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

Esther Horbert, Konstantinos Rematas, Bastian Leibe

UMIC Research Centre, RWTH Aachen University
{horbert,leibe}@umic.rwth-aachen.de

Abstract

In this paper, 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. We experimentally evaluate our proposed approach on two challenging sequences and demonstrate its good performance in practice.

1. Introduction

Level Sets [7, 10] have gained increasing popularity for many segmentation and tracking tasks due to their computational efficiency [20, 3] and their flexibility with respect to topological changes of the contour (which sets them apart from active shapes [5, 17]). In this paper, we consider their use for segmentation-based tracking of articulated objects, as shown in Fig. 1.

Several approaches have been proposed that are targeted at level-set tracking of deformable objects [8, 9, 3]. In particular, the recent formulation by Bibby and Reid has demonstrated robust tracking performance in a variety of real-world scenarios [3], including multi-object tracking from surveillance videos [4] and in busy street scenes [21]. However, while the results obtained there are satisfactory from a tracking perspective, they typically sacrifice segmentation accuracy in order to provide robust tracking under strong articulations. Detailed segmentations are however often required for later processing stages, *e.g.* for video editing [1] or to provide detailed input for body pose analysis and articulated tracking (*e.g.* [26, 27, 14]). Improvements in the way shape and appearance can be represented to cope with such challenging tasks are therefore of considerable interest for many applications.

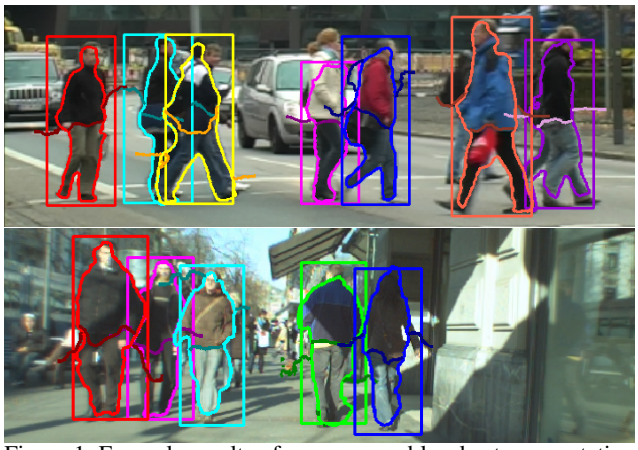

Figure 1. Example results of our proposed level-set segmentation and tracking approach. The combination of localized appearance models and top-down shape information considerably improves the accuracy of the tracked contours.

Surprisingly, the appearance models used in level set tracking have so far been relatively weak, mainly encoding global object properties such as color, texture, motion, or 3D depth [10]. Indeed, it is difficult to use more powerful localized appearance models (such as, *e.g.*, HOG features [11]) in a level-set formalism, since the level-set framework aims at optimally grouping regions whose pixels have *similar* feature signatures. This makes it difficult for level-set approaches to reliably segment and track multi-colored, articulated objects such as pedestrians in front of complex, cluttered backgrounds.

Similarly, shape information in the context of level-set formulations has primarily been considered in the form of static [19, 28, 23] or dynamic [8, 9] priors. Those try to constrain the embedding function using the statistics of a set of training shapes, either by performing the optimization in a subspace [19, 28] or by bringing in the shape constraints on the variational level [23]. The problem with such priors

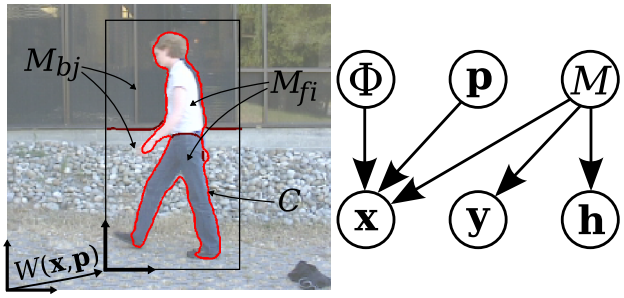

Figure 2. (Left): The tracked person with the contour $\mathbf{C}$, the foreground and background regions M_{fi}, M_{bj} and the warp $W(\mathbf{x}, \mathbf{p})$ representing the position in the image. (Right): The generative model used in our approach, treating the image as bag of pixels.

$\mathbf{x}$	Pixel's coordinates inside reference frame
$\mathbf{y}$	Pixel's color
$\mathbf{p}$	Reference frame position
$\mathbf{h}$	Shape model
$W(\mathbf{x}, \mathbf{p})$	Warp with parameters $\mathbf{p}$
M_{f1}, M_{f2}	Foreground regions
M_{b1}, M_{b2}	Background regions
$P(\mathbf{y}\|M_k)$	Appearance models
$\mathbf{\Phi}$	Level set embedding function
$\{\mathbf{\Phi}_c, \mathbf{\Phi}_f, \mathbf{\Phi}_b\}$	Embeddings for person and fore/background
$\mathbf{C}_k$	Contour represented by the zero level set
$H_\epsilon(z)$	Smoothed Heaviside step function
$\delta_\epsilon(z)$	Smoothed Dirac delta function

Table 1. Notation used in this paper

for tracking articulated body shapes is that they are both too weak and too strong. They are too weak, because they need to encompass the entire intra-category variability, in addition to the variability in articulation. And they are too strong, because the priors tend to continue tracking even in the absence of an object to track (as is also visible in the results of [8]).

