Coupled Region-Edge Shape Priors for Simultaneous Localization and Figure-ground Segmentation

Cheng Chen and Guoliang Fan

School of Electrical and Computer Engineering
Oklahoma State University, Stillwater, OK 74078, USA
Tel/fax: +1 (405) 744-1547/+1 (405) 744-9198
Email: guoliang.fan@okstate.edu

Abstract

We propose a new algorithm for simultaneous localization and figure-ground segmentation where coupled region-edge shape priors are involved with two different but complementary roles. We resort to a segmentation-based hypothesis-and-test paradigm to solve the problem, where the region prior is used to form a segmentation and the edge prior is used to evaluate the validity of the formed segmentation. Our fundamental assumption is that the optimal shape-constrained segmentation that maximizes the agreement with the edge prior occurs at the correctly hypothesized location. Essentially, the proposed algorithm addresses a mid-level vision issue that aims at producing a map image for part detection can be further used for high-level vision tasks. Our experiments demonstrated that this algorithm offers promising results in terms of both localization and segmentation.

Key words: figure-ground segmentation, shape priors, segmentation, localization, watersheds, online learning, kernel-based color modeling

1. Introduction

Segmentation is an important and long-standing research topic in the fields of image analysis and computer vision, and it can be done at different levels. At low-level vision, it is called image segmentation that is to group pixels into regions of homogeneous properties based on various low-level region-based cues (e.g., intensity, color, or texture) and/or edge-based cues (e.g., boundaries or local gradients). Combining both region-based and edge-based cues has led to significant successes for image segmentation due
to their complementary nature [28]. At mid-level and or high-level vision, segmentation is usually referred to as *figure/ground segmentation* that is to partition an image into foreground and background regions [45], where object-specific priors are usually involved, such as a shape prior. Most current shape constrained segmentation methods, such as [1, 13, 23, 3, 8], require manual initialization of the object configuration (position and orientation). As a separate but related topic, object localization is usually discussed outside the context of segmentation. Our research goal is to integrate localization and figure-ground segmentation into one unified framework where the two tasks can be accomplished simultaneously in a synergistic way.

In this paper, the issue of simultaneous localization and figure/ground segmentation is formulated as a Bayesian estimation problem where we search for the optimal configuration and segmentation of an object of interest (OOI) in the sense of maximum *a posteriori* (MAP). Different from some recent techniques where figure/ground segmentation is usually optimized in a spatially implicit way, our objective function is directly defined and optimized in the 2-D spatial space and provides spatially explicit indication of the existence of an OOI. In particular, we resort to a segmentation-based *hypothesis-and-test paradigm* where the coupled region-edge shape priors are involved with two different but complementary roles. Specifically, the region-based shape prior is used to form a segmentation (given a configuration hypothesis), while the edge-based shape prior is used to evaluate the validity of the formed segmentation (in terms of the similarity and smoothness of the boundary). It is believed that *a correct location hypothesis will encourage a valid shape-constrained segmentation while a valid segmentation will enhance the confidence of the location hypothesis*. This makes the proposed algorithm a suitable tool for mid-level vision computation in two ways. First, the prior knowledge about object configuration can be directly used to prune the search space, such as in the case of video tracking, where the object configuration at the previous frame provides useful contextual information for the present frame. Second, the algorithm outputs a *map* image that indicates the likelihood of an OOI at each pixel location in an image.

Additionally, we propose two techniques that ensures the efficiency and

---

1 According to [44], mid-level vision refers to the intermediate-level processing, situated between the analysis of the image (low-level vision) and the recognition of specific objects and events (high-level vision). For example, in the context of part-based pictorial structure model [7], mid-level vision is referred to as part detection.
effectiveness of the proposed algorithm. Specifically, at the hypothesis stage where the region-based shape prior is used, a new semi-parametric kernel-based color model learning method is proposed that can efficiently learn the figure/ground color models on-the-fly at each hypothesized location and support effective figure/ground segmentation. At the test stage where the edge-based shape prior is used, we develop a mixed edge-based evaluation criterion that measures both the similarity and smoothness of the formed boundary and is helpful to reject false positives with rugged boundaries for an OOI with a smooth boundary, as shown in Fig. 6. Our study is focused on mid-level vision, where intermediate results can be obtained that can support various high-level vision tasks. As a case study, the proposed method is examined in the context of body part detection that has many applications for human detection and tracking as well as pose recognition and localization.

