Shape Guided Object Segmentation

Eran Borenstein
Division of Applied Mathematics
Brown University
eranb@dam.brown.edu

Jitendra Malik
Department of Electrical Engineering
and Computer Science
U.C. Berkeley
malik@eecs.berkeley.edu

Abstract

We construct a Bayesian model that integrates top-down with bottom-up criteria, capitalizing on their relative merits to obtain figure-ground segmentation that is shape-specific and texture invariant. A hierarchy of bottom-up segments in multiple scales is used to construct a prior on all possible figure-ground segmentations of the image. This prior is used by our top-down part to query and detect object parts in the image using stored shape templates. The detected parts are integrated to produce a global approximation for the object's shape, which is then used by an inference algorithm to produce the final segmentation. Experiments with a large sample of horse and runner images demonstrate strong figure-ground segmentation despite high object and background variability. The segmentations are robust to changes in appearance since the matching component depends on shape criteria alone. The model may be useful for additional visual tasks requiring labeling, such as the segmentation of multiple scene objects.

1. Introduction

Identifying and separating objects within images (figure-ground segmentation) represents a significant challenge due to high object and background variability. One approach to segmentation, the *bottom-up* approach, is to first segment the image into homogenous regions and then identify those corresponding to a single object. Relying mainly on continuity principles, this approach groups pixels according to their gray level or texture uniformity, as well as the smoothness and continuity of bounding contours. The main difficulty of this approach is that an object may be segmented into multiple regions, some of which may merge it with its background. This difficulty, as well as evidence from human vision (e.g. [10, 11]), suggests that object recognition facilitates segmentation. A complementary approach, *top-down segmentation*, is therefore to apply learned properties about an object – such as its possible shape, color, or texture – to guide the segmentation [2]. The main difficulty in this approach stems from the large variability in the shape and appearance of objects within a given class. Consequently, the segmentation may not accurately delineate the object's figure-ground boundary.

Works addressing these challenges include [8], where deformable templates are combined with bottom-up segmentation. The image is first over-segmented, and then various groupings and splittings are considered to best match a shape represented by a deformable template. This method faces difficult minimization in a high dimensional parameter space. Mori et al [9] take a similar approach to segment baseball players from their background. Another related work by [16] uses the spectral graph partitioning technique to jointly optimize a top-down and bottom-up grouping process. However, in [8, 9, 16] the bottom-up is limited to grouping specific segments in one predetermined scale. The work of [6, 3, 15] demonstrates strong results using a MRF model, yet poses difficult computational issues. Ren *et al* [12, 13] combine low-, mid- and high-level information in a conditional random field formalism. Leibe *et al* [7] use an MDL segmentation-based verification stage to detect and localize pedestrians in crowded scenes.

Borenstein and Ullman [2] apply image fragments for top-down segmentation; as well as its combination with bottom-up criteria [1], using a class of message-passing algorithms suggested by [5]. The fragments are detected in image regions that are sufficiently close in terms of a given intensity-based similarity measure. However, similarly shaped objects might have different intensities, thus requiring more templates for efficient representation. Furthermore, in many cases, an ob-