In this paper, we propose two improvements that address both of the above issues. (1) We propose to use top-down segmentation information fed back from object detection [15, 18, 16] in order to incorporate category-specific shape information that is conditioned on the current image content. This idea takes advantage of the flexible part layout employed by state-of-the-art detectors, which can adapt to different shapes, while still enforcing global consistency. In addition, our formulation does not place hard constraints on the object shape, but brings in soft information in the form of a per-pixel figure/ground probability map that can be directly integrated into the object model. (2) We propose a mechanism to include localized appearance information into the level-set framework. Our approach uses additional level sets to enforce a hierarchical subdivision of the object shape into multiple connected regions with distinct appearance models. The advantage of this formulation is that the separating contour itself can be optimized and tracked to provide the best fit for each frame. This allows our approach to automatically adapt to multi-colored pedestrian clothing and deliver good tracking results in challenging scenarios, such as the ones shown in Fig. 1.

Both contributions are seamlessly integrated into the level-set tracking framework. Together, they considerably improve the accuracy of the tracked segmentations, as our experimental results will demonstrate.

Related Work. Even though level sets do not provide a globally optimal solution (in contrast to, *e.g.*, [25]), they are still widely used in practice because of their efficiency and other advantages [10]. Consequently, many approaches have been proposed that use them for segmentation or tracking. In the following, we only focus on those approaches most closely related to our contributions.

[29] and [6] propose multiphase level set formulations composed of N distinct level set functions to represent complex boundaries between up to 2^N regions. Our appearance modeling approach is similar to this idea in that we also consider an overlay of several level set functions, but our motivation is to create a hierarchical subdivision capturing the detailed appearance of *e.g.* pedestrian clothing, while still preserving the outer silhouette as a single contour.

Schmaltz *et al*. [24] propose a Localized Mixture Model (LMM) for object segmentation and 3D model-based tracking with non-uniform backgrounds. Their approach partitions the fore- and background into several subregions that are modeled with their own PDFs. In principle, such an appearance model could also be used here. However, the main issue in this case is how to localize the foreground subregions with respect to the moving, articulated person. For this step, [24] apply a model-based tracking approach that employs a 3D person model with limb-specific appearance distributions, which is not available for our task. Instead, our approach estimates and tracks the subregions automatically, while staying inside the level-set framework.

Niebles *et al*. [22] combine top-down shape priors and bottom up appearance and optical flow based segmentation for person segmentation in YouTube videos. In contrast to our approach, their method performs a joint optimization on the entire video sequence and is thus not suitable for online analysis.

The rest of the paper is structured as follows. The next section introduces the basic level set tracking framework our approach operates in. Section 3 presents our localized appearance model and Section 4 adds the top-down shape information. Section 5 discusses how all components are combined. Finally, Section 6 presents experimental results.

2. Level Set Segmentation and Tracking

We use a probabilistic level-set framework to perform a segmentation of the target object and track it through the following frames, similar to [3]. Fig. 2 shows the generative model that is the foundation of our proposed framework

and Table 1 summarizes the notation. The tracked object is represented by its contour $\mathbf{C}$ (represented with level sets $\mathbf{\Phi}$) and its position $\mathbf{p}$ in the image. It consists of pixels at coordinates $\mathbf{x}$ with color $\mathbf{y}$. Foreground and background regions M are distinguished by appearance models consisting of color histograms and we additionally incorporate a class specific shape model $\mathbf{h}$. Thus, given an initialization for $\mathbf{x}, \mathbf{y}, \mathbf{h}$, and M, the task is to infer shape $\mathbf{\Phi}$ and position $\mathbf{p}$. The joint distribution for one pixel given by the model is

$$P(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i, \mathbf{\Phi}, \mathbf{p}, M) = P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M)P(\mathbf{y}_i|M)P(\mathbf{h}_i|M)P(M)P(\mathbf{\Phi})P(\mathbf{p}) \quad (1)$$

Conditioning on $\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i$ and marginalizing over M yields

$$P(\mathbf{\Phi}, \mathbf{p}|\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i) = \frac{1}{P(\mathbf{x}_i)} \sum_k \Big\{ P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M_k) \frac{P(\mathbf{y}_i|M_k)P(M_k)}{\sum_l P(\mathbf{y}_i|M_l)P(M_l)} P(M_k|\mathbf{h}_i) \Big\} P(\mathbf{\Phi})P(\mathbf{p}) \quad (2)$$

where the M_k denote the different (foreground and background) regions. We use $P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i, \mathbf{h}_i)$ as shorthand notation for the sum in (2) and fuse the pixels i in a logarithmic opinion pool, which yields:

$$P(\mathbf{\Phi}, \mathbf{p}|\mathbf{x}, \mathbf{y}, \mathbf{h}) = \prod_{i=1}^{N} P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i, \mathbf{h}_i)P(\mathbf{\Phi})P(\mathbf{p}) \quad (3)$$

