and Soraia Raupp Musse
Faculdade de Informática
Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS
Porto Aletre, RS - Brazil
julio.jacques@acad.pucrs.br - soraia.musse@pucrs.br

***Abstract*—Pedestrian segmentation is a problem of considerable practical interest. In this work we propose a shape-based model for pedestrian segmentation. Our model is initialized by a bounding-box of the person under analysis, which can be estimated by a person detector. The basic idea of the proposed model is to create a graph around the detected person, based on a scale invariant shape model and the estimated contour is given by a path in the graph that maximizes certain boundary energy. In practice, such energy should be large in the boundary between the foreground/background. To cope with pose/shape variations, the final estimate is given by a selection scheme, which takes into consideration the individual estimate given by different generated graphs. Experimental results indicated that the proposed technique works well in non trivial images, with comparable accuracy to the state-of-the-art.**

***Keywords*-pedestrian segmentation; contour estimation;**

I. Introduction

The automatic segmentation of human subjects in static images is still a challenge, mainly due to the influence of numerous real-world factors such as shading, image noise, occlusions, background clutter as well as other factors associated with the dynamics of the human being (great variability of poses, shapes, clothes, etc). Pedestrian segmentation can be considered a special case of person/human segmentation. As related in the work of Dollar et al. [1], people analysis, by computer vision techniques, makes the use of datasets containing people in unconstrained pose in a wide range of domains whilst the area of pedestrian analysis uses datasets containing upright people (standing or walking), typically viewed from more restricted viewpoints.

Pedestrian segmentation can be used in several applications, including robotics, surveillance systems, driver assistance models, among others. Farenzena and his group [2], for example, assume the presence of the silhouette of an individual, obtained for each person by inferring over the STEL generative model [3], to extract appearance features applied in a person re-identification problem.

In this work we focus on the case where an external pedestrian detector is used to provide regions of interest to our segmentation model (i.e., bounding boxes, as in [4], [5], [6], [7], [8], which can be estimated by a person detector [9]). In a nutshell, the basic idea of the proposed model is to create structured graphs around the detected person (with different shapes, in order to deal with pose/shape variations), then defining the contour by a path in one of these graphs, which is given by a selection scheme.

The main contributions of the proposed model is an automatic and scale invariant approach to estimate pedestrian contours in still images, proposed to be robust to occlusion, shape/pose variations, as well as to cope with pedestrians captured by different views (i.e., frontal/back foot closed, rightwards feet open, etc.). In addition, the proposed model is executed without appearance cues (e.g., color or texture-based features) as most state-of-the-art models do. Such characteristic could be taken into consideration when color cues are missed or weakened (e.g., at night conditions or when infrared/thermal cameras are employed [10]).

II. Related Work

Numerous approaches have been proposed for pedestrian segmentation in the last years, employing global segmentation [11], [4], [5] or part-based schemes [6], [7], [8]. Gao et al. [11] presented a pedestrian detector approach exploring the Local Segmentation Self-Similarity (LSSS) descriptor. The shape segmentation is the base of their work. To attain satisfactory pedestrian contour, they employ a strategy using color features in the Lab colorspace, which is adopted from [12]. Firstly, a saliency map is obtained by a histogram based contrast method (HC) which integrate spatial relationships into region level contrast computation, then GrabCut [13] is applied to refine the segmentation result initially obtained by thresholding the saliency map. Flohr and Gavrial [4] presented an iterative EM-like framework (Expectation-Maximization-like) for accurate pedestrian segmentation, combining generative shape models and multiple data cues. In the initialization phase, pedestrian shape exemplars are obtained manually and processed in order to obtain 12 clusters. In the segmentation step, EM-like framework is proposed. In E-step, a Conditional Random Field formulation is used combining color, texture and disparity cues (when given). Active Shape Model is used in M-step and then EM process iteratively.

