Hybrid Body Representation for Integrated Pose Recognition, Localization and Segmentation

Cheng Chen and Guoliang Fan
School of Electrical and Computer Engineering
Oklahoma State University, Stillwater, OK 74078
{c.chen, guoliang.fan}@okstate.edu

Abstract

We propose a hybrid body representation that represents each typical pose by both template-like view information and part-based structural information. Specifically, each body part as well as the whole body are represented by an off-line learned shape model where both region-based and edge-based priors are combined in a coupled shape representation. Part-based spatial priors are represented by a "star" graphical model. This hybrid body representation can synergistically integrate pose recognition, localization and segmentation into one computational flow. Moreover, as an important step for feature extraction and model inference, segmentation is involved in the low-level, mid-level and high-level vision stages, where top-down prior knowledge and bottom-up data processing is well integrated via the proposed hybrid body representation.

1. Introduction

We consider pose recognition, localization and segmentation of the whole body as well as body parts in a single image. This research is a fundamental step toward video-based human motion analysis that have been intensively studied recently. Pose recognition, localization and segmentation in a still image are challenging problems due to the variability of human body shapes and poses as well as the inherent ambiguity in the observed image. Our goal is to develop a hybrid human representation and the corresponding processing to assemble three tasks into one integrated framework. We propose a *hybrid body representation*, as shown in Fig. 1, where the four images show *the input image represented by watershed cells*, *the part-based body representation*, *the online learned whole shape prior*, and *the part/whole segmentation results* respectively. Particularly, the segmentation process that has been found useful for object recognition and localization is involved for learning and inference in this work.

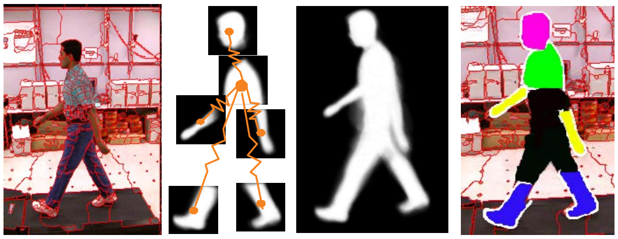


Figure 1. The input image represented by watershed cells, the part-based body representation, the online learned whole shape prior, and the part-whole segmentation results (from left to right).

The proposed research is deeply inspired and motivated by shape representation theories in cognitive psychology where there are two prevailing theories, i.e., the structural description-based and the view-based representations [10]. The former one suggests that a complex object is represented by a collection of simpler elements with specific inter-relationships. The latter one postulates a very simple template-like representation in which an objects is holistically represented by a simple vector or matrix feature without an intermediate representational step. Current cognitive studies indicate that none of these two representation schemes alone can provide a complete characterization of the human vision system for object recognition [7].

Similarly, existing shape representations in computer vision can be roughly grouped into two categories. One is template-like or silhouette-based methods, which are suitable for shape prior-based segmentation. The other is the part-based methods, which can capture the intra-class variability. *The main idea of our research is to integrate both view-based and structural description-based models into a hybrid body representation to support integrated pose recognition, localization, and segmentation*. Particularly, it can facilitate shape prior guided segmentation, by which bottom-up features can be extracted to drive the top-down inference in a cascade fashion. Additionally, both off-line and online learning are involved to learn general and subject-specific knowledge respectively, including the colors, shapes and spatial structure.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

2. Related work

Existing pose recognition, localization, and segmentation methods can be broadly grouped into three major categories according to the way how the body is represented: *the representation-free methods*, *the view-based methods*, and *the structural descriptions-based methods*.

The first category mainly contains some bottom-up approaches, in which there is no explicit shape prior representation, [15] and [19]. All the information used is a series of region grouping rules established according to physical constraints such as the body part proximity. In general, these approaches focus on exploiting bottom-up cues.

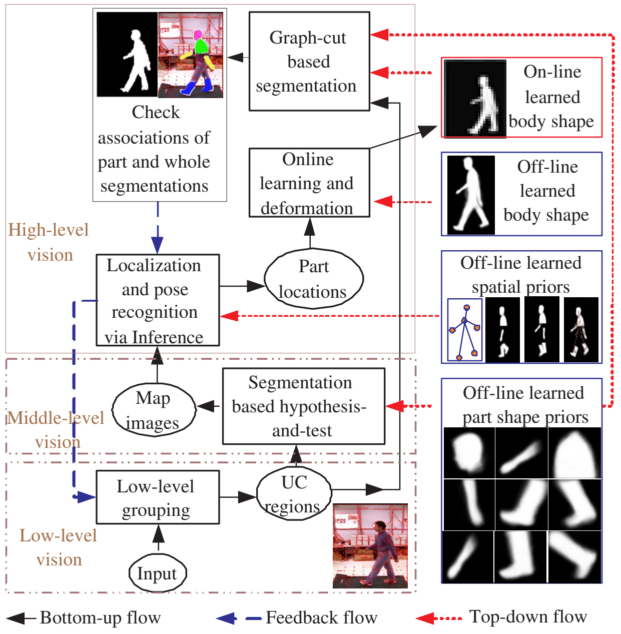
The second category includes all silhouette-based pose analysis methods. In [1], a specific view-based approach was proposed where pose information is implicitly embodied into a classifier learned from SIFT-like features. In general, no intermediate feature or color is used in these approaches. All view-based approaches normally aim at detecting particular body pose without extracting body parts. Thus it cannot recover anthropometric information.