We use the term $P(\mathbf{\Phi})$ to specify some desired internal properties of the contour: First, a geometric prior that rewards a signed distance function (eliminating the need for periodic re-initializations [20, 3]) and second, a prior for the length of the contour [10], rewarding a smoother contour.

$$P(\mathbf{\Phi}) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(|\nabla\mathbf{\Phi}|-1)^2}{2\sigma^2}\Big) \exp\Big(-\lambda|\nabla H_\epsilon(\mathbf{\Phi})|\Big) \quad (4)$$

where σ and λ are the weights of the priors. Maximizing the posterior is equivalent to minimizing its negative logarithm:

$$\begin{aligned}\mathcal{E}(\mathbf{\Phi}) &= -\log(P(\mathbf{\Phi}, \mathbf{p}|\mathbf{x}, \mathbf{y}, \mathbf{h})) \\ &\propto -\sum_{i=1}^{N} \Big\{ \log(P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i, \mathbf{h}_i)) - \frac{(|\nabla\mathbf{\Phi}|-1)^2}{2\sigma^2} \\ &\quad -\lambda|\nabla H_\epsilon(\mathbf{\Phi})| \Big\} + N\log\Big(\frac{1}{\sigma\sqrt{2\pi}}\Big) + \log(P(\mathbf{p})) \quad (5)\end{aligned}$$

This equation gives us the probability for the desired values $\mathbf{\Phi}$ and $\mathbf{p}$. In order to optimize both, we first optimize the shape and keep the position constant (the *segmentation* step), then optimize the position while keeping the shape constant (resulting in a *rigid registration* step).

Segmentation. For segmentation, we optimize (5) w.r.t $\mathbf{\Phi}$ while keeping $\mathbf{p}$ constant, using the Euler-Lagrange equation which minimizes $\mathcal{E}(\mathbf{\Phi})$ [10]:

$$\begin{aligned}\frac{\partial\mathbf{\Phi}_k}{\partial t} &= -\frac{\partial\mathcal{E}(\mathbf{\Phi})}{\partial\mathbf{\Phi}_k} = \frac{\frac{\partial}{\partial\mathbf{\Phi}_k}P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h})}{P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h})} + \\ &\frac{1}{\sigma^2}\left[\nabla^2(\mathbf{\Phi}_k) - \mathrm{div}\left(\frac{\nabla\mathbf{\Phi}_k}{|\nabla\mathbf{\Phi}_k|}\right)\right] + \lambda\delta_\epsilon(\mathbf{\Phi}_k)\mathrm{div}\left(\frac{\nabla\mathbf{\Phi}_k}{|\nabla\mathbf{\Phi}_k|}\right)\end{aligned} \quad (6)$$

The gradient flow (6) consists of three terms: The first term penalizes the deviation from the appearance models and from the shape model. The second term penalizes the deviation from a signed distance function and thus makes the embedding function numerically stable without the need for periodic re-initializations. The third term penalizes the length of the contour, thus smoothing it. This is particularly useful in cluttered scenes, since the background can contain many scattered pixels with foreground appearance, which would otherwise lead to a very uneven contour.

We will specify $P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h})$ later; at this point it is sufficient to know that the contour evolves such that it encloses pixels with foreground appearance and high figure probability.

Rigid Registration. Having obtained the target object's shape, we track it through the following frames by performing a rigid registration. The new position of $\mathbf{\Phi}$ is described with a warp $\mathbf{p}$, which can be any transformation that forms a group. [3] use translation+scale+rotation, but since we want to track pedestrians, we only use translation+scale here. For this, we introduce the warp $\mathbf{W}(\mathbf{x}_i, \Delta\mathbf{p})$ with parameters $\mathbf{p}$. $P(\mathbf{p})$ is dropped here and is handled with drift correction, as in [3]:

$$\log(P(\mathbf{\Phi},\mathbf{p}|\mathbf{x},\mathbf{y},\mathbf{h})) \propto \sum_{i=1}^{N} \log\Big\{P(\mathbf{W}(\mathbf{x}_i,\Delta\mathbf{p})|\mathbf{\Phi},\mathbf{p},\mathbf{y}_i,\mathbf{h}_i)\Big\} \quad (7)$$

We maximize this equation w.r.t $\mathbf{p}$:

$$\mathbf{p} = \arg\max_{\mathbf{p}} \Big\{ \sum_{i=1}^{N} \log P(\mathbf{W}(\mathbf{x}_i, \Delta\mathbf{p})|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i, \mathbf{h}_i) \Big\} \quad (8)$$

Optimization is performed using a second-order Newton optimization scheme, as in [4]. With the short-hand notation $P(...) = P(\mathbf{W}(\mathbf{x}_i, \Delta\mathbf{p})|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i, \mathbf{h}_i)$, we obtain

$$\Delta\mathbf{p} = \Big[\sum_{i=1}^{N} \frac{\Big(\frac{\partial P(...)}{\partial\mathbf{p}}\Big)^2}{P(...)}\Big]^{-1} \sum_{i=1}^{N} \frac{\partial P(...)}{\partial\mathbf{p}} \quad (9)$$

Keeping the shape $\mathbf{\Phi}$ constant, the position $\mathbf{p}$ is optimized by incrementally warping it with $\Delta\mathbf{p}$.