Li et al. [5] presented an approach to segment pedestrians in still images, combining shape and appearance cues. A hierarchical shape matching is employed to extract pedestrian silhouette and skeleton (transferred from the template,

978-1-5090-0379-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISM.2015.25

learned in a previous stage, to the image), used to refine the segmentation via Graph Cuts. Head-torso detections and parsing are applied to fine tuning the initial shape matching. Eslami and Williams [6] extend the Shape Boltzmann Machine [14] model (SBM), applied for the task of modeling binary shape images, to account for the object's parts. Their Multinomial SBM is combined with an appearance model to form a fully generative model of images of objects. Parts-based object segmentations are obtained by performing probabilistic inference in the model.

The work presented by Bo and Fowlkes [7] described a model for pedestrian parsing based on his shape. This technique assembles candidate parts from an oversegmentation of the image and matches them to a library of exemplars. Authors use a hierarchical decomposition into a variable number of parts and computes scores on partial matching in order to prune the search space of candidate segment. In addition, color and texture histograms are used to capture the appearance of specific body parts. Luo et al. [8] proposed a Deep Decompositional Network (DDN), using HOG features [9] as input, for parsing pedestrian images into semantic regions, such as hair, head, body, arms, and legs. They argue that DDN can jointly estimate occluded regions and segment body parts by stacking three types of hidden layers: occlusion estimation layers, completion layers and decomposition layers.

III. The Proposed Model

The main idea of the proposed model is to create a directed weighted graph around the detected person, based on a scale invariant shape model (i.e., the graph can be resized according to the height of the detected person, which is derived from his/her bounding-box size), and to find a path in the graph that maximizes certain boundary energy, describing the person's contour. This formulation was derived from our previous work [15], proposed for the problem of head-shoulder human contour estimation (considering people captured in frontal poses, initialized by a face detector and assuming one single omega-like shape). We propose to extend the original work [15] to deal with the full human body (more specifically, pedestrian figures). The main differences from our previous work are related to the shape model generation (different shape models are learned), the initialization procedure (which is bounding-box based), the way the graph is built (considering the full body) and the way the final segmentation is obtained (using a selection scheme), as described next.

The proposed bounding-box based initialization method enables pedestrian segmentation when the usage of face detectors are usually considered impracticable (e.g., when low resolution images are employed and many facial features are missed, making difficult its detection, or when the object of interest is viewed from the back). In addition, the usage of different learned shapes combined with a selection scheme provides an efficient way to deal with a wide variation in shapes and poses, as well as to give a initial guess of the 2D pedestrian's pose/orientation.

A. Shape Model Generation

In this work we use a shape model to guide the computation of the best path in the graph. Next, we describe how the pedestrian shape model is generated as well as how the segmentation is performed.

1) Learning Dataset: The shape model of the pedestrian is generated based on manual annotations associated to the adopted dataset. The exemplars from the walking actions in the HumanEVA dataset [16], used in the work of Bo and Fowlkes [7][1], are defined in the proposed model as training data. It contains a total of 937 manually segmented exemplars from 4 individuals into the 6 body parts, captured by 8 different viewpoints, providing a range of people's shape, image resolution, poses and occlusions problems.

2) Learning the Pedestrian Shape Model: As we are not interested in segmenting the individual body parts, the training images are initially labeled as foreground and background. Then, images are manually grouped into 9 classes c, as follows: c_i (for $i = 1$ to 4) representing pedestrians oriented to the left; c_5 representing the frontal/back view; and c_i (for $i = 6$ to 9) representing pedestrians oriented to the right. Images from the left view classes are flipped in the y axis (vertical), aiming to increase the number of samples and to deal with small angle orientations on the image plane, and added to the respective right view class (and vice versa - generating, at the end, the same shape for each left/right class, but rotated in the vertical axis). Such flipping procedure is done for the frontal/back view class, but in this case the flipped images are added in the same class. In a second stage we compute, for each class, the average person's height μ_{c_i} (related to the height h_p of each person, derived from its bounding-box). For simplicity, the rest of the shape model generation is described considering the c_5 class (the procedure is the same for the others).