2. Related works

Broadly speaking, recent works on figure/ground segmentation and localization can be classified into three categories, i.e., the bottom-up dominant approaches, the top-down dominant ones, and the combined bottom-up/top-down ones. As a bottom-up dominant approach, [36] provided a set of bottom-up parsing rules to segment human body parts guided by a parse tree. [27] used the contour, shape, shading, and focus cues to find the body parts by searching the optimal segmentation from all possible combinations of super-pixel segments according to a scale constraint defined by a rectangular-shaped bounding box. [43] proposed a shape-based object recognition and image segmentation algorithm where a shape prior is represented in a multi-scale curvature form. Target objects are identified and segmented by grouping over-segmented image regions in a probabilistic way that is influenced by the image information and the shape similarity constraint. Generally speaking, the bottom-up dominant approaches do not depend on a well defined object model, so it is robust to shape variability due to different views or poses. The other side of the same coin is that false positives or negatives may occur because of the limitation of using low-level features only. In contrast, the top-down dominant approaches rely on how good an object model matches the OOI in an image. Borenstein and Ullman proposed a fragment-based object representation in [2] that is to cover as closely as possible the images of different OOIs from a given class using a set of primitive shapes. This representation was used for combined object recognition and segmenta-
tion in [22] where a probabilistic segmentation map can be computed as the output of top-down inference. As noted in [2], bottom-up cues could be used to refine the boundary of a segmented OOI. This kind of extension naturally leads to the combined approaches.

A combined approach usually involves bottom-up features as well as an informative object representation that can be encoded in a global template-like view [26] or a set of fragments [45]. The template-like representation has global shape information that well suits figure/ground segmentation. In order to accommodate more shape variability, various deformable template models have been developed recently [17]. The fragment-based object representation introduced in [2] was used for combined top-down/bottom-up segmentation in [45], where Yu and Shi proposed an integration model to integrate bottom-up pixel grouping and top-down patch matching. It was shown that incorporating bottom-up constraints improves the boundary smoothness of the segmented OOI compared with top-down dominant methods and reduces the false positives/negatives compared with bottom-up dominant approaches.

The combined approaches could be further classified into two classes according to how the OOI is localized. The methods of the first class require manual initialization, such as the active contour model based approaches [19], which is widely used in medical image analysis. The methods of the second class can obtain segmentation and localization simultaneously. Usually, these approaches formulate the two tasks by defining one optimization problem where both low-level features and top-down priors are integrated into an objective (energy) function, e.g., [38, 32, 39]. There are two ways to optimize the energy function. The first way is to optimize it in a spatially implicit space via statistical modeling and inference, such as Conditional Random Field (CRF) [38] [32], or Monte Carlo Markov Chain (MCMC) methods [39]. The second way is to optimize it in a spatially explicit space, such as the hypothesis-and-test approach proposed in [29] that searches through the 2D space to find the global solution. This kind of optimization will facilitate the incorporation of spatial priors and provide an intermediate and spatially sensible outputs for high-level vision tasks.

As a combined approach, our method is inspired by prior research. First, we adopt a segmentation-based hypothesis-and-test paradigm that is similar in spirit with [29] where region-based segmentation is involved for object detection and recognition. While segmentation here is not only the approach but also the goal where coupled region-edge shape priors are involved. Sec-
ond, we use the super-pixel-based image representation. Unlike [33, 27, 36] where Normalized-cut is used, and we adopt the watershed transform to create super-pixels with well defined boundaries that are essential for edge-based evaluation. Third, we involve edge-based segmentation evaluation in the 2-D spatial space, unlike the one in [27] where a bounding box is used for segmentation optimization in a spatially implicit space. Our algorithm outputs a spatially sensible map image at the part-level that can be further used for various high-level vision tasks at the whole-level.

3. Overview of the approach

Our fundamental assumption is that the optimal shape-constrained segmentation that maximizes the agreement with the edge-based shape prior occurs at the correctly hypothesized location. The algorithm overview is pre-
sented in Figure 1. The hypothesis-and-test paradigm is adopted that con-
sists of two phases for simultaneous localization and segmentation: hypothesis
generation and hypothesis test. The former one generates a hypothesis of ob-
ject configuration and creates a corresponding segmentation; and the latter
one evaluates the formed segmentation by comparing it boundary with the
edge-based shape prior. In practice, we can use a slide-window method to
scan through the whole image directly. At each pixel position, the hypothesis-
and-test approach is employed to produce a score that shows the confidence
of the existence of an OOI in that pixel. Then a map image can be obtained
from this full-search that can be considered as a mid-level vision output and
used for various high-level vision tasks. It is worth noting the map image
here is different from the confidence map involved in some segmentation lit-
erature, e.g., [40], where the pixel value represents the probability that that
pixel belongs to an OOI. In our map image, only localization information is
encoded in each pixel value that indicates how likely an OOI exists at that
pixel location.