Summary. The contour of the foreground object is described by the zero level set of a level set embedding function $\mathbf{\Phi}_c$. Starting from some initialization, the contour is evolved to maximize its probability given the image, the

Figure 3. Results obtained with a level set tracker as in [3]: (a) With 1 segmentation iteration after rigid registration, the contours lose highly articulated parts such as the legs. (b) With more iterations, the results become very inaccurate due to cluttered background containing the same colors as the foreground regions.

learned appearance models, and the shape model. The appearance models (in our case L*a*b color histograms) are rebuilt in each of the n_1 iterations. In the following frame, the new position of the shape is registered and afterwards the contour is adapted by performing n_2 segmentation iterations. In this case, the appearance models are not rebuilt, but only slightly adapted for greater robustness, as in [3]. In the following two sections we will now describe the used appearance models and shape models in more detail.

3. Localized Appearance Models

[3] use a simple appearance model consisting of two color histograms, one for the foreground region and one for the background region. While this has proven to be robust to viewpoint changes and changes occurring in the background due to a moving camera, we found that this model is too weak if the goal is detailed segmentation in addition to robust tracking. Persons are highly articulated and thus their shapes undergo significant changes between frames, which cannot be traced with only one segmentation iteration (*c.f.* Fig. 3(a)). The shape can of course be adapted by performing more segmentation iterations to account for the shape changes. However, a weak appearance model is not able to handle cluttered background that may contain the same colors as the foreground region, which itself already contains a number of different colors. In this case, foreground and background are hard to distinguish by two color histograms, leading to a dramatically decreased robustness to similar background regions. In consequence, the contour can easily bleed out into background regions, while some foreground regions may be lost (*c.f.* Fig. 3(b)). Consider the case shown in Fig. 4(a). Here, the tracked person's trousers have the same color as the background region behind his head. The two regions do not adjoin and yet it is difficult to distinguish them with a single appearance models.

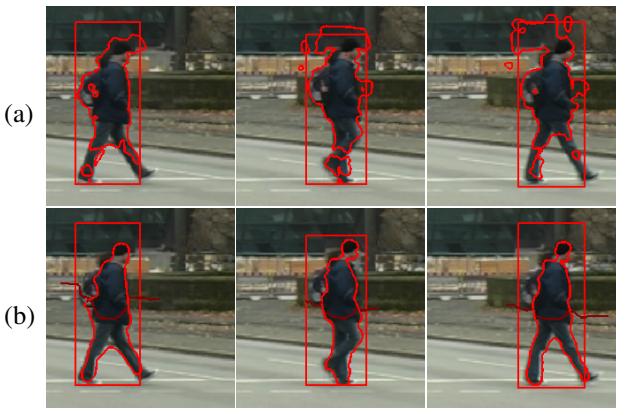

Figure 4. (a): Failure case of the level set tracker as in [3] with more segmentation iterations: the background of the head has the same color as the trousers resulting in the contour to drift into the background and away from the legs. (b): Results obtained with our Localized Appearance Models: foreground and background regions are distinguished much more accurately.

Hence, we propose to incorporate spatial information into the appearance models. For pedestrians, this is very intuitive for the lower and upper body with different clothing. The background can also be separated into different regions, for example at the horizon. One option would thus be to track lower and upper body independently with two different contours. However, this strategy does not make use of the information that the two parts belong together and would lead to problems, since the two regions could overlap or diverge. Another option would be a fixed grid subdivision, where every grid cell has its own appearance model. The problem is that narrow-band techniques then become difficult to use, since the contour can evolve into previously empty cells. Moreover, this could have negative effects for small regions that are part of a larger region and only partially lie in one cell. Instead, we propose to use a subdivision that is specific to the target object by optimizing the subdivision with the contour.

To this end, we use two additional level set embedding functions $\mathbf{\Phi}_f$ and $\mathbf{\Phi}_b$ for two additional contours that separate foreground and background into two regions, respectively. Consequently, we use four color histograms to describe these four regions. The person's shape is still described by $\mathbf{\Phi}_c$, but $\mathbf{\Phi}_f$ and $\mathbf{\Phi}_b$ are used to determine which appearance model to use. Fig. 5(a) shows the four regions with the four color histograms. Fig. 5(b) depicts the level set embedding function $\mathbf{\Phi}_c$ and the person's contour at the zero level. $\mathbf{\Phi}_f$, whose zero level describes the foreground's additional separating contour, is shown in Fig. 5(c).