The region of each person is represented by a binary image (Fig. 1(a)), limited by a bounding-box, smoothed with a Gaussian filter (to deal with inaccuracies coming from the manual annotation) and resized by a factor $f_l = \mu_{c_5}/h_p$.

The head position $\mathbf{H}_p$ of each person (illustrated in Fig. 1(a) by a red cross) is estimated by projecting the pixels of its binary image on the horizontal axis and retrieving the point with maximum value as x coordinate and its respective projected value as y coordinate (adjusted to consider the origin of the Cartesian plane in the upper left corner of the image). The head position is estimated considering only the upper body region of the binary image (illustrated in Fig. 1(a) by a dotted line), defined experimentally as the top half region of the bounding-box. The estimated head

[1] http://vision.ics.uci.edu/datasets

positions are then used to compute the following reference values (used to align the final shape model to the image under analysis, detailed in Sec. III-B): $\mathbf{H}_\mu$ and $\mathbf{H}_\sigma$ (each one is a 2D vector composed by x and y values, normalized by μ_{c_5}), related to the average head position and standard deviation.

The resized images are then projected into a plane (aligned by $\mathbf{H}_p$), accumulating the value of each pixel and generating the initial shape model $\mathbf{S}_0$, as illustrated in Fig. 1(b). The white dot in the top of Fig. 1(b) represents the reference point $\mathbf{R}_s$, which will be used to align the final shape model to the bounding-box of the detected person in the image under analysis. We defined $\mathbf{R}_{s_y} = \mathbf{H}_{\mu_y} h_s$ (where h_s is the height of $\mathbf{S}_0$, derived from the projected pixels interval), and $\mathbf{R}_{s_x}$ is defined by the average x coordinate of those pixels (with highest value) in $\mathbf{S}_0$ at $y = \mathbf{R}_{s_y}$ location.

The initial shape model $\mathbf{S}_0$ is then thresholded through visual inspection. The thresholded image (Fig. 1(c)) is then used to obtain what we call the "average" pedestrian's body contour $\mathbf{B}_f$, defined by the boundary of the thresholded image, as shown in Fig. 1(d), which is used to create the final shape model $\mathbf{S}_f$ as well as to guide the graph generation. The final shape model $\mathbf{S}_f$ (Fig. 1(f)) is computed using a Gaussian function, as defined on Eq. 1.

$$S_f(x,y) = e^{-\frac{D_t(x,y)^2}{(\tau_1 \mu_c)^2}}, \tag{1}$$

where x, y are the spatial coordinates of each pixel, $\mathbf{D}_t$ is the Distance Transform (Fig. 1(e)) computed using $\mathbf{B}_f$ and the scale factor of the Gaussian is given by $\tau_1 \mu_c$ (set based on experiments[2]). The shape model can be viewed as a prior confidence map on the location of the pedestrian's body contour, and it is combined with image data to obtain the final contour, as explained next. For the sake of illustration, Fig. 1(g-j) shows the generated average contours for the left view classes (or for the right view when flipped in y axis).

B. Pedestrian segmentation

For simplicity, the segmentation procedure is firstly described considering the c_5 class. Let $G = (V, E)$ be the generated graph, consisting of a finite set V of vertices and a set of edges E. The vertices form a grid-like structure, and they are placed along a region where the contour of the person is expected to appear (Fig. 2(a)). The number of the levels of the graph, the length of each level, as well as the number of vertices along the levels is parameterized according to the bounding-box of the detected person, as described next. As the bounding-box does not give us a well defined reference of the person's head, we firstly need to define such point, to further align the shape model to it by $\mathbf{R}_s$, as well as to guide the construction of the grid-like structure of the graph. The bounding-box reference point $\mathbf{R}_b$ is defined as follows:

[2]In all experiments we used $\tau_1 = 0.08$ and $\tau_2 = 0.025$.