4. Proposed Algorithm

The flowchart of the proposed algorithm is depicted in Fig. 2. There are
four major steps involved that will be discussed in details in the following.

4.1. Watershed-based Super-pixels

There are several commonly used algorithms for super-pixel generation,
such as Normalized-cut [33], watershed [42] and mean-shift [6]. Specifically,
we choose the watershed transform due to its many “biologically plausible”
properties [34]. Moreover, it is fast, local, and has the potential for parallel
processing. However, the severe over-segmentation problem is the main con-
cern of using the watershed method. Many studies showed that this problem
can be largely mitigated by some preprocessing techniques, such as geodesic
reconstruction [24].

Given an input image $G$, the immersion-based watershed algorithm [42]
and geodesic reconstruction preprocessing [41] are used to obtain $Z$ watershed
cells $G = \{C_i | i = 1, 2, \ldots, Z\}$. Each watershed cell $C_i$ consist of its pixel
members $C_i = \{p_1^{(i)}, p_2^{(i)}, \ldots, p_{\eta_i}^{(i)}\}$, where $\eta_i$ is the number of pixels in the cell.
For each watershed cell $C_i$, we also record its edge pixels by $\Gamma(C_i)$ that will
be used for edge-based evaluation. Moreover, we use a 3-D Gaussian model
$\mathcal{N}(x | \mu_i, \Sigma_i)$ to represent the color distribution in the $L*a*b$ color space
Figure 2: The algorithm flowchart.
for pixels in cell $C_i$. $(\mu_i, \Sigma_i)$ are estimated simply by a maximum-likelihood estimator (MLE) that will be used to online learn the color models for figure-ground segmentation.

4.2. Offline Learning of Shape Priors

Inspired by the MetaMorphs model in [15], we develop an implicit shape model for each part where both region-based and edge-based shape priors are holistically represented. We use the shape histogram to represent the shape prior, in which the shape prior is embedded implicitly into an “image” [23]. Given a set of manually aligned and segmented OOIs defined in a window $\Omega$, the shape histogram $SH(p)$ can be obtained by adding and these binary image windows followed by appropriate normalization, i.e., $SH(p) \in [0, 1]$ and $p \in \Omega$ is a pixel location in the window $\Omega$ where the shape prior is defined. $SH(p)$ reflects the the probability that pixel $p \in \Omega$ belongs to the object, and $1 - SH(p)$ indicates the the probability that the pixel $p$ belongs to the background. Given a threshold $\varepsilon$ (say 0.5), an average object boundary $\mathcal{M}$ can be extracted from $SH(p)$ by a level-set like method, as shown in Fig. 3.

\[ \mathcal{M} = \{ p | SH(p) = \varepsilon \}. \] (1)

Figure 3: The top row shows the training images without alignment. The bottom row show the aligned training images, the learned shape prior, and the extracted edge-based shape prior.

Therefore, $\mathcal{M}$ defines two regions in $\Omega$, namely the object region $\mathcal{R}_M$.
enclosed by \( \mathcal{M} \) and the background region \( \Omega \setminus \mathcal{R}_M \), as defined below:

\[
SH(p) = \begin{cases} 
\varepsilon, & \text{if } p \in \mathcal{M}; \\
> \varepsilon, & \text{if } p \in \mathcal{R}_M; \\
< \varepsilon, & \text{if } p \in \Omega \setminus \mathcal{R}_M.
\end{cases}
\] (2)

Given \( SH(p) \), pixels in \( \mathcal{R}_M \) more likely belong to the foreground, and those in \( \Omega \setminus \mathcal{R}_M \) the background. Therefore, \( \mathcal{M} \) can be used as an edge-based shape prior. Such coupled shape representation by \( SH(p) \) and \( \mathcal{M} \) facilitates the interface between region-based segmentation (bottom-up) and edge-based evaluation (top-down).