In order to formally define this model, we now specify the term $P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h})$ from Sec. 2. Using a smoothed Heaviside step function H_ϵ to select the respective regions and a smoothed Dirac function δ_ϵ to select the contours

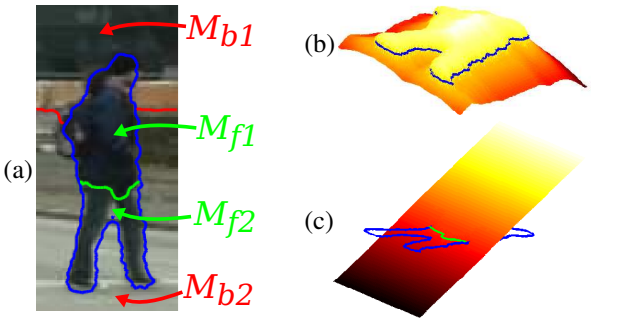

Figure 5. (a) Four appearance models for four regions. (b) 3D visualization of level set embedding function $\mathbf{\Phi}_c$, contour in blue, note the legs at the lower left. (c) 3D visualization of $\mathbf{\Phi}_f$, foreground division line in green.

$$\begin{aligned} H_c &= H_\epsilon(\mathbf{\Phi}_c(\mathbf{x}_i)), \quad \tilde{H}_c = 1 - H_\epsilon(\mathbf{\Phi}_c(\mathbf{x}_i)) \\ H_f &= H_\epsilon(\mathbf{\Phi}_f(\mathbf{x}_i)), \quad \tilde{H}_f = 1 - H_\epsilon(\mathbf{\Phi}_f(\mathbf{x}_i)) \\ H_b &= H_\epsilon(\mathbf{\Phi}_b(\mathbf{x}_i)), \quad \tilde{H}_b = 1 - H_\epsilon(\mathbf{\Phi}_b(\mathbf{x}_i)) \\ \delta_c &= \delta_\epsilon(\mathbf{\Phi}_c(\mathbf{x}_i)), \quad \delta_f = \delta_\epsilon(\mathbf{\Phi}_f(\mathbf{x}_i)), \quad \delta_b = \delta_\epsilon(\mathbf{\Phi}_b(\mathbf{x}_i)) \end{aligned} \tag{10}$$

we can write the number of pixels in the four regions as follows:

$$\begin{aligned} N = \sum \eta_k,\ \eta_{f1} = \sum_{i=1}^{N} H_c H_f,\ \eta_{f2} = \sum_{i=1}^{N} H_c \tilde{H}_f, \\ \eta_{b1} = \sum_{i=1}^{N} \tilde{H}_c H_b,\ \eta_{b2} = \sum_{i=1}^{N} \tilde{H}_c \tilde{H}_b\ . \end{aligned} \tag{11}$$

This way, we can express the terms

$$P(M_k) = \frac{\eta_k}{N}, \quad k \in \{f1, f2, b1, b2\}, \tag{12}$$

$$\begin{aligned} P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M_{f1}) &= \frac{H_c H_f}{\eta_{f1}}, \quad P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M_{f2}) = \frac{H_c \tilde{H}_f}{\eta_{f2}}, \\ P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M_{b1}) &= \frac{\tilde{H}_c H_b}{\eta_{b1}}, \quad P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, M_{b2}) = \frac{\tilde{H}_c \tilde{H}_b}{\eta_{b2}}. \end{aligned}$$

Having specified all these terms, we obtain

$$\begin{aligned} &P(\mathbf{x}_i|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}_i, \mathbf{h}_i) = \\ &\quad H_c H_f P_{f1} + H_c \tilde{H}_f P_{f2} + \tilde{H}_c H_b P_{b1} + \tilde{H}_c \tilde{H}_b P_{b2} \end{aligned} \tag{13}$$

where

$$P_k = \frac{P(\mathbf{y}_i|M_k) P(M_k|\mathbf{h}_i)}{\sum_l \eta_l P(\mathbf{y}_i|M_l)}, \quad k, l \in \{f1, f2, b1, b2\}. \tag{14}$$

The segmentation gradient flow is thus:

$$\begin{aligned} \frac{\partial P(\mathbf{\Phi}, \mathbf{p}|\mathbf{x}, \mathbf{y}, \mathbf{h})}{\partial \mathbf{\Phi}_k} &= \frac{\delta_k \left(H_f P_{f1} + \tilde{H}_f P_{f2} - H_b P_{b1} - \tilde{H}_b P_{b2} \right)}{P(\mathbf{x}|\mathbf{\Phi}, \mathbf{p}, \mathbf{y}, \mathbf{h})} \\ + \frac{1}{\sigma^2} \left[\nabla^2(\mathbf{\Phi}_k) - \mathrm{div}\left(\frac{\nabla \mathbf{\Phi}_k}{|\nabla \mathbf{\Phi}_k|} \right) \right] &+ \lambda \delta_\epsilon(\mathbf{\Phi}_k) \mathrm{div}\left(\frac{\nabla \mathbf{\Phi}_k}{|\nabla \mathbf{\Phi}_k|} \right) \end{aligned} \tag{15}$$

We initialize the additional contours with horizontal lines at 50% of the bounding box height for the foreground and 60% for the background. We evolve the two additional level set functions interleaved with the original level set function $\mathbf{\Phi}_c$. In this way, the four appearance models are optimized at the same time, which leads to more robust and accurate segmentation results (*c.f.* Fig. 4(b)).