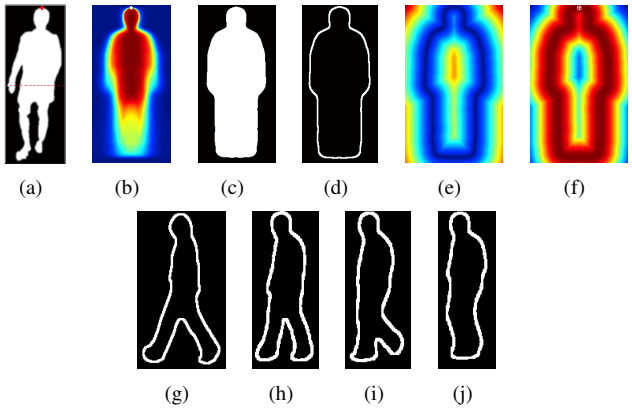


Figure 1. Shape model generation. (a) input training image sample and the estimated head position $\mathbf{H}_p$; (b) initial shape model $\mathbf{S}_0$ and the reference point $\mathbf{R}_s$; (c) thresholded image; (d) estimated "average" pedestrian's boundary $\mathbf{B}_f$; (e) distance transform $\mathbf{D}_t$ of (d); (f) the final shape model $\mathbf{S}_f$; (g-j) obtained average contours for the left view classes.

$\mathbf{R}_{b_y} = h_p \mathbf{H}_{\mu_y}$ and $\mathbf{R}_{b_x} = 0.5 w_p$, where w_p is the bounding-box width. To deal with misalignment we include two other values as its x coordinate, considering the standard deviation of the head positions, computed in the learning stage, defined as follows: $\mathbf{R}_{b_{x'}} = \mathbf{R}_{b_x} - h_p \mathbf{H}_{\sigma_x}$ and $\mathbf{R}_{b_{x''}} = \mathbf{R}_{b_x} + h_p \mathbf{H}_{\sigma_x}$.

Consider the estimated "average" pedestrian's boundary $\mathbf{B}_f$ (Fig. 1(d)), resized according to $f_s = h_p / \mu_{c_5}$, to deal with scale variations and aligned to the reference point of the bounding-box $\mathbf{R}_b$ by $\mathbf{R}_s$ (also resized by f_s). The goal of using in this stage f_s instead of f_l is to adapt the learned shape model to the input image resolution, avoiding the loss of data when the resolution is reduced. Next, the points of this curve ($\mathbf{B}_f$) are discretized from one another by $d = \tau_2 h_p$ pixels. Then, for each two consecutive points of this curve, a level of the graph is generated orthogonal to the line segment that connects these points. Each level, with a length $L = \tau_1 h_p$ pixels, is centered on its respective line segment. The vertices are labeled $V_{m,n}$, where $m = 1, ..., M$ denotes the level of the vertex, and $n = 1, ..., N$ is the position of the vertex in such level. The number of levels M are based on the number of points of the estimated "average" pedestrian's body boundary and the d value, and the number of nodes in each level is set to $N = L/2$. The levels of the entire graph, for a given image, are illustrated by blue lines in Fig. 2(a) and a detailed illustration of it is shown in Fig. 2(b).

The edges of the graph relate to line segments connecting two nodes belonging to adjacent levels. More precisely, each node in a level m can be connected to the $k = 3$ (up to) nearest nodes in the level $m + 1$, as illustrated in Fig. 2(b). The weight $w(e_k)$ of each edge e_k is computed as:

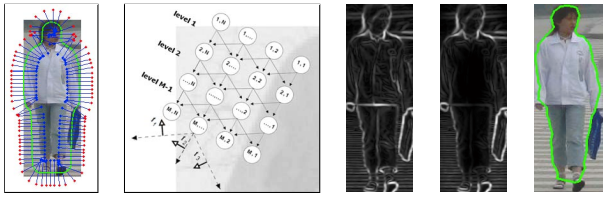
$$w(e_k) = \frac{1}{q_k} \sum_{j=1}^{q_k} E_k(x_j, y_j), \tag{2}$$

where q_k is the number of image pixels in a raster scan

along edge e_k, E_k is the energy function, and (x_j, y_j) are the coordinates of the pixels along such scan. The proposed energy function is composed by several factors: edge, shape mask and angular constraints [15]. The energy map for pixels related to graph edges is given by Eq. 3.