### 4.3. Hypothesis step: region-based segmentation

Given the shape prior \( SP(p) \) \( p \in \Omega \) for an OOI where \( \Omega \) is a rectangular window, we can use \( \Omega \) as a sliding-window to scan through the whole image to examine the existence of the OOI at each location. For a hypothesized location, we use \( SP(p) \) to induce a local figure-ground segmentation that is composed by some watershed cells covered by \( \Omega \). This segmentation will be used to validate the existence of the object at that location. In order to take advantage of watershed cells and their built-in color models, we propose a new *semi-parametric* kernel-based model learning techniques to online learn the figure/ground color models from the watershed cells directly. We treat the Gaussian model learned from a watershed cell as a kernel center, and learn the figure/ground color models as follows,

\[
\hat{f}_{ob}(x) = \sum_{i=1, \mathcal{C}_i \cap \Omega \neq \emptyset} \alpha_i K_i(x),
\]

\[
\hat{f}_{bg}(x) = \sum_{i=1, \mathcal{C}_i \cap \Omega \neq \emptyset} \beta_i K_i(x),
\]

(3)

where \( x \) is a color vector; \( \mathcal{C}_i \) is one of \( Z \) watershed cells that overlap with window \( \Omega \); \( K_i(x) = \mathcal{N}(x | \mu_i, \Sigma_i) \) is the color model associated with \( \mathcal{C}_i \); \( \alpha_i \) and \( \beta_i \) denote the contribution of cell \( \mathcal{C}_i \) to the object and background respectively that can be calculated from \( SP(p) \) \( p \in \Omega \) and the overlapping watershed cells as

\[
\alpha_i = \frac{1}{T} \sum_{p \in (\mathcal{C}_i \cap \Omega)} SP(p),
\]

(4)
\[
\beta_i = \frac{1}{T} \sum_{p \in (C_i \cap \Omega)} (1 - SP(p)),
\]

where \(T\) is the size of shape prior window \(\Omega\). Based on the figure/ground color models, we can use the maximum a posteriori (MAP) criterion to identify the watershed cells that belong to the object. Let \(\tau_i\) be the class label for \(C_i\):

\[
\tau_i = \begin{cases} 
1 \text{ (object),} & \alpha_i \hat{f}_{ob}(x = \mu_i) > \beta_i \hat{f}_{bg}(x = \mu_i); \\
0 \text{ (background),} & \alpha_i \hat{f}_{ob}(x = \mu_i) < \beta_i \hat{f}_{bg}(x = \mu_i). 
\end{cases}
\]

where we use the average color \(\mu_i\) of cell \(C_i\) as well as the background/foreground color models to decide its class label. Therefore, we can obtain the corresponding segmentation for a position hypothesis as, \(X = \{ \bigcup C_i | \tau_i = 1 \}\). Different from the one in [29] where the shape prior is used once for online figure-ground color model learning, here we use the region-based shape prior \(SP(p)\) twice (they are combined in implementation). The first time is for the online color model learning as defined in (3), and the second time is for MAP-based segmentation as defined in (6). Considering the false negative is more detrimental than the false positive at mid-level vision, we encourage more object-like segmentations by fully incorporating the region-base shape prior into the segmentation process. This may lead to some false positives due to the double usage of the region-based shape prior. However, the later edge-based evaluation will mitigate this problem.

4.4. Test step: edge-based evaluation

After segmentation \(X\) is formed, we evaluate it according the edge prior \(M\). Let \(\Gamma(X)\) to be the boundary of \(X\), we compare \(\Gamma(X)\) with \(M\) in terms of shape similarity and boundary smoothness. The score of \(X\) with respect to its compliance with \(M\), i.e., \(\rho_M(X)\) is given by,

\[
\rho_M(X) = \exp(-d_{chamfer}(\Gamma(X), M)) + \zeta(1 - S(\Gamma(X), M)),
\]

where the first term is the chamfer distance indicating the shape similarity; the second term measures the boundary smoothness; and \(\zeta\) balances the relative importance between the two terms. It is expected that a valid segmentation should have a smooth boundary that matches with \(M\) well. The first term is sensitive to the transition, rotation and scale. This is desired for rejecting false hypotheses. The second term aims to reject false segmentations with rugged boundaries for the OOI with smooth boundary.
4.4.1. Boundary similarity

In order to eliminate the effect of outliers, we use a modified Chamfer distance [37],

$$d_{chamfer}(U, V) = \frac{1}{n} \sum_{u_i \in U} \max_{v_j \in V} \min \| u_i - v_j \|, \eta, \quad (8)$$

where $\eta$ is a factor controlling the tolerance of mismatching. Furthermore, Eqn. (8) can be efficiently computed using the distance transform (DT) [12].