The pictorial structure model proposed [5] is a typical approach belonging to the third category, in which the human body is described by several parts with their appearances and spatial relationships. This kind of approach usually requires a robust part detector. The edge histogram [3] and other SIFT-like features are widely used to represent parts. Very recently, a region-based deformable model is used to represent parts [17] where segmentation was used to verify the object hypothesis. The method in [17] is similar in spirit to the part-level inference proposed here. However, in our approach, where an image is represented by small building blocks (watershed cells), the coupled shape model is involved in a hypothesis-and-test paradigm where the region prior forms a segmentation given a position hypothesis and the edge prior evaluates the formed segmentation.

As the name suggests, the hybrid human body representation proposed here absorbs recent multifaceted advances in the field. The proposed representation involves shape prior guided segmentation and inference in a multi-stage fashion. Unlike previous methods, we use segmentation to extract bottom-up features to drive the top-down inference. Our contributions in this work include: (1) *a hybrid human body representation* that supports the online color model learning and involves an online learned deformable shape model to segment the whole body and parts, (2) *an effective hypothesis-and-test paradigm* for the part-level inference that involves the coupled region-edge shape priors, (3) *a three-stage cascade computational flow* to integrate pose recognition, localization and segmentation into a "biologically plausible" framework, and (4) *a new watershed-based Graphic-cut segmentation* where both region and edge shape priors are used for *optimal* segmentation.

3. Overview of our approach

The proposed hybrid body representation synergistically integrates pose recognition, localization and segmentation of the whole body as well as body parts in an image, as shown in Fig. 2. Several key issues are addressed.



Graph-cut based segmentation
On-line learned body shape
Check associations of part and whole segmentations
Online learning and deformation
Off-line learned body shape
High-level vision
Localization and pose recognition via Inference
Part locations
Off-line learned spatial priors
Middle-level vision
Map images
Segmentation based hypothesis-and-test
Off-line learned part shape priors
Low-level grouping
UC regions
Low-level vision
Input
Bottom-up flow
Feedback flow
Top-down flow

Figure 2. Overview of our approach.

Off-line and online learning: Off-line and online learning are used to obtain both general and subject-specific information, respectively. The former one acquires the general shape and spatial priors for both body parts and the whole body, and the latter one captures the subject specific information, including colors and shapes.

Part-whole organization: Parts and the whole are two complementary components for object representation. The part-level inference produces the *map* images that are assembled to localize the whole body as well as body parts.

Coupled region-edge shape model: The coupled region-edge shape representation supports a hypothesis-and-test paradigm, where the region-based prior is used to form a segmentation and the edge-based prior is used to evaluate the formed segmentation. After the online learning of the whole body, both priors are used in a new Graph-cut segmentation framework for an *optimal* segmentation.

Integration of bottom-up and top-down: Both the data-driven bottom-up and knowledge-driven top-down flows are integrated at low-level vision (watershed segmentation), mid-level (part-level inference/segmentation) and high-level (whole-level inference/segmentation). Potentially, this work can lead to a dynamic computational framework by incorporating two feedback flows (from high-level vision to mid-level and low-level vision).

4. Hybrid human body representation

Consider a walking cycle with K typical poses $\mathcal{W} = \{W^{(k)}|k = 1,...,K\}$, We model each pose $W^{(k)}$ by both part-based and whole-based statistical representations $W^{(k)} = \{V_{1:d}^{(k)}, \mathcal{L}^{(k)}, SW_{off}^{(k)}\}$, where $V_{1:d}^{(k)}$ are shape priors of d part, $\mathcal{L}^{(k)}$ is a set of statistical parameters that encode the spatial relationships between parts in a star graphical model, and $SW_{off}^{(k)}$ is the off-line learned shape prior of the whole body. The shape prior of each part $V_i^{(k)}$ is represented by the region-based shape prior $SP_i^{(k)}$, the edge-based shape prior $\mathcal{M}_i^{(k)}$, and the average orientation $\bar{\theta}_i^{(k)}$, i.e., $V_i^{(k)} = \{SP_i^{(k)}, \mathcal{M}_i^{(k)}, \bar{\theta}_i^{(k)}\}$. Moreover, during the inference processes, the part-based and whole-based color models as well as the subject specific whole shape model will be online learned as part of the hybrid body representation. For clearness, we may omit the pose index (k), in some places below.

4.1. Part-based shape prior

Inspired by the *MetaMorphs* model in [9], we develop an implicit shape model for each part where both region-based and edge-based shape priors are holistically represented. The learning process is similar to the one in [13].

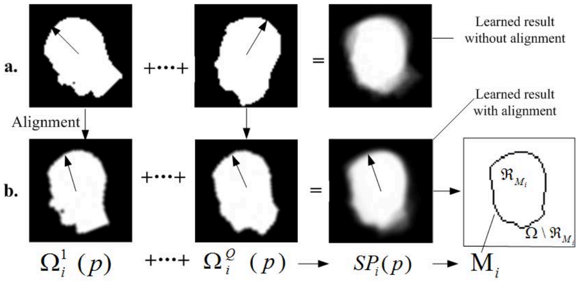

a.
+···+
=
Learned result without alignment
Alignment
Learned result with alignment
b.
+···+
=
$\Re_{M_i}$
$\Omega \setminus \Re_{M_i}$
$\Omega_i^1\ (p)$
+···+
$\Omega_i^Q\ (p)$
$SP_i(p)$
M_i

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.