0-7695-2646-2/06 $20.00 (c) 2006 IEEE

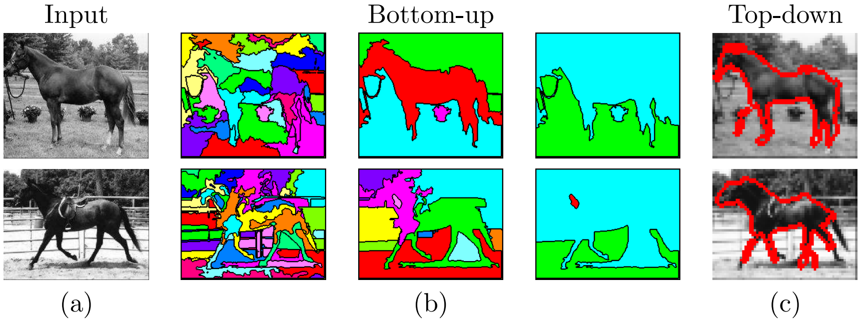

Figure 1. Relative merits of top-down (td) and bottom-up (bu) segmentation. (a) Input image, (b) Bu (here at 3 scales) can be applied to any given image to detect salient image discontinuities that potentially correspond to object boundaries. However, objects are usually segmented into multiple regions, some of which may merge with the background, (c) Td groups together dissimilar image regions corresponding to a single object and separates similar regions generated by different objects. However, it may not precisely follow image discontinuities.

ject's shape might be highly characteristic, thus offering information that is more robust than its intensity and texture.

In the present work, we construct a probabilistic model that relies on shape information (contour) to obtain concrete segmentation information that is invariant to changes in appearance. The model integrates bottom-up and top-down criteria to draw on their respective merits (Figure 1). It quickly queries and detects the grouping and splitting of bottom-up segments likely to form a specific top-down shape. The bottom-up part uses a hierarchical segmentation based on general image properties, such as uniformity in color and texture [4], to derive a segmentation (or grouping) prior. The top-down part applies the derived prior to match a set of stored *shape templates* (such as limbs, head and shoulders) to the image. The appropriate grouping of these detected templates then creates a consistent top-down shape that identifies the final segmentation (e.g. runner). In matching, the templates are compared to segments identified at multiple scales. This is in contrast to other approaches, which compare the template to the image itself or to one of its particular segmentations at a predetermined scale. Computations under the model are efficient, avoiding iterative calculations and their convergence issues. In particular, the matching is linear in template and image size and the inference of the final segmentation is also linear in the image size. Our results, tested on 328 gray level horses and 180 runner images, demonstrate that the model can deal with a large variability of object shapes, appearance and cluttered backgrounds.

2. Overview of the Approach

The main idea of our method is to use top-down shape templates to guide the grouping of homogenous bottom-up regions: The bottom-up process segments the image into homogenous regions V_j and for a given shape template S_i the goal is to form a shape $F_i = \cup_j V_j$ that is as similar as possible to the shape of S_i. However, there is a tradeoff between the top-down and bottom-up criteria. When the bottom-up process provides very small regions, it is possible to group them to form any specific shape anywhere in the image. In this case the top-down dominates the detection since template detections are not highly correlated with image content. On the other hand, when the building blocks represent large regions, it is difficult to use them to form a desired shape – they may be too big for that. In this case, the bottom-up dominates (Figure 2). We therefore use a hierarchy of segments in multiple scales to construct a prior on all figure-ground segmentation events of the image (Figure 3). This prior increases as more pixels that are strongly connected to the same salient segments are grouped together as figure or ground. It decreases as more salient regions are split to figure and background parts. We then apply a top-down process to guide the segmentation in forming a specific desired shape while maintaining a high prior for it. Initially, shape templates representing object parts are locally detected in the image and then integrated to form the global top-down segmentation Y (Figure 4). This approach can be represented by the

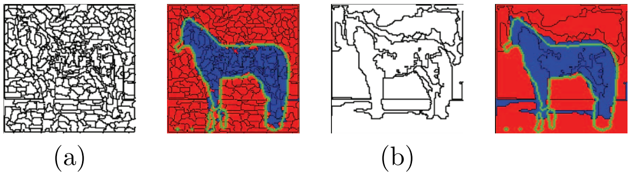

Figure 2. Tradeoff between td and bu: (a) It is easy to form a desired shape (green contour) using small segments (oversegmentation), but then it could be found anywhere. (b) However, it may be impossible to form the desired shape using large segments, which may merge the object with its background ("bleeding").