Given two sets of edge points $U = \{u_i\}$ and $V = \{v_i\}$ in a window $\Omega$, the distance transform $d_V(p)$ specifies the distance from each pixel $p \in \Omega$ to the nearest pixel $v_i \in V$ (as shown in Fig. 4). Therefore the chamfer distance based shape similarity between $U$ and $V$ can be calculated by

$$d_{chamfer}(U, V) = \frac{1}{\#(U)} \sum_{u_i \in U} d_V(u_i), \quad (9)$$

where $\#(U)$ denotes the number of pixels in $U$, and $d_{chamfer}(U, V)$ the average distance between $U$ and $V$.

4.4.2. Boundary smoothness

As shown in Fig. 6, assume that $\Gamma(X)$ touches $n$ cells $\{C_1, \ldots, C_n\}$, and we define $H_i = \{h_{1}^{i}, \ldots, h_{n_i}^{i}\}$ to be the set of $n_i$ boundary pixels shared between $C_i$ and $\Gamma(X)$. Let $\phi_M(p) : \mathbb{R}^2 \rightarrow \mathbb{R}$ be the signed Euclidian distance transform that is “+” or “-” for $p$ inside or outside $M$, respectively. The maximum and minimum distances from $H_i$ to $M$ are obtained by

![Figure 4: The computation of the Chamfer distance via the distance transform. (a) and (b) show two shapes for matching, and (c) shows the distance transform between them.](image-url)
Figure 5: The computation of edge smoothness based on the signed Euclidean distance transform of the edge prior \( \mathcal{M} \), i.e., \( \phi_{\mathcal{M}}(p) \). \( X \) is a segmentation and \( \Gamma(X) \) is the boundary of \( X \) that touches several watershed cells. \( H_i \) is the set of edge pixels shared by watershed cell \( C_i \) and \( \Gamma(X) \). Specifically, \( H_2 \) has the high parallelness (good smoothness), while \( H_3 \) has low parallelness (bad smoothness).

\[
d_{\text{max}}^{(i)} = \max(\phi_{\mathcal{M}}(h_1^{(i)}), ..., \phi_{\mathcal{M}}(h_n^{(i)})), \quad \text{and} \quad d_{\text{min}}^{(i)} = \min(\phi_{\mathcal{M}}(h_1^{(i)}), ..., \phi_{\mathcal{M}}(h_n^{(i)})),
\]

respectively. The degree of parallelness between \( \Gamma(X) \) and \( H_i \) is defined as

\[
S_{\mathcal{M}}(H_i) = \frac{d_{\text{max}}^{(i)} - d_{\text{min}}^{(i)}}{n_i}.
\]

(10)

When \( H_i \) is parallel to \( \mathcal{M} \) (e.g., \( H_2 \) in Fig. 6(b)), \( S_{\mathcal{M}}(H_i) \cong 0 \), indicating good local smoothness. When \( H_i \) is perpendicular to \( \mathcal{M} \) (e.g., \( H_3 \) in Fig. 6(b)), \( S_{\mathcal{M}}(H_i) \cong 1 \), indicating poor local smoothness. In general, the smaller the value, the more parallel between \( H_i \) and \( \Gamma(X) \). Therefore, we define the overall smoothness of \( \Gamma(X) \) as

\[
S(\Gamma(X), \mathcal{M}) = \frac{1}{n} \sum_{i=1}^{n} S_{\mathcal{M}}(H_i).
\]

(11)

If a full search is involved, the score function (7) will return a map image that records the existence possibility of the OOI at every pixel location. The larger the value, the more likely there is an OOI. It is worth noting that at each position hypothesis, the shape prior is hypothesized with different angles around the mean orientation, and we use the winner-take-all strategy to generate the map image. The optimal angle at each location is also recorded. Some examples of edge-based segmentation evaluation is shown in Fig. 6.