For each part, we have obtained a set of training images (pre-segmented binary images with a fixed window size and the measured orientation). Let $\bar{\theta}_i$ be the average orientation of part V_i. All training images have been aligned to the average orientation first. Let $\{\Omega_i(p)|j = 1,...,Q\}$ denote the aligned training images, where $p = (x, y)$ is the location of one pixel in the window where the shape prior is defined. The shape prior $SP_i(p)$ can be obtained by adding all aligned training images as shown in Fig. 3,

$$SP_i(p) = \frac{1}{Q}\sum_{j=1}^{Q}\Omega_i(p).$$

$SP_i(p)$ and $1 - SP_i(p)$ reflect the the probability of pixel p belonging to the object and background respectively. Given a threshold ε, an *average* object boundary $\mathcal{M}_i$ can be extracted from the learned region-based shape prior $SP_i(p)$ by a level-set like method,

$$\mathcal{M}_i = \{p|SP_i(p) = \varepsilon\}. \tag{1}$$

Since both the region-based and edge-based priors are embedded in V_i, the two priors can be learned simultaneously in the training process. More importantly, this coupled representation allows a hypothesis-and-test paradigm for the inference at the part-level where the region prior induces a segmentation given a position hypothesis and the edge prior is used to validate the segmentation, resulting part-based possibility maps for whole-based inference.

4.2. Part-based spatial prior

We use the spatial prior model proposed in [3] to characterize the variability of spatial configuration of body parts. For pose k, we define the part-based spatial prior by a start graphical model as shown in the second figure of Fig. 1 that is parameterized by $\mathcal{L}^{(k)} = \{\mu_i^{(k)}, \Sigma_i^{(k)}|i = 1,....,d, i \neq r\}$. Specifically, $\{\mu_i^{(k)}, \Sigma_i^{(k)}|i \neq r\}$ denote the Gaussian priors for the *relative* locations between the non-reference part i and the reference part r. These statistical parameters can be obtained by a maximum-likelihood estimator (MLE) from labeled training data. Given a particular spatial configuration of d parts, $L = (l_1,...,l_d)$, the joint distribution of d parts with respect to pose k can be written as the following:

$$p_{\mathcal{L}^{(k)}}(L) = p_{\mathcal{L}^{(k)}}(l_1,...,l_d) = p_{\mathcal{L}^{(k)}}(l_r)\prod_{i\neq r}p_{\mathcal{L}^{(k)}}(l_i|l_r). \tag{2}$$

The same as in [3], we assume that $p_{\mathcal{L}^{(k)}}(L)$ is Gaussian. Therefore, the conditional distribution $p_{\mathcal{L}^{(k)}}(l_i|l_r)$ is still Gaussian. As defined above, $\mu_i^{(k)}$ and $\Sigma_i^{(k)}$ are the mean and covariance for the spatial distribution (relative) of part i in pose k. Then, for each non-reference part i, the conditional distribution of its position with respect to pose k is defined below,

$$p_{\mathcal{L}^{(k)}}(l_i|l_r) = \mathcal{N}(l_i - l_r|\mu_i^{(k)}, \Sigma_i^{(k)}). \tag{3}$$

4.3. Whole body shape prior

For each pose, a whole body shape prior is needed for body segmentation after pose recognition and localization. Both off-line and online learning are involved for generating shape models that capture the general representation as well as the subject specific information, as shown in Fig. 4.

4.3.1 Off-line learning

The off-line learning is similar to that of parts, except that a part-based alignment is needed due to the spatial variability

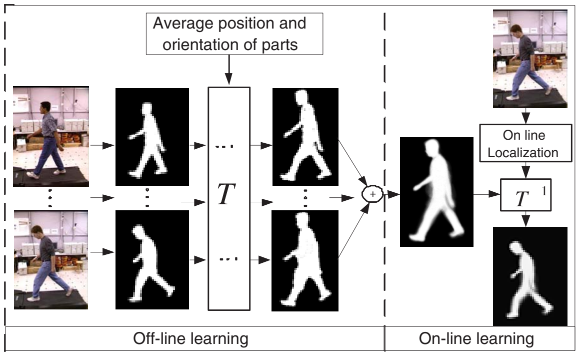


Figure 4. The learning of the whole body shape prior.

of each pose. For each pose, we can compute the average location and orientation for all parts. For each training image with a segmented body and parts, we want to find a set of control points based on which the training image can be deformed in a way that all parts are transformed to the average location and orientation. To preserve the shape information of parts, we will use the edge points of all parts to be control points which can be transformed to a target image via 2-D rigid transformations obtained from the averaged locations and orientations. After we get control points in both source and target images, the multilevel B-spline method [11] is used to obtain non-linear transformations by which all pixels in the source image are mapped to the target image. Small holes can be filled by simple morphological operations. These aligned biliary images are used construct a whole body shape prior, i.e., $SW_{off}(p)$.