0-7695-2646-2/06 $20.00 (c) 2006 IEEE

following Bayesian model:

$$P(X^0|G,Y) = \frac{P(Y|X^0,G)P(X^0|G)}{P(Y|G)} \qquad (1)$$

which represents the probability of the combined segmentation being X^0, given the bottom-up and top-down segmentations G and Y. The first term $P(Y|X^0,G)$ $(= P(Y|X^0))$ models the relationship between the top-down segmentation Y and the combined segmentation X^0. The second term $P(X^0|G)$ models the prior of the combined segmentation X^0, given G. The denominator $P(Y|G)$ models the relationship between the top-down segmentation Y and the bottom-up segmentation G. (Note that the terms in (1) are conditional on the image segmentation G and are therefore image dependent.)

Given an image I, we first segment it to obtain G; then use shape templates S_i to find the Y that optimizes $P(Y|G)$; and finally find X^0 that optimizes $P(X^0|G,Y)$. Figure 5 shows an overview of the approach.

3. Segmentation Prior

The bottom-up process [4] segments the image into a hierarchy of homogenous regions. These regions are identified by a recursive coarsening process in which homogeneous segments at a given level are used to form larger homogeneous segments at the next level. In this manner, the image is segmented into fewer and fewer segments, producing a *segmentation weighted, hierarchical graph* $G(V,E)$, in which each segment V_i^l at a level l is connected with a relating weight E_{ij}^l to another segment V_j^{l+1} at a coarser level $l+1$, providing the first was one of the segments used to define the latter (Figure 3). The weight of an edge connecting two segments represents their similarity, taking into account texture, average intensity and boundary properties. Therefore, the more a segment is similar to a larger segment at the next coarser level, the stronger they are connected. These edges are normalized such that:

$$\sum_j E_{ij}^l = 1 \qquad (2)$$

Each segment V_i^l is represented by its connectivity $V_i^l(q)$ to every image pixel q. These connections are recursively determined by:

$$V_i^l(q) = \sum_j E_{ji}^{l-1} V_j^{l-1}(q) \qquad (3)$$

where a segment V_q^0 at the terminal level $l=0$ is connected with weight one to pixel q and weight zero to the other pixels. Note that a pixel may be connected to multiple segments at a given level so this structure provides a *soft segmentation* of the image.

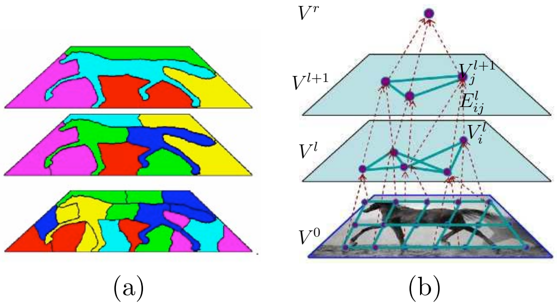

Figure 3. Hierarchical bottom-up segmentation of the image. Segments at multiple scales, identifying salient homogenous regions (a), are represented by a graph $G = (V,E)$ (b). The graph is constructed recursively: Segments V^l at a given level l are used to form larger segments V^{l+1} at the next level, as indicated by their connecting edges E.

The segmentation graph is used to define a generative model $P(X|G)$ for the labeling $X = \{X_i^l\}$ of its nodes $V = \{V_i^l\}$ with $X_i^l = 1$ representing the labeling of segment V_i^l as figure and $X_i^l = 0$ representing its labeling as background. Each segment V_i^l, first chooses a parent V_j^{l+1} from the next coarser level, with probability given by their connecting edge E_{ij}^l. The segment's labeling is then generated according to its saliency and its parent's labeling. The labeling process therefore starts at the root segments X^r and progresses recursively until the finest level X^0, representing image pixels. Formally, if X^l is a vector denoting the labeling of segments at the l-th level, then $X = (X^0, \ldots, X^r)$ and we can write:

$$P(X|G) = P(X^r) \prod_{l=0}^{r-1} P(X^l|X^{l+1}) \qquad (4)$$

with

$$P(X^l|X^{l+1}) = \prod_i P(X_i^l|X^{l+1}) \qquad (5)$$

We model $P(X^r)$ as i.i.d. Bernoulli (0.5) and:

$$P(X_i^l|X^{l+1}) = \sum_j E_{ij}^l P_G(X_i^l|X_j^{l+1}) \qquad (6)$$

The parent-child generation term $P_G(X_i^l|X_j^{l+1})$ is de-

0-7695-2646-2/06 $20.00 (c) 2006 IEEE

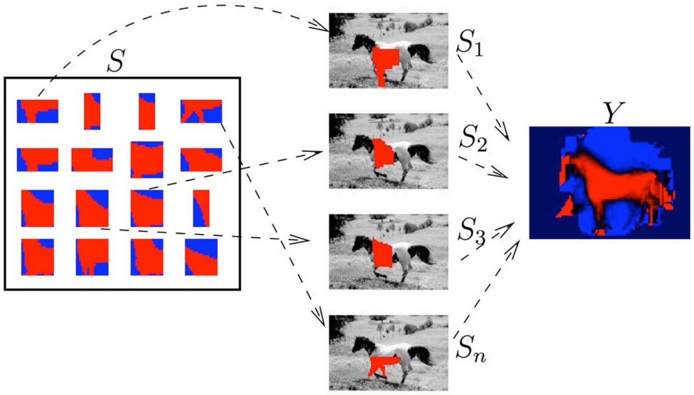

Figure 4. Top-down part: Shape templates representing object parts are detected in different image regions. The individual detections are integrated to derive the top-down segmentation Y. Red regions are labeled as figure, blue as background. (Color intensity represents their likelihood of being figure or background: dark - low likelihood, bright - high likelihood.)

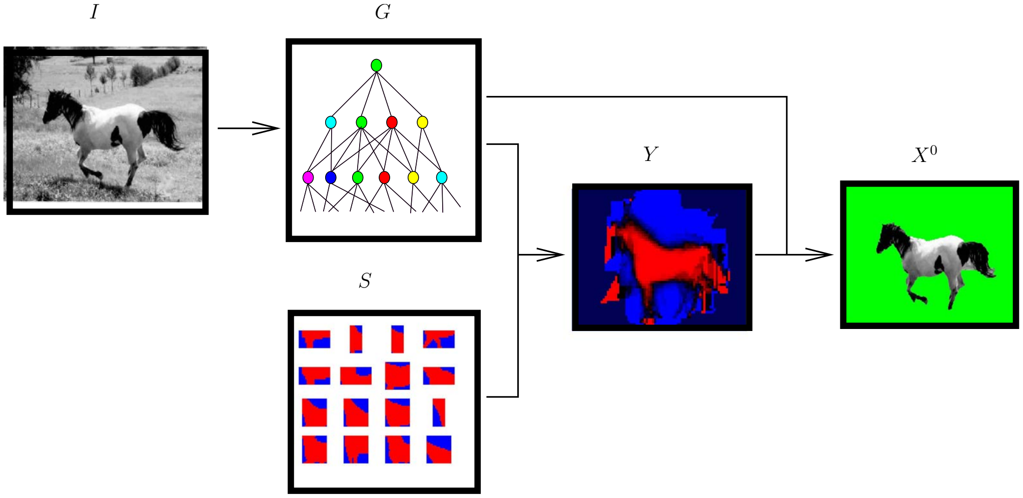

Figure 5. Overview of the approach: Given an input image I, a bottom-up process is used to derive a segmentation graph G which defines a segmentation prior. Top-down templates S are matched to the segmentation graph to detect object parts. These detections are integrated to form the top-down segmentation of the image Y which is then combined with the bottom-up prior to derive the final segmentation X^0.