$$E_k(x, y) = |\mathbf{t}_k \cdot \nabla I(x, y)|\mathbf{S}_f(x, y), \tag{3}$$

where $\mathbf{S}_f$ is the shape model, resized according to f_s (as well as $\mathbf{R}_s$) and aligned by $\mathbf{R}_s$ to the pedestrian under analysis by its bounding-box reference point $\mathbf{R}_b$; $\mathbf{t}_k$ is a unit vector orthogonal to the measured graph edge (to prioritize contour with similar orientation as the graph edge under analysis), and $\nabla I(x, y)$ is the discrete gradient image, which can be computed using different approaches, such as: using the Di Zenzo operator (when using color images); or using the Sobel operator (when the luminance component I of the original image is considered), for example. The results illustrated in Section IV were obtained by using the Di Zenzo operator applied to RGB images (excluding those illustrated in Fig. 6, where the Sobel operator was applied in thermal images). The RGB colorspace was chosen because it is one of the most widely used color representation in image processing applications. As the choice of colorspace can be very context dependent, we intend to perform a comparative analysis about its selection, as well as to test different gradient operators as future work.



(a) (b) (c) (d) (e)

Figure 2. (a) Generated graph; (b) illustration of the nodes/edges of the graph; (c-d) energy terms (∇I), without and with the influence of $\mathbf{S}_f$, respectively; (e) segmentation result.

It is important to emphasize that the people's contour is defined by a maximal cost path in the graph, given by a combination of edge and geometric information. Since the graph is acyclic (the estimated contour is connected by its extremities as a post-processing operation), such path can be computed using dynamic programming, as in Dijkstra's algorithm [17]. As the generated graphs can have different number of levels, according to the $\mathbf{B}_f$ curve of each class c_i, the measured energy of each path is defined by the average energy computed along such path.

We observed in some situations that the desired path is not the one with maximum cost. It usually happens when the estimated path is attracted by a very contrasting background region of the image. To minimize such problem we apply a penalization weight to the computed energy (ε_i), related to the best path of each class c_i, as defined by Eq. 4.

$$\varepsilon_i' = \varepsilon_i(1 - \omega), \tag{4}$$

where ω is the Hausdorff Distance [18], computed using the points of the estimated contour and the ones related to its respective (and resized) average shape $\mathbf{B}_f$ (both aligned by its respective reference points and normalized to $\{0, ..., 1\}$). The goal of such procedure is to penalize a path with high energy value, which is very different from its respective average shape class. The procedure described above is computed for each class c_i (for $i = 1$ to 9) and for each bounding-box reference point $\mathbf{R}_b$ shift j (for $j = 1$ to 3). Then, the final contour is defined by the graph path with maximum ε_i' energy.

In case of partial occlusions edge information tends to be weak along the desired contour, but the geometric cue tends to attract the maximal cost path to the desired location. Fig. 3(e) illustrates such situation, where the lower body part of one person was occluded by a bag.

IV. Experimental Results

In this section we illustrate some results of the proposed model, also presenting a quantitative comparison with the state-of-the-art on two public datasets: the Penn-Fudan [19] and the PPSS dataset [8]. In addition, we also illutrate some experimental results to motivate the usage of the proposed model on thermal data [10]. Regarding the comparative analysis, the segmentation accuracy is measured by the intersection/union criterion (as in [4]) of the PASCAL VOC challenge [20].

1) The Penn-Fudan dataset: It contains 170 color images with 345 box/shape-labeled pedestrians from which 169 labels (defined in [7]) compose our test dataset.