Essentially, this computation is at mid-level vision, and it could support high-level vision by incorporating high-level knowledge. We focus on simultaneous localization and segmentation of a non-articulated object with a
relatively well defined shape, such as human body parts. In [4], we proposed a hybrid body representation for integrated pose recognition, localization and segmentation, where this algorithm is used as the inference engine at mid-level vision to locate and segment the body parts.

5. Experiments

A set of experiments were conducted to validate the proposed algorithm. The algorithm was programmed in C++, and the test platform is a PC with Pentium-IV 3.2GHz CPU and 1GB RAM. Our experiments were based on the CMU Mobo database [14], which contains image sequences of 25 individuals walking on a treadmill. Each image is resized to $240 \times 320$ pixels. In particular, we are interested in simultaneous localization and segmentation of six body parts, i.e., (the head, torso, left-arm, right-arm, left-leg, and right-leg) as our OOI’s, each of which is defined in a window of $61 \times 61$ pixels $^2$. For each OOI, 200 manually segmented images from six individuals were used for learning the coupled region-edge shape priors, and all training images share similar poses (i.e., recoil/contact) and the same side-view. The reason of using one pose is because that the shape of some body parts (legs and arms) may deform under different poses, and that of using the recoil/contact poses is due to the fact that they have the least occlusion problem. How to handle shape deformation and occlusion is beyond our scope. The coupled region-edge shape priors of six body parts are shown in Fig. 7.

$^2$For simplicity, Left and right here are defined according to the relative position of two arms or legs to the viewer.
Our algorithm was evaluated in two aspects, i.e., localization, segmentation and tracking. For localization, the competing algorithm is the state-of-the-art edge histogram (EH) method in [7] that is a mid-level computation and generates an intermediate localization map image for an OOI. For segmentation where the online learning of figure/ground color models is the key issue, we compare our semi-parametric technique with the Fast Gaussian transform (FGT) that is a non-parametric learning technique in [10]. 180 test images are from six individuals that have the similar poses with the training data, and we also manually obtained the ground-truth segmentation and localization information for all test images with respect to six body parts. For tracking, we reported some experimental results from our recent work ([5]) where the proposed algorithm is used for part detection in articulate human tracking.

5.1. Localization

As a mid-level vision task, our algorithm outputs a map image of a given OOI in an image. Its pixel value indicates the likelihood or confidence of the OOI at each pixel location in the image. Given a map image that is the mid-level vision output, we expect that the ground truth position locates at (or close enough to) a local maximum. In order to evaluate the quality or saliency of map images, we define a localization accuracy function that shows the relationship between the tolerance error and the hit rate. A tolerance error defines an acceptable region around the ground truth position, and the hit rate records the percentage of the local maxima falling in the acceptable region. We believe a good map image should provide high localization precision locally (rather than globally). As shown in Figure 8, our algorithm is compared with the EH method in terms of localization accuracy for six body parts.
body parts. Specifically, we further analyze our algorithm by considering the case with (Seg-H-T-WS) and without the smoothness term (Seg-H-T-WOS) in the edge-based segmentation evaluation in (7).

![Comparison of localization accuracy for six body parts.](image)

There are two observations from Figure 8. (1) the Seg-H-T offers better localization accuracy than the EH, especially for two legs. (2) The smoothness term is more useful for the OOI with smooth boundaries, such as the head and two arms. As for the OOIIs with less well defined boundaries, such as the two legs (due to different shoes and pants worn by the subjects), the smoothness term is less useful. It is also possible that the usefulness of the smoothness term may not be fully exploited because the imperfect segmentation (with rugged boundaries) at the true location could be mis-judged.
by edge-based evaluation. On the other hand, the EH method is more efficient due to a direct 2D convolution involved, while the Seg-H-T involves a segmentation at each pixel location. Some preprocessing could be used to trim the candidate locations for speeding-up. For example, we used a simple shape matching method that convolves the edge map (consists of all watershed boundaries) with the edge-based shape prior (under different orientations) and selects only 2% pixel locations of the best shape matching. Then it takes about 10 seconds per image for the Seg-H-T, and about 1 second for the EH method. It is interesting to find out that there is little chance for both methods to achieve successful localization in the same image at low tolerance errors. It implies that there is a complementary nature between the two methods, and they could be combined together for more accurate (due to Seg-H-T) and efficient (due to EH) localization. We also compared Seg-H-T with the HOG method [9] that is normally used for object detection. The two methods are comparable in terms of position estimation. Our method also yields the segmentation result that is not available from other methods (e.g., EH and HOG). More importantly, HOG is robust to scale and rotation that makes it unable to detect the object orientation that is also needed for object localization, especially when dealing with multiple similar-looking objects.