$\mathbf{\Phi}_f$ and $\mathbf{\Phi}_b$'s contours are only optimized where they are "visible", *i.e.* inside or outside the person's contour, respectively. This means that only these visible parts are meaningful and when the person's contour grows or shrinks, it reveals parts of $\mathbf{\Phi}_f$ and $\mathbf{\Phi}_b$ that were meaningless before. However, as described in Section 2, the level set embedding functions are kept in a numerically stable form and the contour is smoothed slightly, thus the newly appearing regions are handled implicitly.

This hierarchical subdivision could easily be extended to even more than four regions by dividing foreground and/or background into multiple regions as, *e.g.*, in [29] (n level set embedding functions for 2^n regions), each with its own color histogram.

We can now also specify the rigid registration:

$$\begin{aligned} \frac{\partial P(...)}{\partial \mathbf{p}} &= \left(\mathbf{J}_c H_f + H_c \mathbf{J}_f\right) P_{f1} + \left(\mathbf{J}_c \tilde{H}_f - H_c \mathbf{J}_f\right) P_{f2} - \mathbf{J}_c H_b P_{b1} - \mathbf{J}_c \tilde{H}_b P_{b2} \\ &= \mathbf{J}_c \left(H_f P_{f1} + \tilde{H}_f P_{f2} - H_b P_{b1} - \tilde{H}_b P_{b2}\right) + \mathbf{J}_f \left(H_c P_{f1} - H_c P_{f2}\right) \end{aligned} \tag{16}$$

with the Jacobians of the warp

$$\mathbf{J}_k = \frac{\partial H_k}{\partial \mathbf{\Phi}_k} \frac{\partial \mathbf{\Phi}_k}{\partial \mathbf{x}} \frac{\partial \mathbf{W}}{\partial \Delta \mathbf{p}} = \delta_\epsilon(\mathbf{\Phi}_k(\mathbf{x}_i)) \nabla \mathbf{\Phi}_k(\mathbf{x}_i) \frac{\partial \mathbf{W}}{\partial \Delta \mathbf{p}}\ . \tag{17}$$

We assume that the background does not move with the foreground, thus $H_\epsilon(\Phi_b(\mathbf{x}_i))$ is constant w.r.t the derivation $\frac{\partial}{\partial \mathbf{p}}$. Eq. (16) illustrates that both the person's contour and the division line of the foreground contribute to the warp, whereas the division line of the background does not contribute: $\mathbf{J}_k$ contain the factor $\delta_\epsilon(\mathbf{\Phi}_k)$, the Dirac delta function of the respective level set function, which is only greater than zero in a narrow band around the contour (with width 2ϵ), and $\mathbf{J}_b$ does not occur in (17).

Even with more exact appearance models, the initial segmentation still poses problems, since it is not possible to distinguish foreground regions from same-colored background regions. Therefore we incorporate class specific information to be able to distinguish persons from similar looking background, as described in the next section.

4. Detection-Based Top-Down Segmentation

In order to achieve robust tracking performance, we want to make use of the information that we track objects of a certain category (*e.g.*, pedestrians). This is also the motivation behind work on category-specific shape priors [19, 28, 23, 8]. However, such priors do not take into account image-specific information and have difficulties modeling the dynamic shape of strongly articulated objects. Instead, we propose to use top-down segmentation information fed back from object detection.

For this, we build upon class-specific Hough Forest detectors [13], which have been shown to reach state-of-the-art detection performance on several benchmark datasets. This approach uses a random forest structure to efficiently match densely sampled input patches to a discriminatively trained visual vocabulary represented by the trees' leaf nodes. It then takes up the voting idea of Implicit Shape Models (ISM) [18] in order to let each activated leaf node vote for possible locations of the object center using a learned spatial occurrence distribution. The votes are collected in a Hough voting space. Local maxima in this space correspond to object hypotheses, which are passed to a final non-maximum suppression stage. The advantage of Hough Forests is that they can be evaluated very fast, so that they are suitable for processing densely sampled image patches.

As shown in [18], the votes corresponding to a local maximum in the Hough space can be backprojected to the image in order to infer top-down segmentation information. Intuitively, this step derives a local figure-ground label for the patch $\mathbf{X} = \{\mathbf{x}_k\}$ conditioned on the patch appearance and the detected object location. As each vote v_j originating from patch $\mathbf{X}$ hypothesizes a relative object location, it can be augmented with a figure-ground label (of the size of the patch and learned from training data) $Seg(v_j)$ consistent with that location. When a Hough-space maximum is selected and its constituent votes are backprojected, the effect is that only consistent segmentation labels survive. The resulting patch segmentation label can then be inferred by summing the backprojected figure-ground labels, weighted by the weight of the corresponding vote w_{v_j}. The figure-ground probability of a pixel $\mathbf{x}$ is obtained by averaging over all patches $\mathbf{X}_i$ containing this pixel:

$$P(M_f|\mathbf{h}) = \frac{1}{z} \sum_{\mathbf{X}_i(\mathbf{x})} \frac{1}{|\mathbf{X}_i|} \sum_{v_j \in votes(\mathbf{X}_i)} w_{v_j} Seg(v_j) \qquad (18)$$

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

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

Fig. 6 visualizes the *figure* and *ground* probability maps for an example detection.