4.3.2 Online learning

The online learning is used to create a subject specific shape model $SW_{on}(p)$ after all parts are localized. The goal is to deform $SW_{off}(p)$ in a way that the locations of all detected parts are reflected in the shape prior. The similar technique described above for off-line learning is used here to find the non-linear transformation functions for every pixel in $SW_{off}(p)$ by which $SW_{off}(p)$ is converted to $SW_{on}(p)$ that carries a subject specific shape model. It is worth noting that both $SW_{off}(p)$ and $SW_{on}(p)$ are not binary images. Appropriate interpolation is needed to fill the possible holes in $SW_{on}(p)$ during the deformation process.

5. Low-level vision: watershed transform

Grouping pixels into small homogenous regions is becoming a popular pre-processing for many computer vision tasks. This is well supported by the cognitive theory proposed by [16] that considers uniform connectedness (UC) regions as the building block for object representation. In this work, we chose the watershed transform [20] because of its many "biologically plausible" properties, such as fast, local computation. More importantly, both boundary and regional information are available for each cell. To overcome the over-segmentation problem, the geodesic reconstruction preprocessing [14] is used to control the watershed size through some morphological parameters, which can be dynamically adjusted according to the feedback from the high-level vision (as shown in Fig. 2).

A given an image I, is represented by Z watershed cells $I = \{\mathcal{C}_i | i = 1, 2, ..., Z\}$. Each cell consists of a set of pixels $\mathcal{C}_i = \{p_1^{(i)}, p_2^{(i)}, ..., p_{\eta_i}^{(i)}\}$, where η_i is the number of pixels. Moreover, we use a 3-D Gaussian model $\{\mu_i^{(c)}, \Sigma_i^{(c)}\}$ to represent the color distribution in the $L * a * b$ color space for cell $\mathcal{C}_i$. The watershed cells are used as the building blocks in the following processes.

6. Mid-level vision: part-based inference

The goal of the mid-level vision is to generate immediate part detection results that will be useful for the high-level vision. What we need here is a *map* image that indicates how likely there is an object (i.e., a body part) at each location. Recently, the idea of using segmentation to verify object hypotheses has achieved remarkable success in object detection [18, 12]. We hereby propose a new hypotheses-and-test paradigm, as shown in Fig. 5, where the region-based shape prior is used to form a segmentation for a given position hypothesis and the edge-based shape prior is used to validate the formed segmentation.

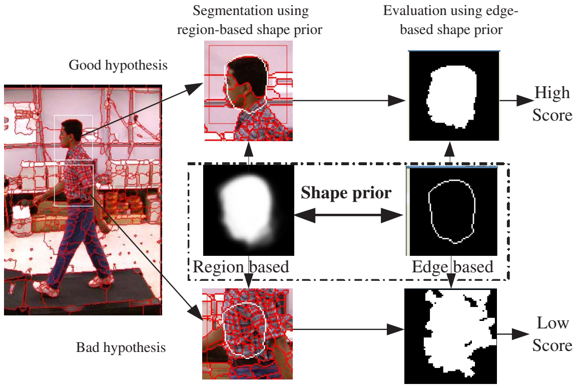

Segmentation using region-based shape prior
Evaluation using edge-based shape prior
Good hypothesis
High Score
Shape prior
Region based
Edge based
Bad hypothesis
Low Score

Figure 5. The hypothesis-and-test paradigm for part inference.

6.1. Hypothesis step: region-based segmentation

Given the shape prior $SP_i(p)$ $p \in \Omega$ for part i where Ω is a rectangular window, we can use Ω as a sliding-window to scan through the whole image to examine the existence of part i at each location. For a hypothesized location, we use $SP_i(p)$ to induce a local figure-ground segmentation that is composed by some watershed cells covered by Ω. This segmentation will be used to validate the existence of part i at that location. In order to take advantage of watershed cells and their color models, we propose a new *semi-parametric* kernel-based model learning techniques to online learn the figure/ground color models from the watershed results directly, as shown in Fig. 6(a). We treat the Gaussian model learned from a watershed cell as a kernel center, and learn the figure/ground color models as follows,

$$\hat{f_{ob}}(x) = \sum_{i=1,\mathcal{C}_i\cap\ \Omega\neq\emptyset}^{Z} \alpha_i K_i(x), \tag{4}$$

$$\hat{f_{bg}}(x) = \sum_{i=1,\mathcal{C}_i\cap\ \Omega\neq\emptyset}^{Z} \beta_i K_i(x), \tag{5}$$

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

$$\alpha_i = \frac{1}{\mathcal{T}} \sum_{p\in(\mathcal{C}_i\cap\Omega)} SP(p), \tag{6}$$

$$\beta_i = \frac{1}{\mathcal{T}} \sum_{p\in(\mathcal{C}_i\cap\Omega)} (1 - SP(p)), \tag{7}$$

where $\mathcal{T}$ is the size of shape prior window Ω. Based on the figure/ground color models, we can use the maximum *a posterior* (MAP) criterion to identify the watershed cells that belong to the object. Let τ_i be the class label for $\mathcal{C}_i$:

$$\tau_i = \begin{cases} 1 \text{ (object)}, & \alpha_i \hat{f_{ob}}(\mu_i^{(c)}) > \beta_i \hat{f_{bg}}(\mu_i^{(c)}); \\ 0 \text{ (background)}, & \alpha_i \hat{f_{ob}}(\mu_i^{(c)}) < \beta_i \hat{f_{bg}}(\mu_i^{(c)}). \end{cases} \tag{8}$$

Therefore, we can obtain the corresponding segmentation for the position hypothesis l_i as, $X = \{\bigcup \mathcal{C}_i | \tau_i = 1\}$. Different from the one in [18] where the shape prior is used once for online color model learning, here we use the shape-based prior twice. The first time is for the online color model learning and the second time is for MAP-based segmentation. Considering the false negative is more detrimental than the false positive in mid-level vision, we fully incorporate the region-base shape prior into segmentation. This may lead to more false positives due to more object-like segmentations. However, the sequent edge-based evaluation will mitigate this problem.