The first columns in Table I show the measured accuracy of the proposed model without (ε) and with (ε') the penalization weight (Eq. 4). As we can see in Table I, the penalization weight slightly improved the obtained results.

It is important to highlight that we used a significant smaller number of shape exemplars for training (931 images) when compared to [4] (in which 10946 images were used). In addition, multiple data cues were used in [4], [7], whereas the proposed model is color/texture independent, being useful for nighttime applications, in which such information could be weakened due to illumination conditions, or when infrared/thermal cameras are employed (see Figure 6). On the other hand, the proposed model, which we consider a very simple approach (and with a small set of input parameters), achieved results (in this first experiment) below to the ones obtained by the Deep Decompositional Network (DDN) proposed in the work of Luo et al. [8]. As related by the authors [8], DDN is trained by estimating a set of weight matrices and corresponding biases, which can be very

challenging due to the huge amount of parameters. Thus, considering the values presented in Table I, we consider the obtained results can be comparable to the state-of-the-art.

Table I
RESULTS ON THE PENN-FUDAN DATASET.

	Our model		State-of-the-art			
	ε	ε'	[8]	[4]	[7]	[6]
FG	73.46	75.04	78.4	78.5	73.3	71.6
BG	76.52	77.41	85.0	81.5	81.1	73.8
AVG	74.99	**76.22**	**81.7**	**80.0**	**77.2**	**72.7**

Figure 3(a-e) illustrates some results of the proposed model we consider very promising, whereas Figure 3(f) illustrates a bad obtained results.

(a) 90.6 (b) 90.1 (c) 86.9 (d) 85.8 (e) 84.7 (f) 62.5

Figure 3. Results (with average accuracy) on Penn-Fudan dataset [19].

2) The PPSS dataset: It contains 3.673 images, from 171 videos of different surveillance scenes, where 2.064 images are occluded and 1.609 are not. The ground truth of label maps for all these images is provided. As we are not interested in pedestrian parsing, the body labels were ignored in this experiment (the ground truth is used as background/foreground mask). To make a fair comparison against [8], we performed two experiments on this dataset. In the first experiment, the same evaluation protocol as in [8] was adopted, i.e., the last 71 scenes were used for testing, containing 1.892 images. As we can see in Table II, regarding the first experiment, the achieved results are still slightly below to the ones obtained by [8].

Table II
COMPARISON WITH THE STATE-OF-THE-ART ON PPSS DATASET.

	Our model (ε')		Luo et al. [8]	
Experiment	1	2	1	2
FG	64.85	67.19	71.4	67.56
BG	80.72	83.02	80.0	80.86
AVG	**72.79**	**75.1**	**75.7**	**74.21**

It is important to emphasize that the PPSS testing dataset includes a lot of people riding bicycles and motorcycles (about 344 people), as well as some seated people (about 23), which we consider a big challenge, as their poses usually are very different from an expected pedestrian pose (upright people, standing or walking, as mentioned in the work of Dollar et al. [1]). Figure 4 illustrates some of these images and its respective ground truth data.

To validate our model in the PPSS dataset with only upright people (standing or walking), we conducted a second experiment in which these 367 images of bikers and seated people were ignored. In this second experiment on a subset of the PPSS dataset, the accuracy of the state-of-the-art [8] was slightly improved by the proposed model, as we can see on Table II (the segmentation results of [8] were computed using the implementation provided by the authors[3]).

Figure 4. Undesirable "pedestrian" samples (people riding motorcycles or seated) of the PPSS dataset [8] and respective ground truth data.

Figure 5 illustrates some experimental results on PPSS dataset (some of them with severe occlusion).

Figure 5. Results of our model on PPSS dataset [8].

3) Thermal dataset: Figure 6 illustrates the segmentation results in a few images captured by a thermal camera [10], where color information is not available. In this experiment, the bounding-box of each person were extracted by a person detector [9] directly from the thermal images. The visible images associated to each thermal image are shown on their left side (for the sake of illustration - they were not used in the segmentation stage). In this experiment, the gradient image (used in Eq. 3) was obtained through Sobel operator, considering the luminance component of the respective thermal image.