As mentioned before, the proposed Seg-H-T method accomplishes a mid-level computation, and Fig. 8 only partially reveals its advantages over the EH method for localization. We have compared the two methods in terms of their usefulness for high-level vision tasks, such as pose recognition and localization, in [4], where we incorporated the spatial prior of six body parts represented by a “star” model proposed in [7]. It was shown that the new method provides significant advantages over the EH method for pose recognition and whole-part localization. Here, we show part localization results in Table 1 which is obtained by averaging the localization results of 144 tests images from 21 persons in the database. We can see that significant improvements are achieved for two arms and two legs that undergo major movement during walking. The proposed Seg-H-T method offers more salient mid-level outputs, i.e., map images, which ensures precise localization. At the same time, we can achieve the segmentation results for each body part given the correct localization which are not available from the EH-based method [7].
Table 1: The comparison of localization errors (in pixel) in one walking cycle.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Head</th>
<th>Torso</th>
<th>L-arm</th>
<th>R-arm</th>
<th>L-leg</th>
<th>R-leg</th>
</tr>
</thead>
<tbody>
<tr>
<td>EH</td>
<td>5.23</td>
<td>6.83</td>
<td>12.17</td>
<td>11.07</td>
<td>12.73</td>
<td>12.77</td>
</tr>
<tr>
<td>Seg-H-T</td>
<td>5.07</td>
<td>6.03</td>
<td>9.80</td>
<td>8.63</td>
<td>4.83</td>
<td>5.63</td>
</tr>
<tr>
<td>Improvement</td>
<td>3%</td>
<td>11.7%</td>
<td>19.5%</td>
<td>22.0%</td>
<td>62.1%</td>
<td>55.9%</td>
</tr>
</tbody>
</table>

5.2. Segmentation

One key idea in our Seg-H-T approach is the semi-parametric kernel-based method for online color model learning that takes advantage of the super-pixel representation and supports efficient figure/ground segmentation at each hypothesized position. This is essential to the computation at mid-level vision. We compare the proposed color model learning technique with the Fast Gauss Transform (FGT) [10]. We have downloaded the online FGT code from [20], and used it for comparative studies. Also, we have implemented the pixel-wise FGT (FGT-P) and super-pixel-wise FGT (FGT-SP). We evaluate the three segmentation algorithms on 180 test images for six body parts at the ground truth position. Segmentation results are evaluated by the ratio of the falsely detected region size (including both false positives and false negatives) to the object size ([35]). The experimental results are shown in the Table 2.

Table 2: The comparison of segmentation errors (%) and speed (per segmentation) for different shape prior guided methods at ground truth location.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Time</th>
<th>Torso</th>
<th>Head</th>
<th>L-arm</th>
<th>R-arm</th>
<th>L-leg</th>
<th>R-leg</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGT-P</td>
<td>150 ms</td>
<td>35.8</td>
<td>31.18</td>
<td>43.59</td>
<td>51.63</td>
<td>34.04</td>
<td>33.35</td>
</tr>
<tr>
<td>FGT-SP</td>
<td>25 ms</td>
<td>3.83</td>
<td>1.85</td>
<td>17.74</td>
<td>9.89</td>
<td>7.72</td>
<td>6.55</td>
</tr>
<tr>
<td>Our method</td>
<td>2.3 ms</td>
<td>3.08</td>
<td>1.49</td>
<td>15.90</td>
<td>7.77</td>
<td>7.26</td>
<td>6.36</td>
</tr>
</tbody>
</table>

From Table 2, we can see that significant improvements can be achieved by using super-pixels in stead of raw image pixels and our method slightly outperforms the FGT-SP approach. One possible reason is that our method uses soft decision while the FGT involves hard decision to learn the color models. Moreover, it costs only about 2.3 ms for our method, and 150 and
25 ms for the FGT-P and FGT-SP methods respectively, for each segmentation, making the new method more appropriate in the hypothesis-and-test paradigm. Some segmentation examples of the head for the three methods are shown in Figure 9, where the first column shows the input images where the watershed cells and the edge-based shape prior are shown.