6.2. Test step: edge-based evaluation

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

$$\rho_{\mathcal{M}}(X) = \exp(-d_c(\Gamma(X), \mathcal{M})) + \zeta(1 - \mathcal{S}(\Gamma(X), \mathcal{M})), \tag{9}$$

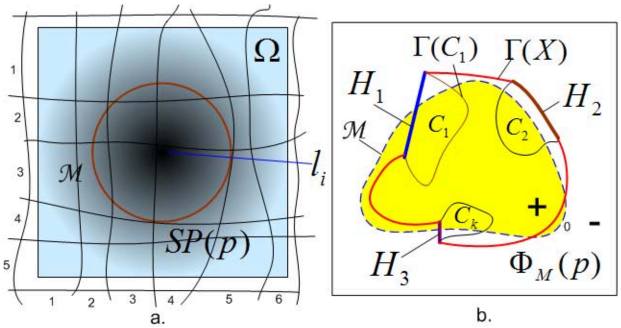


Figure 6. (a) Kernel-based color model learning using the region-based priors and watershed cells.(b) Smoothness measurement.

where the first term is the chamfer distance indicating the shape similarity; the second term measures the boundary smoothness; and ζ balances the relative importance between the two terms. It is expected that a valid segmentation should be smooth and match with $\mathcal{M}$ well. The chamfer distance is sensitive to the transition, rotation and scale. This is desired for rejecting false hypotheses. The second term aims to reject segmentations with ragged boundaries. We define an effective smoothness measurement as follows.

As shown in Fig. 6(b), assume that $\Gamma(X)$ touches n cells $\{\mathcal{C}_1, ..., \mathcal{C}_n\}$, and we define $H_i = \{h_1^{(i)}, ..., h_{n_i}^{(i)}\}$ to be the set of n_i boundary pixels shared by cell $\mathcal{C}_i$ and $\Gamma(X)$. Let $\phi_{\mathcal{M}}(p) : \mathrm{R}^2 \rightarrow \mathrm{R}$ to be the signed Euclidian distance transform that is "+" or "-" for p inside or outside $\mathcal{M}$, respectively. The maximum and minimum distances from H_i to $\mathcal{M}$ are obtained by $d_{max}^{(i)} = \max(\phi_{\mathcal{M}}(h_1^{(i)}), ..., \phi_{\mathcal{M}}(h_{n_i}^{(i)}))$, and $d_{min}^{(i)} = \min(\phi_{\mathcal{M}}(h_1^{(i)}), ..., \phi_{\mathcal{M}}(h_m^{(i)}))$, respectively. The degree of parallelness between $\Gamma(X)$ and H_i is measured by

$$\mathcal{S}_{\mathcal{M}}(H_i) = \frac{d_{max}^{(i)} - d_{min}^{(i)}}{n_i}. \tag{10}$$

When H_i is parallel to $\mathcal{M}$ (e.g., H_2 in Fig. 6(b)), $\mathcal{S}_{\mathcal{M}}(H_i) \cong 0$, indicating good local smoothness. When H_i is perpendicular to $\mathcal{M}$ (e.g., H_3 in Fig. 6(b)), $\mathcal{S}_{\mathcal{M}}(H_i) \cong 1$, 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

$$\mathcal{S}(\Gamma(X), \mathcal{M}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\mathcal{M}}(H_i). \tag{11}$$

For each of d parts of pose k, the score function (9) will return a *map* image that records the existence possibility of a part in every location and will be the input for the high-level vision. It is worth noting that at each position hypothesis, the shape prior is hypothesized with different orientations around the mean angle, and we use the winner-take-all strategy to generate the *map* image. The optimal angle at each location is also recorded.

7. High-level vision: recognition/localization

We use $g_i^{(k)}(I, l_i)$ to represent the *map* image for part i of pose k in image I, and l_i denotes an arbitrary position. Let $I^{(k)}_{maps} = \{g_i^{(k)}(I, l_i), ..., g_d^{(k)}(I, l_d)\}$ denotes the set of d *map* images, part localization and pose recognition are formulated as an inference process guided by the spatial priors of different poses represented by $\{\mathcal{L}^{(k)} | k = 1, ..., K\}$. Using Bayes law, the posterior distribution for pose k can be written in terms of the map images $I^{(k)}_{maps}$ and the spatial prior defined in (2) as ,

$$p_{\mathcal{L}^{(k)}}(L|I^{(k)}_{maps}) \propto p_{\mathcal{L}^{(k)}}(I^{(k)}_{maps}|L).p_{\mathcal{L}^{(k)}}(L). \tag{12}$$