termined by the segment's saliency $\Gamma(V_i) \in [0,1]$[1]:

$$P_G(X_i^l | X_j^{l+1}) = \left\{ \begin{array}{c|c} 1 - \frac{\Gamma(V_i^l)}{2} & X_i^l = X_j^{l+1} \\ \frac{\Gamma(V_i^l)}{2} & X_i^l \neq X_j^{l+1} \end{array} \right. \quad (7)$$

where $\Gamma(V_i^l)$ compares the segment's interior homogeneity (similarity between the segment points) and its contrast with the surrounding. For example, a uniform black segment V_i surrounded by a white background has a high saliency $\Gamma(V_i)$. Therefore, the more a segment is salient ($\Gamma(V_i^l) \to 1$), the more independent it is and the less likely to inherent its parent labeling. Other top-down information may be incorporated into this probability. For example, we can use it to give a higher prior for labeling a green segment as grass – i.e. background – in horse images.

It follows from this generative model that there is a high prior $P(X|G)$ for the uniform labeling of pixels that are strongly connected to some salient homogenous segment at coarser levels. Note that in this setting (4) defines a probability measure without any additional normalization terms.

[1] The saliency Γ is derived from Γ', the measure of saliency defined in [4], equation (2). However, this measure is inversely proportional to the segment's saliency and it also satisfies $\Gamma' \in [0, \infty]$. For our implementation we therefore re-normalize it to be proportional and in $[0, 1]$

0-7695-2646-2/06 $20.00 (c) 2006 IEEE

4. Shape-based Similarity Measure

To determine the top-down segmentation Y we use a set of detected shape templates. A shape template S_i is a binary (figure/ground) patch representing a shape part, in which each point is labeled as either figure or background (Figure 4, left). To detect a template S in an image region I_R it is necessary to define a shape-based similarity measure $\phi(S, I_R)$ between them. Assuming that the labeling I_R of a region R is generated by its corresponding terminal nodes X_R^0 in the segmentation graph G, we define the following labeling likelihood:

$$P(I_R|X) = \prod_{q \in R} P(I_q|X_q^0) = \\ = \prod_{q \in R} p^{1-|I_q - X_q^0|}(1-p)^{|I_q - X_q^0|} \tag{8}$$

where I_q represents the template labeling of pixel q and X_q^0 represents its corresponding node in G. Setting the Bernoulli parameter p to higher values defines a high correlation between the template and graph labeling (we set $p = 0.95$). This likelihood is then used to define the likelihood $P(I_R|G)$ for observing a specific labeling event $I_R = S$ in an image represented by G:

$$P(I_R = S|G) = \sum_X P(I_R = S|X)P(X|G) \tag{9}$$

where $P(X|G)$ is given by (4). However, the graph topology alone may influence the resulting likelihood (e.g. segments tangent to the image boundary usually have a smaller number of children than internal segments) and therefore, to define the similarity measure such that it is comparable for different regions, we use the following normalized likelihood ratio:

$$\phi(S, I_R) = \left[\frac{P(I_R = S|G)}{P(I_R \equiv 1|G)}\right]^{\frac{1}{|S|}} \tag{10}$$

where $P(I_R \equiv 1|G)$ represents the likelihood of a uniform labeling for R and the $1/|S|$ term normalizes the ratio to be independent of the template's size. We compare $P(I_R|G)$ to a uniform labeling $P(I_R \equiv 1|G)$, since this is the maximum value for $P(I_R|G)$. (Note that when $X \equiv 1$, all parent-child pairs are consistently labeled as figure and therefore all the $P_G(X_i^l|X_j^{l+1})$ terms in (6) are maximized). Therefore, for any S and I_R we have $\phi(S, I_R) \in [0, 1]$.