More segmentation results for other body parts are shown in Figure 10. There are two main possible causes to the segmentation errors. (1) Some super-pixels are under-segmented which contains both the foreground and background regions (e.g., the arms in the second row). The assumption of our algorithm is that there is no or little under-segmented super-pixels at low-level vision. (2) The OOI has an irregular shape that violates the shape prior (e.g., the legs in the first row). We also examined our algorithm in some real-world images in Figure 10 where the proposed algorithm is applied locally due to its nature of mid-level vision. In the first row, one head is detected and segmented out despite some occlusion by the leaves. The reason two other heads (viewed from top) are not detected are mainly due to the fact that the shape prior of head is mainly trained from the side-view images. In the second row, both the legs and the head can be detected for two pedestrians despite moderate scale and view variation between them. In practice, the result of watershed transform is critical to the localization performance since it produces the building blocks for the final formed segmentation.

5.3. Tracking

We also applied the proposed approach as the part detection technique for articulated human tracking in a video sequence ([5]), where both spatial and temporal priors are combined to detect, track and segment each body part with continuous pose variation. Specifically, the temporal prior is represented by the motion trajectory in a low-dimensional latent space learned from past tracking history by the Back Constrained-GPLVM proposed by [21], and it predicts the configuration of each body part in the next frame. The spatial prior is encoded by the start-structured graphical model proposed in [7], and it can be constructed “on-the-fly” from the predicted pose and used to evaluate and correct the prediction by assembling part detection results. Significant improvements are achieved by combining both spatial and temporal priors compared with the case of using the spatial prior only ([4]), as shown in Fig. 12. We can also obtain the part-level appearance model for each body part by that corresponds to the optimal localization in the present frame and can facilitate the tracking in the next frame (Fig. 13).
Figure 9: Segmentation results for the head in six images where the shape prior is imposed at the ground-truth location (a). The segmentation results using FGT-P (b) and FGT-SP (c), and our method (d).
Figure 10: Segmentation examples of five body parts (the left/right arms, torso, and left/right legs) where super-pixels and the edge-based shape prior are shown.

Figure 11: Segmentation and localization examples for real-world images.
Moreover, the occlusion problem can be well-handled due to the use of spatial prior that imposes some spatial constraints among different body parts for each pose (the right foot in the first three images and the right arm in the last three images, Fig. 12).

5.4. Limitation and Discussion

There are two major limitations in this work. The first is that it cannot deal with the deformation or scale changes, and the second is that it cannot handle occlusion. However, it is possible to modify the evaluation term defined in (7) to account for the scale change. For example, we can modify (9) to be the variance of distance values instead of summation of the distance values. Because a small variance value indicates that the two boundaries for comparison are likely to be parallel (i.e., only scale variation). We can also revise (9) with more tolerance for shape mis-matching in order to cope with the deformation, possibly leading to more false positives. On the other hand, if the deformable object has a few well defined parts, like the human body, we can use a part-based approach for object detection where the proposed method can be used for part detection. Occlusion handling usually involves some contextual information or prior knowledge that are normally not available at mid-level vision. In [5], we have used the proposed algorithm for part detection in articulate human tracking where both spatial and temporal priors are combined for part localization and segmentation. The occlusion
problem is effectively handled by using the multi-object tracking theory proposed by [18] along with the combined spatial and temporal modeling.

6. Conclusion and Future Research

We have presented a new simultaneous localization and figure/ground segmentation algorithm that involves coupled region-edge shape priors and is implemented in a segmentation-based hypothesis-and-test paradigm. Our research focuses on mid-level vision and can produce a spatially sensible map image that reveals the possibility of the existence of an OOI in each pixel location and can be used for part-based object recognition and detection. One possible improvement of this work is to introduce a deformable shape model or a more powerful edge-based shape representation, such as the multiscale-curvature shape model used in [43], which could enhance the adaptability of edge-based evaluation. The proposed method can be incorporated into other high-level vision research tasks, such as [16, 11, 31, 25, 30], where the human body is modeled as an assembly of body parts. It can also be directly used for part-based human detection, pose recognition, localization, tracking and segmentation, i.e., [5]
Acknowledgments

This work is supported in part by the National Science Foundation (NSF) under Grant IIS-0347613, and an Oklahoma NASA EPSCoR 2009 Research Initiation Grant, and an OHRS award (HR09-030) from the Oklahoma Center for the Advancement of Science and Technology (OCAST). The authors also thank the anonymous reviewers for their comments and suggestions.

References