Let $P_{\mathcal{L}^{(k)}}(I^{(k)}_{maps}|L) = \prod_{i=1}^{i=d} g_i^{(k)}(I, l_i)$, by manipulating the terms in (12), we have

$$p_{\mathcal{L}^{(k)}}(L|I^{(k)}_{maps}) \propto p_{\mathcal{L}^{(k)}}(l_r) g_r^{(k)}(I, l_r) \prod_{i \neq r} p_{\mathcal{L}_k}(l_i|l_r) g_i^{(k)}(I, l_i). \tag{13}$$

Then pose recognition and part localization can be jointly obtained by the following optimization:

$$\{k^*, L^*\} = \arg\max_{k,L} p_{\mathcal{L}^{(k)}}(L|I^{(k)}_{maps}). \tag{14}$$

However, the direct evaluation of (14) is computationally prohibitive. We use the efficient inference engine proposed in [3] to obtain the solution here. For any non-reference part i of pose k, the quality of an optimal location can be,

$$\epsilon^*_{k,i}(l_r) = \max_{l_i} p_{\mathcal{L}^{(k)}}(l_i|l_r) g_i^{(k)}(I, l_i). \tag{15}$$

Given $p_{\mathcal{L}^{(k)}}(l_i|l_r)$ is Gaussian, $\epsilon^*_{i,k}(l_r)$ can be computed by the generalized distance transform. Then, the posterior probability of an optimal configuration for pose k can be expressed in terms of the reference location l_r and ϵ^*_i. Then the posterior probability in (13) will become,

$$p_{\mathcal{L}_k}(L|I^{(k)}_{maps}) \propto p_{\mathcal{L}_k}(l_r) g_r^{(k)}(I, l_r) \prod_{i \neq r} \epsilon^*_{k,i}(l_r), \tag{16}$$

which will lead to a new *map* image $G_k(I, l_r)$ that indicates how likely the reference part of pose k is in each location. This new *map* image $G_k(I, l_r)$ is the pooling results of all the *map* images in $I^{(k)}_{maps}$ via the spatial prior model of pose k, i.e., $\mathcal{L}^{(k)}$. Therefore, pose recognition and reference part localization can be efficiently implemented by

$$\{k^*, l^*_r\} = \arg\max_{k,l_r} G_{k=1:K}(I, l_r). \tag{17}$$

After the reference part is located, the position of each non-reference part can be obtained by

$$l^*_i = \arg\max_{l_i} p(l_i|l^*_r) g_i^{(k^*)}(I, l_i). \tag{18}$$

According to the maximum value of obtained $G_{k^*}(I, l^*_r)$, we could design a feedback loop to adjust the size of watershed cells in low-level vision, as shown in Fig. 2.

8. Whole body Segmentation via Graph-cut

Recently, the graph-cut approach has achieved considerable success in image segmentation. It has the capacity to fuse both boundary and regional cues in an unified optimization framework [2]. Several existing methods, such as [6], only incorporate a single shape prior (edge-based or region-based) into the segmentation process. Our contribution here is to combine two shape priors into segmentation where the image is represented by watershed cells.

After pose recognition and localization, online learned whole body shape priors, $SW^{L^*}_{on}(p)$, $\mathcal{M}^{L^*}_w$ and pose configuration L^* can be obtained. Given image $I = \{\mathcal{C}_i | i = 1, ..., Z\}$, $\tau = \{\tau_i | i = 1, ..., Z\}$ denotes the set of binary class labels for all watershed cells ($\tau_i = 0$: background and $\tau_i = 1$: object). Following the segmentation energy definition from [2]

$$E(\tau) = \lambda. \sum_{i=1}^{Z} R(\tau_i) + \sum_{\mathcal{C}_i \bigcap \mathcal{C}_j \neq \emptyset} E(H_{i,j}) \delta(\tau_i, \tau_j), \tag{19}$$

where $R(\tau_i)$ is the regional term, which relates to the posteriori probability of $\mathcal{C}_i$ belonging to class τ_i; $E(H_{i,j})$ is the boundary term, which represents the consistence between the edge-based shape prior $\mathcal{M}^{L^*}_w$ and local boundary formed by two cells, $H_{i,j} = \mathcal{C}_i \bigcap \mathcal{C}_j$; $\delta(\tau_i, \tau_j) = 1$ when $\tau_i \neq \tau_j$ otherwise $\delta(\tau_i, \tau_j) = 0$; λ specifies a relative importance between two terms.

The calculation of $R(\tau_i)$ involves online learning of figure/ground color models where region-based shape prior $SW^{L^*}_{on}(p)$ is involved for kernel-based density estimation, as discussed in Section 6.1. Let $\hat{f}^{(w)}_{ob}(x)$ and $\hat{f}^{(w)}_{bg}(x)$ be the figure/ground color models, and $\alpha^{(w)}_i$ and $\beta^{(w)}_i$ are computed from $SW_{on}(p)$ that denote the prior probabilities of $\mathcal{C}_i$ belonging to the object and background respectively. Therefore, $R(\tau_i)$ is defined as

$$R(\tau_i = 1) \quad = \quad -\ln \alpha^{(w)}_i \hat{f}^{(w)}_{ob}(\mu^{(c)}_i), \tag{20}$$

$$R(\tau_i = 0) \quad = \quad -\ln \beta^{(w)}_i \hat{f}^{(w)}_{bg}(\mu^{(c)}_i), \tag{21}$$

where $\mu^{(c)}_i$ is the mean color vector of $\mathcal{C}_i$. Using the same idea of edge-based shape evaluation defined in (9), let $X = H_{i,j}$, and we can define $E(H_{i,j}) = \rho_{\mathcal{M}}(X)$, which evaluates the consistence between $H_{i,j}$ and edge-based shape prior $\mathcal{M}^{L^*}_w$ in terms of the degree of parallelness and the shape similarity.