5. Top-down Segmentation

Given an image to segment, the top-down algorithm uses a set of stored templates, learned automatically from training examples (See appendix), to construct an object cover. The templates are first detected (Figure 4, middle) in image regions where the similarity measure (10) is sufficiently high (we set a detection threshold to give 0.05% false detection rate). The detected templates are used in a naïve base approach to identify a *Region Of Interest* (ROI) likely to contain the object: template detections are assumed to be independent given the ROI's position and we look for one (or more) ROI that maximize the posterior $P(\text{ROI}\,|\,\text{template detections})$. Detections that are spatially consistent with the ROI's position identify the initial object cover. The stored templates also provide an over-complete representation and therefore the detected fragments are likely to be overlapping and cover the entire object. A figure-ground consistency term is used to remove inconsistent templates and add consistent, overlapping ones. This object-cover construction is based on [2], which, however, uses image fragments rather than shape templates as the cover building blocks. Each covering template S_i^c provides a labeling event I_{R_i} for the region R_i it is covering. The final top-down segmentation Y (Figure 4, right) is then given by the weighted combination of these labeling events:

$$Y_q = \frac{\sum_{i:q \in R_i} \phi(S_i, I_{R_i}) S_i(q)}{\sum_{i:q \in R_i} \phi(S_i, I_{R_i})} \tag{11}$$

where $S_i(q)$ represents the labeling given to pixel q by the i-th template, which covers region R_i. In this manner we have $Y_q \in [0, 1]$ for every point q.

Using the resulting top-down labeling Y, we derive the combined segmentation $\hat{X}$ by the following MAP:

$$\hat{X} = \arg\max_X P(X|G, Y) = \arg\max_X P(Y|X, G)P(X|G) \tag{12}$$

with $P(Y|X, G) = P(Y|X^0)$, since the final segmentation of the image is determined solely by the labeling of the terminal segments:

$$P(Y|X^0) = \prod_q P(Y_q|X_q^0) = \prod_q \frac{1}{Z(\lambda)} \exp^{-\lambda|X_q^0 - Y_q|} \tag{13}$$

where $Z(\lambda)$ is the partition function of an exponential distribution with support inside the interval $[0, 1]$. In our experiments we used $\lambda = 1$.

6. Efficient computations

The computations of (10) and (12) are complex. We therefore use another generative model $P(X|T)$ to approximate $P(X|G)$, with T representing a spanning tree of G. The main idea is to express the labeling dependency of X_i^l upon X^{l+1} through a single parent $X_j^{l+1} = \pi(X_i^l)$ connected to the segment with the strongest edge E_{ij}^l:

$$P(X|T) = P(X^r) \prod_{l=0}^{r-1} \prod_i P(X_i^l | \pi(X_i^l)) \qquad (14)$$

However, while single-parent labeling simplifies the computation, we do not want to lose the information from the other parents. We therefore use the $P(X|G)$ prior to find $P(X_i^l|\pi(X_i^l))$, noting that it is exactly the conditional expectation:

$$P(X_i^l|\pi(X_i^l)) = \mathbb{E}\left[P(X_i^l|X^{l+1})|\pi(X_i^l)\right] = \qquad (15)$$
$$\sum_k E_{ik}^l \sum_{X_k^{l+1}} P_G(X_i^l|X_k^{l+1}) P(X_k^{l+1}|\pi(X_i^l))$$

which is obtained by substituting (6) for $P(X_i^l|X^{l+1})$ and using the linearity of the expectation operator. The labeling $P(X_i^l|\pi(X_i^l))$ therefore takes into account the average contribution of the other parents through the conditional probabilities $P(X_k^{l+1}|\pi(X_i^l)) = P(X_k^{l+1}|X_j^{l+1})$. These are computed by:

$$P(X_k^{l+1}|X_j^{l+1}) =$$
$$\sum_m E_{km}^l \sum_{X_m^{l+2}} P_G(X_k^{l+1}|X_m^{l+2}) P(X_m^{l+2}|X_j^{l+1}) =$$
$$\sum_m E_{km}^l \sum_{X_m^{l+2}} P_G(X_k^{l+1}|X_m^{l+2}) \frac{P(X_j^{l+1}, X_m^{l+2})}{P(X_j^{l+1})} \qquad (16)$$

The last two equations (15),(16) are computed recursively, starting with (15) at level $l = r-1$, recalling that $X^r \sim$ i.i.d Bernoulli, and then using (16) to compute the conditional lateral dependencies at level $l = r - 1$. This completes one recursive cycle and we repeat it until $l = 0$. Note that it may also require a recursive computation of $P(X_i^l)$ for all segments. In our case however, the prior is symmetric $P(X|G) = P(1-X|G)$ and therefore $P(X_i^l) = 0.5$ for all segments.

