Level Set Person Segmentation and Tracking with Multi-Region Appearance Models and Top-Down Shape Information

Esther Horbert, Konstantinos Rematas, Bastian Leibe {horbert,leibe}@umic.rwth-aachen.de

MMP Mobile Multimedia Processing

RWTH AACHEN UNIVERSITY

ABSTRACT

We address the problem of segmentation-based tracking of multiple articulated persons. We propose two improvements to current level-set tracking formulations. The first is a localized appearance model that uses additional level-sets in order to enforce a hierarchical subdivision of the object shape into multiple connected regions with distinct appearance models. The second is a novel mechanism to include detailed object shape information in the form of a per-pixel figure/ground probability map obtained from an object detection process. Both contributions are seamlessly integrated into the level-set framework. Together, they considerably improve the accuracy of the tracked segmentations.

MOTIVATION

Accurately segment articulated persons in the presence of similar background colors and clutter for:

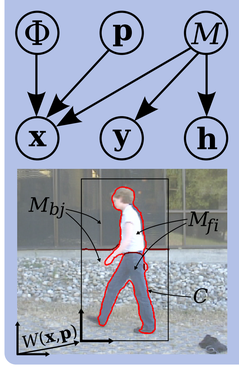
- Articulated tracking
- Better appearance models for tracking
- Video editing

CONTRIBUTIONS

Level set segmentation and tracking with seamless integration of

- Multi-region appearance models
- Detailed class specific information from an object detector

LEVEL SET SEGMENTATION AND TRACKING



$\mathbf{x}$: pixel's coordinates in reference frame, $\mathbf{y}$: pixel's color, $\mathbf{p}$: reference frame position, $\mathbf{h}$: shape model

$\mathbf{\Phi}$: level set embedding function, M: foreground and background regions with appearance model $P(\mathbf{y}|M)$

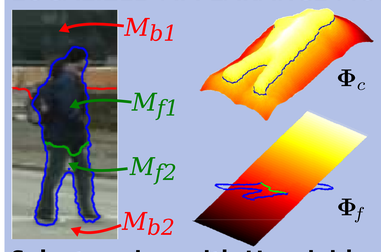
Segmentation: maximize probability of level set function (extension of [1]): $\mathcal{E}(\mathbf{\Phi}) = -log\left(P(\mathbf{\Phi}, \mathbf{p}|\mathbf{x}, \mathbf{y}, \mathbf{h})\right)$

Evolve contour by optimizing energy functional with gradient descent

$$\frac{\partial \mathbf{\Phi}_m}{\partial t} = -\frac{\partial \mathcal{E}(\mathbf{\Phi}_m)}{\partial \mathbf{\Phi}_m} = \underbrace{\frac{\frac{\partial}{\partial \mathbf{\Phi}_m} P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h})}{P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h})}}_{\text{data term}} + \underbrace{\frac{1}{\sigma^2}\left[\nabla^2(\mathbf{\Phi}_m) - \mathrm{div}\left(\frac{\nabla \mathbf{\Phi}_m}{|\nabla \mathbf{\Phi}_m|}\right)\right] + \lambda \delta_\epsilon(\mathbf{\Phi}_m)\mathrm{div}\left(\frac{\nabla \mathbf{\Phi}_m}{|\nabla \mathbf{\Phi}_m|}\right)}_{\text{smoothness term}} \qquad m \in \{c, f, b\}$$

Tracking with rigid registration step: optimize position $\mathbf{p}$ while contour $\mathbf{\Phi}$ stays constant

LOCALIZED APPEARANCE MODELS



4 regions
$k \in \{f1, f2, b1, b2\}$

3 LS embedding functions
$\mathbf{\Phi}_c$: person's contour
$\mathbf{\Phi}_f$: foreground subregions
$\mathbf{\Phi}_b$: background subregions

Select region with Heaviside step function

$$H_m = H_\epsilon(\mathbf{\Phi}_m(\mathbf{x})), \quad \tilde{H}_m = 1 - H_\epsilon(\mathbf{\Phi}_m(\mathbf{x})), \quad m \in \{c, f, b\}$$

$$\eta_{f1} = \Sigma_{i=1}^{N} H_c H_f \quad \#\text{ pixels in region } f1 \quad (f2, b1, b2 \text{ similarly})$$

$$P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h}) = H_c H_f P_{f1} + H_c \tilde{H}_f P_{f2} + \tilde{H}_c H_b P_{b1} + \tilde{H}_c \tilde{H}_b P_{b2}$$

where $$P_k = \frac{P(\mathbf{y}|M_k)P(M_k|\mathbf{h})}{\sum_l \eta_l P(\mathbf{y}|M_l)}, \quad k, l \in \{f1, f2, b1, b2\}$$

The data term is thus $$\frac{\delta_m\left(H_f P_{f1} + \tilde{H}_f P_{f2} - H_b P_{b1} - \tilde{H}_b P_{b2}\right)}{P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h})}$$

DETECTION BASED TOP-DOWN SEGMENTATION

Detector

- Match patches to learned vocabulary with random forest
- Leaf nodes vote for location (as in ISM)
- Collect votes in Hough voting space
- Maxima correspond to object hypotheses
- Back-projection of votes yields top-down segmentation