In a similar way, all body parts can also be segmented. Moreover, the segmentation in the high-level vision stage will help us extract more useful features to prune possible false positives. For example, false positives can be identified by checking the color similarity between the two arms or legs. The feedback loop from segmentation to localization (as shown in in Fig. 2) makes our framework a dynamic system that has potential to be further optimized.

9. Experimental results

Here we validate the effectiveness of the proposed approach on the CMU Mobo database [8], which contains images of 25 individuals walking on a treadmill. Each image is down-sampled to the size of 240×320 pixels, and each body part is defined in a window of 61×61 pixels. For each pose, there are totally six parts (the *head*, *torso*, *left-arm*, *right-arm*, *left-leg*, and *right-leg*), and 200 manually segmented binary images are used for the off-line learning of part-based and whole body shape priors. The algorithm was programmed in C++, and the test platform is Pentium 4 3.2GHz and 1GB RAM.

Totally, the experimental results are reported for three tasks, pose recognition, body/parts localization, and body segmentation. The main competing algorithms are the 1-fan method in [3] for the first two tasks and the graphic-cut algorithm defined in (19) for the last task. Although our approach is a dynamic process with potential for further optimization, we have fixed the watershed transform in all experiments (without the feedback from high-level vision to fine tune the watershed transform as shown in Fig. 2).

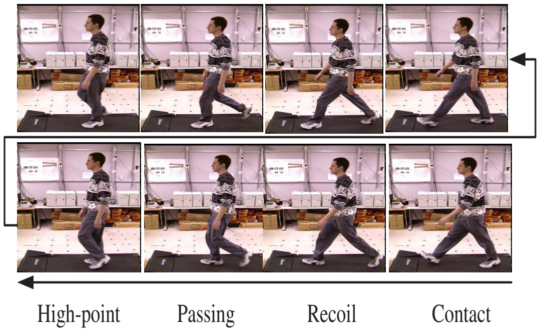


Figure 7. Pose definitions [4].

9.1. Pose recognition

In a walking cycle, the human pose is a continuous time-varying variable. According to [4], a complete walking cycle can be defined by eight poses that are further grouped into four poses due to the symmetric property. They are *Contact*, *Recoil*, *Passing* and *High-point*, as shown in Fig. 7. In our experiments, we combine poses *Recoil* and *Contact* together due to their strong similarity. For each of the three poses, the *torso* is used as the reference part, and 230 labeled training data are collected for learning the part-based spatial prior. It was found that the proposed approach achieves the recognition rate of 98% for the three poses over 144 test images from 21 persons. The mis-classification only occurs for pose *passing* that sometimes is very similar to other two poses. One possible way to improve the recognition for this pose is to incorporate the motion information from video sequences. The “1-fan” method in [3] achieved the recognition rate of 93%.

9.2. Localization

Based on the same test images used for pose recognition, we have tested two methods on the localization of three poses, i.e., *High-point* (H-point), *Contact* and *Passing*. Table 1 compares the proposed method with the 1-fan method in terms of localization error in pixel for six body parts. It is shown that the results on localizing the $head$ and $torso$ are comparable for the two methods, and our approach shows significant advantages on localizing other body parts.

Table 1. The comparison of localization errors (in pixel) between the two methods with respect to three poses.

Poses	Methods	Head	Torso	Larm	Rarm	Lleg	Rleg
H-point	1-fan	6.7	7.9	10.7	13	16.4	15.4
	hybrid	6.2	6.2	6.4	9.2	6.2	9.3
Contact	1-fan	3.4	6.4	13.9	10.4	8.7	10
	hybrid	5.4	5.6	11.8	9.6	3.8	4.4
Passing	1-fan	5.6	6.2	11.9	9.8	13.1	12.9
	hybrid	6.2	6.3	11.2	7.1	4.5	3.2

The reasons for above observations is that the relative position between the $head$ and $torso$ has least variability, and the part-based spatial prior that is shared by the two methods plays the major role for part localization, leading to the similar results. However, there is drastic (relative) spatial variability, both positional and orientational, for the $arms$ and $legs$, and the improvements from the proposed method are much more significant due to the enhanced saliency of the part-based *map* images generated by the segmentation-based hypothesis-and-test paradigm. Some part localization results of three poses are shown in Fig. 8 (the first three rows), where the proposed method successfully detects (and segments) all body parts despite the significant variability.