The tree structure of the derived $P(X|T)$ enables us to use dynamic programming to compute (12) and (10) through a simple message-passing algorithm (e.g. [5]). The complexity of these computations is linear in the number of tree edges, since every segment sends and receives exactly one message from its parent. It is also linear in the number of pixels, since at every level, the number of segments is less than half that of the previous level, giving $|V| < 2|V^0|$. As for the computation of the shape similarity (10) between a template S and a region R, it can be shown that the only informative messages are bottom-up messages sent from segments overlapping with R. The overall complexity of the similarity measure is therefore linear in the number of queried regions and their size.

7. Experiments

We conducted two independent tests: on a database of horse images (side views) and on runners (facing left, right and frontal positions). The runner segmentations are particularly challenging due to their higher average contour-length/region-area ratio. For the horses, 98 templates are chosen from 562 candidates to optimally cover 64 training horse silhouettes, as described in the Appendix. For the runners, 164 templates are chosen from 1401 candidates to optimally cover 36 training runner silhouettes. The templates' size vary between $14\times16 - 27\times27$ points. We then segment 328 gray level images of horses with a total of 4.3×10^6 figure and 14.4×10^6 background pixels and 180 gray-level images of runners with a total of 1.4×10^6 figure and 5.6×10^6 background pixels. The average figure consistency (segmenting a figure pixel as figure) and background consistency (segmenting a background pixel as background) is measured for each image using a manual benchmark. Histograms of these consistencies are plotted in Figure 6 for the top-down part alone and for the final, combined segmentation. The results demonstrate that the top-down segmentation is significantly improved by the combination, particularly in considering the large number of pixels classified. For example, the overall segmentation/benchmark consistency (average percentage of correctly classified pixels) is 93% for horses and 92% for the runners. This is contrast

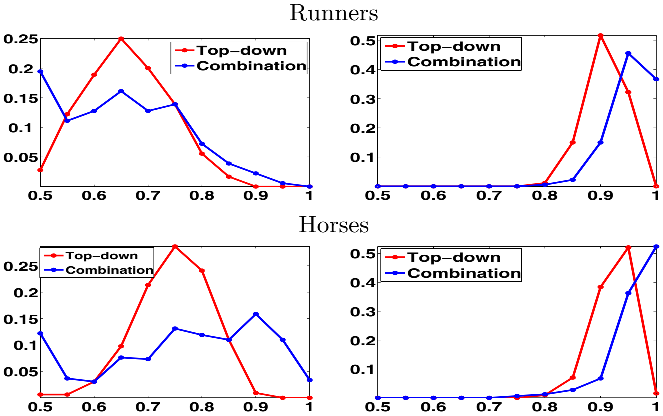

Figure 6. Segmentation consistency. Each bin (x-axis) in these histograms counts the frequency (y-axis) of segmentation results having a specific figure consistency (left) and background consistency (right) with a manual benchmark. Red represent top-down and blue represents combination results. These histograms are measured on 328 horse and 180 runner images.