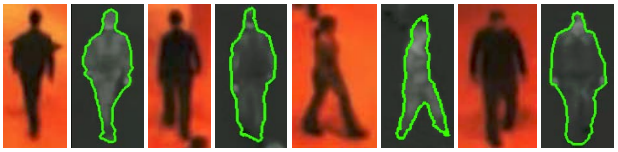


Figure 6. Proposed model applied on thermal images [10].

[3]http://mmlab.ie.cuhk.edu.hk/projects/luoWTiccv2013DDN

4) Computational cost: The average time (in sec. - measured in the first experiment) to obtain the best path for each class was 1.21 (± 0.19), considering an average bounding-box height of 289.79 (± 26.71). The model was implemented using MATLAB (assuming the bounding-box of each person is given; without considering I/O procedures). The hardware used was an HP xw8600 Workstation, with an Intel Xeon processor, Core2 Quad, 2.83GHz and 3Gb of memory.

V. Conclusion

This paper presented a new approach for pedestrian segmentation, captured in a wide range of viewpoints. The proposed model, which is scale invariant, generates a graph around the detected person and the estimated contour is defined by a path in the graph with maximal cost combined with a selection scheme. The computation of the path is guided by a shape model, which was learned from an entirely different dataset. Obtained results indicate the proposed model works well in non trivial images, being comparable to the state-of-the-art. In addition, the proposed model is color/texture independent, which can be useful in nighttime applications (when such information could be weakened due to illumination conditions) or when infrared/thermal cameras are employed. Future work will concentrate on automatic generation of shape model classes.

Acknowledgment

The authors would like to thank Brazilian agencies FAPERGS and CAPES for the financial support.

References

[1] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 2360–2367.

[3] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey, "Stel component analysis: Modeling spatial correlations in image class structure," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2044–2051.

[4] F. Flohr and D. Gavrila, "Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues," in *Proceedings of the British Machine Vision Conference*, 2013.

[5] Y. Li, Z. Zhou, and W. Wu, "Combining shape and appearance for automatic pedestrian segmentation," in *Tools with Artificial Intelligence, 23rd IEEE International Conference on*, Nov 2011, pp. 369–376.

[6] S. M. A. Eslami and C. K. I. Williams, "A generative model for parts-based object segmentation," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 100–107.

[7] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2265–2272.

[8] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 2648–2655.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[10] A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 210 – 221, 2012.

[11] H. Gao, H. Wang, X. Liu, and X. Ma, "High performance pedestrian detector using local segmentation self-similarity in complex scenes," *Pattern Recognittion Image Analysis*, vol. 24, no. 1, pp. 93–101, 2014.

[12] M.-M. Cheng, G.-X. Zhang, N. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Computer Vision and Pattern Recognition, 2011 IEEE Conference on*, June 2011, pp. 409–416.

[13] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut": Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, Aug. 2004.

[14] S. M. A. Eslami, N. Heess, and J. Winn, "The shape boltzmann machine: a strong model of object shape," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 406–413.

[15] J. C. S. J. Junior, C. R. Jung, and S. R. Musse, "Head-shoulder human contour estimation in still images," in *21th IEEE International Conference on Image Processing*, Paris, France, 2014, pp. 278–282.

[16] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, 2010.

[17] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Sec. Ed.* McGraw-Hill Science/Engineering/Math, 2001.

[18] M. Hossain, M. Dewan, K. Ahn, and O. Chae, "A linear time algorithm of computing Hausdorff distance for content-based image analysis," *Circuits, Systems, and Signal Processing*, 2012.

[19] L. Wang, J. Shi, G. Song, and I.-F. Shen, "Object detection combining recognition and segmentation," in *Proceedings of the 8th Asian Conference on Computer Vision*, vol. 1, 2007.

[20] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, 2010.