PedCut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues

Fabian Flohr[1,2]
fabian.flohr@daimler.com

Dariu M. Gavrila[1,2]
www.gavrila.net

[1] Environment Perception Department, Daimler R&D, Ulm, Germany

[2] Intelligent Systems Laboratory, Univ. of Amsterdam, The Netherlands

Person segmentation is a key computer vision problem in a number of application domains, such as image editing, surveillance and intelligent vehicles. This paper presents an iterative, EM-like framework for accurate pedestrian segmentation, combining generative shape models and multiple data cues. It is able to cope with a large variation of pedestrian appearances across cluttered backgrounds. In the E-step, shape priors are introduced in the unary terms of a Conditional Random Field (CRF) formulation, joining other data terms derived from color, texture and disparity cues. In the M-step, the resulting segmentation is used to adapt an Active Shape Model (ASM) [2]. The EM process alternates until the CRF-based segmentation does not appreciably change any more or a maximum number of iterations is reached. Fig. 1 illustrates our framework.

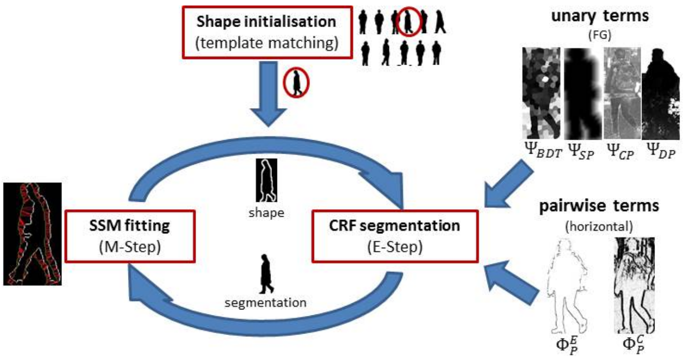

Figure 1: Our EM-like segmentation framework, alternating CRF segmentation (E-step) and SSM fitting (M-step), given shape initialisation.

Shape initialisation Input is an image region of interest (i.e. a bounding box) provided by a pedestrian detector front-end. As ASMs can get stuck in local minima, we obtain our initial shape by matching a set of pedestrian shape exemplars from a training set. We use chamfer matching differentiated by gradient direction (four discretization intervals, not encoding the gradient sign), as in [3]. The best matching shape exemplar is converted to its Statistical Shape Model (SSM) representation (we use several SSMs to model various pose clusters [4], e.g. feet apart/closed); it acts as a shape prior in the following CRF segmentation step.

CRF segmentation Let $\boldsymbol{I}$ and $\boldsymbol{D}$ be the color and disparity values of the image region. We use Semi Global Matching [5] for disparity computation. Furthermore, let $\boldsymbol{S}$ be the superpixel feature vectors of the region.

We specify four unary potentials for the CRF: 1) the sigmoid converted output (Ψ_{BDT}) of a Boosted Decision Tree ensemble trained with dense SIFT and Texton features on the image region, 2) a shape potential (Ψ_{SP}) calculated on a distance transformation obtained from the current shape contour Ω, 3) a GMM-based color potential (Ψ_{CP}) similar to the GrabCut framework [6] based on the current segmentation and 4) a GMM-based disparity potential (Ψ_{DP}) based on the median disparity over the current segmentation.

We further specify two pairwise potentials, which take the form of generalized Potts models [1]: 1) a color-sensitive potential (Φ_P^C), specified such, that it increases the costs of an edge inversely proportional to the color difference in Lab color space of two neighbored pixels i and j, and 2) a contour-sensitive potential (Φ_P^E), which increases the cost inversely proportional to the edge magnitude between pixels i and j, weighted based on the disparity information.