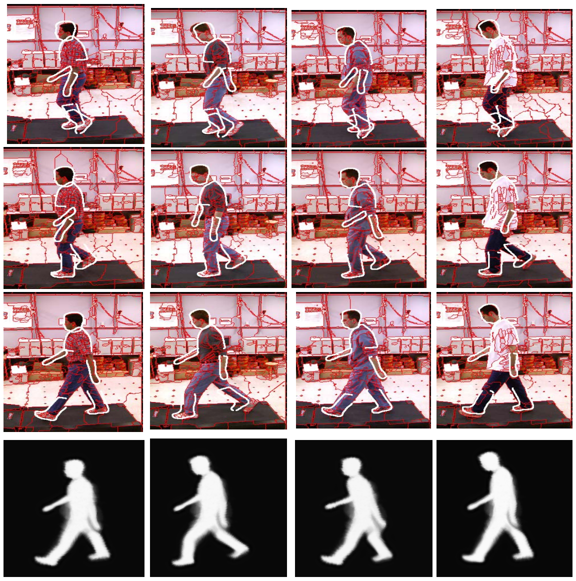


Figure 8. Part localization of three poses and online learned whole body shape models that are used for the whole body segmentation.

9.3. Segmentation

After localization, an online learned subject specific shape prior (as shown in the last row of Fig. 8) is used in the new Graph-cut algorithm where both region and edge priors are involved in the energy function defined in (19). For comparison,the second term of the energy function defined in (19) is replaced by a standard color similarity-based boundary penalty term [2] without using the edge-based shape prior. Currently, we only performed Graph-cut segmentation on pose *Contact* due to its least occlusion. By setting $\lambda = 1/30$ in (19) and using 60 test images, segmentation results are evaluated by the ratio between the falsely detected region size (including both false positives and false negatives) and the ground truth region size. The error rate of the segmentation using both region-based and edge-based priors is 17.2%, while that of the one without using the edge-based shape prior is 38.1%. Therefore, we have obtained more than 50% improvement.

10. Discussion and Conclusion

In this paper, we have proposed a hybrid body representation that supports an integrated pose recognition, localization and segmentation framework. Particularly, segmentation, as a bridge between bottom-up cues and top-down knowledge, plays an important role in all three levels of vision. In our experiment, when the number of poses goes up, the performance of pose recognition will deteriorate due to the overlap between adjacent poses. However, the performance of body/part localization and segmentation is quite stable and is less sensitive to the error of pose recognition. We also found that the proposed approach is very robust to the occlusion of one body part. More experiments need to be done to test its robustness for multiple occluded parts. The proposed framework has potential to be a dynamic system with the feedback loops that can be used for further optimization. In principle, this approach is applicable to any articulated object that has well defined spatial configurations and can be decomposed into different parts that have little shape variability. Our future research will focus on extending the proposed body representation to be a dynamic human body representation that supports video-based pose recognition, localization, tracking and segmentation.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments. This work is supported by the National Science Foundation under Grant IIS-0347613 (CAREER).

References

[1] A. Bissacco, M. H. Yang, and S. Soatto. Detecting humans via their pose. In *Neural Information Processing Systems (NIPS)*, 2006. 2

[2] Y. Boykov and G. Funka-Lea. Graph cuts and efficient n-d image segmentation. *International Journal of Computer Vision*, 70:109–131, 2006. 6, 8

[3] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. of IEEE CVPR*, 2005. 2, 3, 6, 7

[4] Dermot. http://www.idleworm.com/how/anm/02w/walk1.shtml, 2007. 7

[5] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, January 2005. 2

[6] D. Freedman and T. Zhang. Interactive graph cut based segmentation with shape priors. In *Proc. of IEEE CVPR*, 2005. 6

[7] R. L. Goldstone. Object, please remain composed. *Behavioral and brain sciences*, 21:472–473, 1998. 1

[8] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001. 7

[9] X. Huang, D. Metaxas, and T. Chen. Metamorphs: Deformable shape and texture models. In *Proc. of IEEE CVPR*, 2004. 3

[10] J. E. Hummel and B. J. Stankiewicz. Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Congnition*, 5:49–79, 1998. 1

[11] S. Lee, G. Wolberg, and S. Y. Shin. Scattered data interpolation with multilevel b-splines. *IEEE Trans. on Visualization and Computer Graphics*, 3:229–244, 1997. 4

[12] B. Leibe, E. Seemann, and B. Schiele. Pedestrain detection in crowded scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 4

[13] X. Li and G. Hamarneh. Modeling prior shape and appearance knowledge in watershed segmentat. In *The 2nd Canadian Conference on Computer and Robot Vision*, pages 27–33, 2005. 3

[14] Y.-C. Lin, Y.-P. Tsai, Y.-P. Hung, and Z.-C. Shih. Comparison between immersion-based and toboggan-based watershed image segmentation. *IEEE Trans. on Image Processing*, 15:632–640, 2006. 4

[15] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *Proc. of IEEE CVPR*, volume 2, pages 326–333, 2004. 2

[16] S. E. Palmer and I. Rock. Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin and Review*, 1:29–55, 1994. 4

[17] D. Ramanan. Learning to parse images of articulated bodies. In *Proc. of Neural Info. Proc. Systems (NIPS),*, 2006. 2

[18] D. Ramanan. Using segmentation to verify object hypotheses. In *Proc. of IEEE CVPR*, 2007. 4, 5

[19] P. Srinivasan and J. Shi. Bottom-up recognition and parsing of the human body. In *Proc. of IEEE CVPR*, 2007. 2

[20] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based onimmersion simulations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13:583–598, 1991. 4