The resulting procedure provides an object-specific figure and ground probability for every pixel and integrates naturally into the level-set formulation. Moreover we use the top down segmentation as initialization for the level set segmentation. The foreground region is given by all pixels $\mathbf{x}$ with

$$\frac{\theta P(M_f|\mathbf{h})}{\theta P(M_f|\mathbf{h}) + P(M_b|\mathbf{h})} \geq 0.5\,. \qquad (19)$$

We initialize $\mathbf{\Phi}_c$ with a signed distance function of the obtained contour. The factor θ can be used to shrink or enlarge the obtained contour. We use $\theta = 0.9$ to increase precision.

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

Figure 6. Visualization of the top-down segmentation information used in our approach: (a) $p(M_f|\mathbf{h})$ and (b) $p(M_b|\mathbf{h})$ probability maps. (c) Resulting contour just based on this information. (d)(e) Two more examples for the detector's segmentations.

Note that eq. (14) contains $P(M_k|\mathbf{h})$ for four regions, but the detector only provides probabilities for two regions: foreground and background. However, (13) selects one region by use of H, H_f and H_b. Therefore the following always holds: Either $P(M_f|\mathbf{h}) = P(M_{f1}|\mathbf{h})$ or $P(M_f|\mathbf{h}) = P(M_{f2}|\mathbf{h})$ and ditto for $P(M_b|\mathbf{h})$

5. Combined Model

We demonstrate our approach with a simple pedestrian tracker. The combined model can be summarized as follows: The Hough Forest detector is used to initialize tracks for newly appearing persons. These are first segmented using four color histograms in our Localized Appearance Model framework and using figure and ground probabilities provided by the detector by performing $n_1 = 100$ segmentation iterations. In subsequent frames, the persons are tracked and the new object positions are determined with a simple Kalman Filter using the new detection as observation. The shape is adapted to the new image and figure-ground probabilities by performing $n_2 = 100$ segmentation iterations. If there is no detection for a tracked person in a particular frame, $P(M|\mathbf{h})$ does not exist. In this case, we consider the generative model without h, which in practice means $P(M|\mathbf{h})$ is omitted in the equation. Tracklets are discarded if there are no new detections for a specified number of frames.

All persons are tracked independently, their level set embedding functions do not interact directly. We do however infer an approximate depth ordering by comparing the y-positions of the bounding boxes. This approach does not need any further information (as, *e.g.*, a ground plane or depth maps), but it is restricted to simple occlusion cases. It cannot propagate identities through occlusions, but one could apply a high-level tracker as in [21] to achieve this.

6. Experimental Results

We evaluate our approach on two challenging and very different datasets. The TUD CROSSING dataset [2] consists of 201 frames with walking persons mainly seen from the side and was taken with a stationary camera on a cloudy

	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%

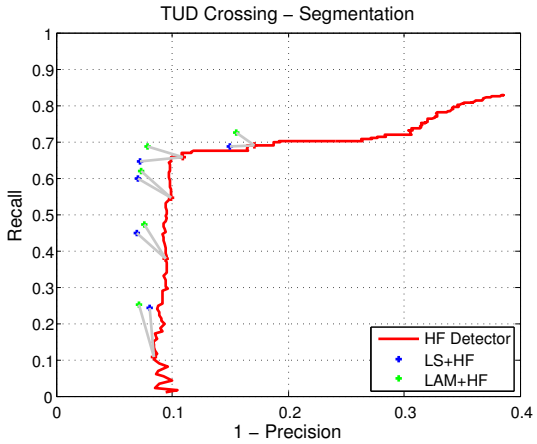

Figure 7. (Left): Segmentation performance for TUD Crossing for different combinations of our approaches. BR: level set tracker as in [3], LS: the level set tracker with 2 appearance models, LAM: with our localized appearance models, HF: Hough Forest detector. Box init: initialization with detector bounding box, hf init: initialization with segmentation, +HF: initialization and optimization with shape probability maps. (Right): Segmentation performance for TUD Crossing for different initializations of LS and LAM.

day. We annotated pixel-wise segmentations of the persons in 21 frames (every 10th frame). The ETH SUNNY DAY sequence [12] was taken with a moving camera on a sunny day and contains persons walking towards the camera. This dataset consists of 354 frames and also provides a ground plane estimate which we use to reject detections that are inconsistent with the ground plane.

Qualitative results. Fig. 8 shows our results for both test sequences. In comparison with the baseline results in Fig. 3, it becomes clear that our segmentation results are more accurate.

Segmentation Performance.

In Fig. 7 (left), we compare the segmentation performance of the Hough Forest detector top-down segmentation, the level set tracker as in [3], and different combinations of our proposed approaches. It can be seen that all parts contribute to improve the segmentation results. The full model without the localized appearance models or without the Hough Forest top-down segmentation both do not reach the performance of the full model, proving that both are necessary to achieve this improvement.