Object-specific figure-ground probability for every pixel

$$P(M_f|\mathbf{h}) = \frac{1}{z}\sum_{\mathbf{X}_i(\mathbf{x})} \frac{1}{|\mathbf{X}_i|} \sum_{v_j \in votes(\mathbf{X}_i)} w_{v_j} Seg(v_j)$$

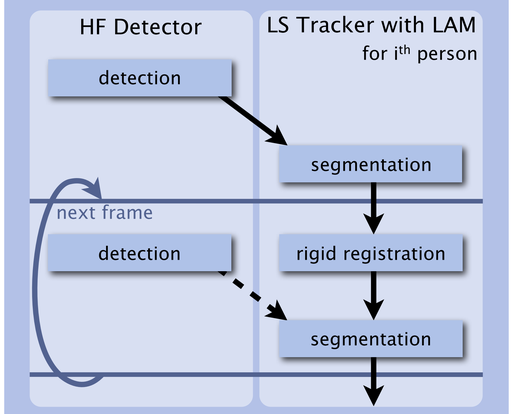
$$P(M_b|\mathbf{h}) = \frac{1}{z}\sum_{\mathbf{X}_i(\mathbf{x})} \frac{1}{|\mathbf{X}_i|} \sum_{v_j \in votes(\mathbf{X}_i)} w_{v_j} (1 - Seg(v_j))$$

$$z = \sum_{\mathbf{X}_i(\mathbf{x})} \sum_{v_j \in votes(\mathbf{X}_i)} w_{v_j}$$

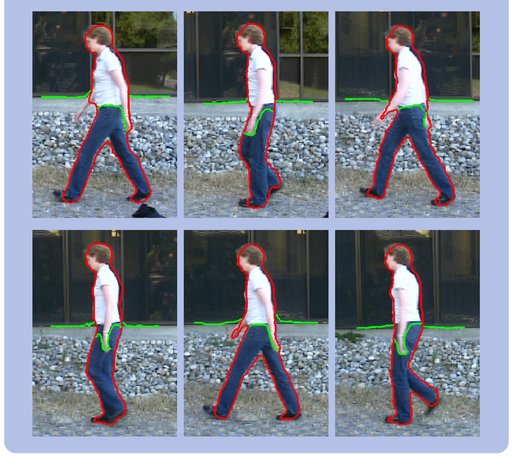
A segmentation is obtained with $$\frac{\theta P(M_f|\mathbf{h})}{\theta P(M_f|\mathbf{h}) + P(M_b|\mathbf{h})} \geq 0.5$$

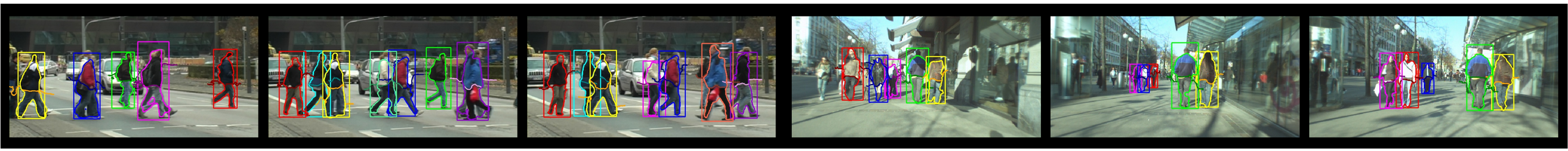
see [2] for details on the detector

COMBINED MODEL



QUALITATIVE RESULTS





EVALUATION OF COMPONENTS

	recall	IOU	prec
BR box init	57.5%	51.5%	83.1%
LS box init	60.0%	55.6%	88.4%
LS hf init	64.1%	58.6%	87.3%
LAM box init	64.5%	58.1%	85.5%
LAM hf init	65.1%	59.8%	88.0%
LS+HF	64.5%	61.4%	**92.7%**
LAM+HF	**68.8%**	**65.0%**	**92.1%**
HF	65.7%	61.3%	90.1%

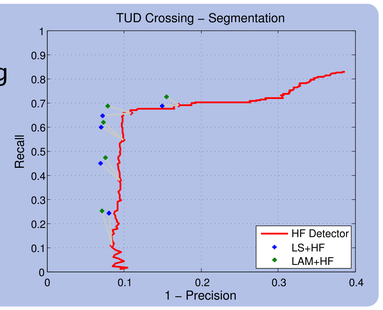
LS Level set tracker
LAM LS tracker with localized appearance models
HF Hough Forest detector

➔ *Both the localized appearance models and the figure/ground probability maps contribute to the improved segmentation results*

SEGMENTATION PERFORMANCE

Segmentation results for TUD Crossing of our complete model starting from different detector thresholds in comparison to the detector alone

➔ *Our localized appearance models improve the performance on top of the integration with probabilistic shape models*



CONCLUSION

Improved segmentation performance through

- Hierarchical subdivision of the segmented regions for more distinctive localized appearance models
- Integration of Hough Forest ISM top-down segmentations as probabilistic shape models

ANNOTATIONS AVAILABLE
http://www.mmp.rwth-aachen.de/people/horbert

[1] C. Bibby, I. Reid: Robust Real-Time Visual Tracking using Pixel-Wise Posteriors. ECCV (2008)
[2] K. Rematas, B. Leibe: Efficient Object Detection and Segmentation with a Cascaded Hough Forest ISM. ICCV, CORP Workshop (2011)