Given the unary and pairwise terms we minimize an energy functional defined on the index set $\mathcal{V}$ with an eight-connected edge neighborhood $\mathcal{E}$, of the following form:

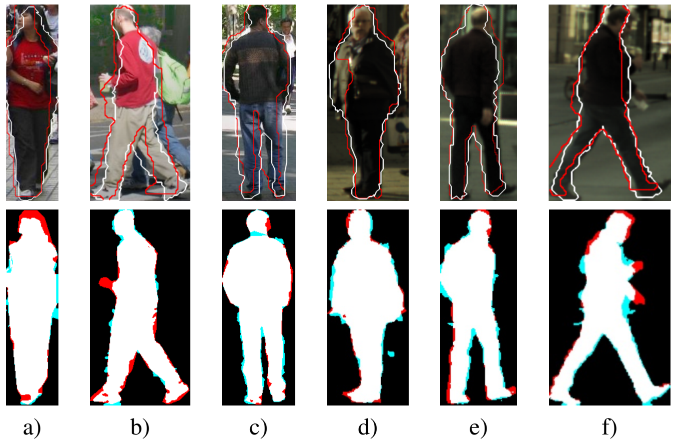

Figure 2: Results after four iterations. First row: input images with initial/final SSM fit (red/white). Second row: correct/missing/excessive segmentation (white/red/cyan). a)-c): Penn-Fudan dataset (BDT+SP+CP); d)-f): Our dataset (BDT+SP+CP+DP)

$$E(\boldsymbol{x},\Omega,\boldsymbol{I},\boldsymbol{D},\boldsymbol{S},\boldsymbol{\omega}) = \sum_{i\in\mathcal{V}} \omega_{BDT}\Psi_{BDT}(x_i,\boldsymbol{S}) + \omega_{SP}\Psi_{SP}(x_i,\Omega) + \omega_{CP}\Psi_{CP}(x_i,\boldsymbol{I})$$
$$+\,\omega_{DP}\Psi_{DP}(x_i,\boldsymbol{D}) + \sum_{i,j\in\mathcal{E}} \omega_P^C\Phi_P^C(x_i,x_j,\boldsymbol{I}) + \omega_P^E\Phi_P^E(x_i,x_j,\boldsymbol{I}.\boldsymbol{D}). \quad (1)$$

Main CRF parameters $\boldsymbol{\omega}$ are the weights for the specified unary and pairwise terms (ω_{BDT}, ω_{SP}, ω_{CP}, ω_{DP}, ω_P^C and ω_P^E). As our pairwise terms stay submodular, we can perform inference with Graph Cut [1] methods.

SSM fitting We use an ASM approach [2] for fitting the SSM model to the obtained CRF segmentation. Point correspondences between SSM and image are given by chamfer matching [3]. As in shape initialisation, we can differentiate chamfer matching by gradient direction. Since we have a binary segmentation, we can here utilize information about the gradient sign to improve matching (i.e. eight discretization intervals for gradient direction).

Results We show the benefit of different cue combinations and the ability of the framework to improve results with each additional cue. On the public Penn-Fudan dataset [7] (Fig. 2 a-c), we outperform the state-of-the-art by more than 5% on foreground accuracy while remaining ahead on background accuracy. Further we provide results on our own pedestrian dataset (Fig. 2 d-f), captured from on-board a vehicle, which includes disparity data. This dataset is made public for non-commercial research purposes to facilitate benchmarking.

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE PAMI*, 23(11):1222–1239, 2001.

[2] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 61(1):38–59, 1995.

[3] D. M. Gavrila. A Bayesian, exemplar-based approach to hierarchical shape matching. *IEEE PAMI*, 29(8):1408–1421, 2007.

[4] J. Giebel and D. M. Gavrila. Multimodal shape tracking with point distribution models. *Pat. Rec.*, pages 1–8, 2002.

[5] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE PAMI*, 30(2):328–341, 2008.

[6] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *Proc. ACM TOG (SIGGRAPH)*, 23(3):309–314, 2004.

[7] L. Wang, J. Shi, G. Song, and I.-F. Shen. Object detection combining recognition and segmentation. In *Proc. ACCV*, pages 189–199. Springer, 2007.