Fig. 7 (right) shows how the segmentation performance of the raw Hough Forest detector (red line) is improved through the integration of the level set tracker with the probabilistic shape models and our localized appearance models (green crosses) for different detector thresholds (as indicated by the gray lines). As can be seen, the localized appearance models improve performance on top of the integration with the probabilistic shape models (blue crosses).

7. Conclusion

In conclusion, we have presented a level set framework for segmentation and tracking of multiple articulated persons. We have shown how to use a hierarchical subdivision of the segmented regions, that is optimized within the framework, to use more distinctive localized appearance models. Furthermore we have demonstrated how to incorporate detailed class specific information obtained from an object detector. As our experiments have shown both of those approaches contribute to an increased segmentation performance.

Acknowledgments This project has been funded, in parts, by the EU project EUROPA (ICT-2008-231888), the cluster of excellence UMIC (DFG EXC 89) and the RWTH Seed Fund.

References

[1] A. Agarwala, A. Hertzmann, D. Salesin, and S. Seitz. Keyframe-Based Tracking for Rotoscoping and Animation. In *SIGGRAPH*, 2004.

[2] M. Andriluka, S. Roth, and B. Schiele. People Tracking-by-Detection and People Detection-by-Tracking. In *CVPR*, 2008.

[3] C. Bibby and I. Reid. Robust Real-Time Visual Tracking using Pixel-Wise Posteriors. In *ECCV*, 2008.

[4] C. Bibby and I. Reid. Real-time Tracking of Multiple Occluding Objects using Level Sets. In *CVPR*, 2010.

[5] A. Blake and M. Isard. *Active Contours*. Springer, London, 1998.

[6] T. Brox and J. Weickert. Level Set Based Image Segmentation with Multiple Regions. In *DAGM*, 2004.

[7] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic Active Contours. *IJCV*, 22(1):61–79, 1997.

[8] D. Cremers. Dynamical Statistical Priors for Level Set Based Tracking. *PAMI*, 28:1262–1273, 2006.

[9] D. Cremers. Nonlinear Dynamical Shape Priors for Level Set Segmentation. *J. Sci. Comput.*, 35:132–143, 2008.

[10] D. Cremers, M. Rousson, and R. Deriche. A Review of Statistical Approaches to Level Set Segmentation Integrating Color, Texture, Motion and Shape. *IJCV*, 72:195–215, 2007.

[11] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

Figure 8. Qualitative results for our full model for sequences TUD CROSSING (top) and ETH SUNNY DAY (bottom).

[12] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. Robust Multi-Person Tracking from a Mobile Platform. *PAMI*, 31(10):1831–1846, 2009.

[13] J. Gall and V. Lempitsky. Class-Specific Hough Forests for Object Detection. In *CVPR*, 2009.

[14] S. Gammeter, A. Ess, T. Jaeggli, K. Schindler, B. Leibe, and L. Van Gool. Articulated Multi-Body Tracking under Egomotion. In *ECCV*, 2008.

[15] M. Kumar, P. Torr, and A. Zisserman. OBJ CUT. In *CVPR*, 2005.

[16] D. Larlus, J. Verbeek, and F. Jurie. Category Level Object Segmentation by Combining Bag-of-Words Models and Markov Random Fields. In *CVPR*, 2008.

[17] S. Lefevre and N. Vincent. Real Time Multiple Object Tracking Based on Active Contours. In *ICIAR*, 2004.

[18] B. Leibe, A. Leonardis, and B. Schiele. Robust Object Detection with Interleaved Categorization and Segmentation. *IJCV*, 77(1-3):259–289, 2008.

[19] M. Leventon, W. Grimson, and O. Faugeras. Statistical Shape Influence in Geodesic Active Contours. In *CVPR*, 2000.

[20] C. Li, C. Xu, C. Gui, and M. Fox. Level Set Evolution without Re-initialization: A New Variational Formulation. In *CVPR*, 2005.

[21] D. Mitzel, E. Horbert, A. Ess, and B. Leibe. Multi-Person Tracking with Sparse Detection and Continuous Segmentation. In *ECCV*, 2010.

[22] J. C. Niebles, B. Han, and L. Fei-Fei. Efficient Extraction of Human Motion Volumes by Tracking. In *CVPR*, 2010.

[23] M. Rousson and N. Paragios. Shape Priors for Level Set Representations. In *ECCV*, 2002.

[24] C. Schmaltz, B. Rosenhahn, T. Brox, and J. Weickert. Localised Mixture Models in Region-based Tracking. In *DAGM*, 2009.

[25] T. Schoenemann and D. Cremers. Matching Non-rigidly Deformable Shapes Across Images: A Globally Optimal Solution. In *CVPR*, 2008.

[26] H. Sidenbladh and M. J. Black. Learning the statistics of people in images and video. *IJCV*, 54(1-3):183–209, 2003.

[27] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *CVPR*, 2005.

[28] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, E. Grimson, and A. Willsky. Model-Based Curve Evolution Technique for Image Segmentation. In *CVPR*, 2001.

[29] L. Vese and T. Chan. A Multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model. *IJCV*, 50(3):271–293, 2002.