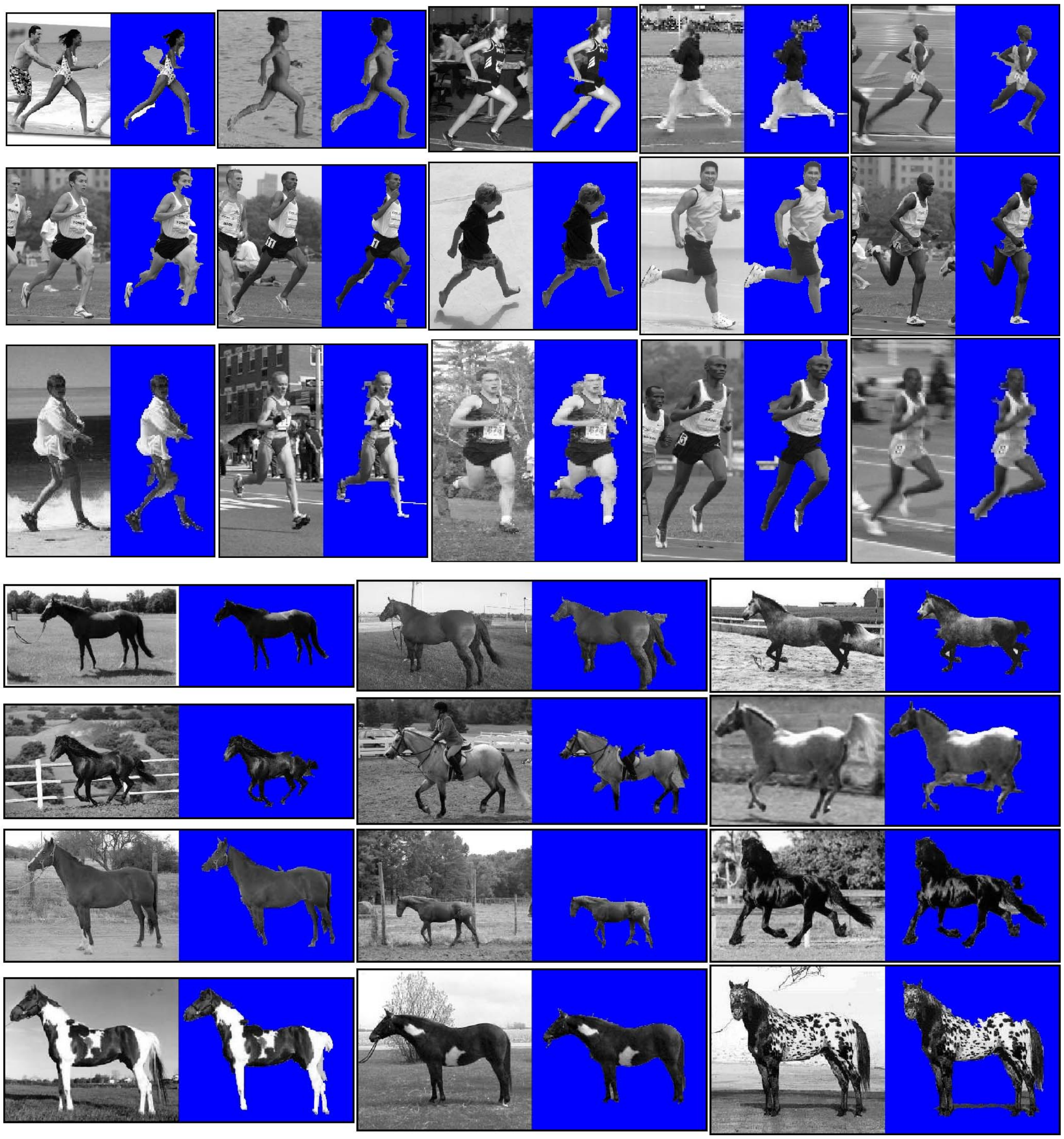

Figure 7. Results of the combined segmentation scheme.

to the top-down segmentation alone, where the results are 89% for horses and 87% for runners. A sample of the results are shown in Figure 7.

8. Conclusions

Shape is a powerful cue for recognizing and segmenting objects, offering specific advantages in various conditions, such as poor figure-ground contrast or illumination. In the present work we use shape-based object representation to obtain efficient class-specific figure-ground segmentation that is robust to changes in appearance. This appearance invariance property makes it possible to address images characterized by high object and background variability, using only a small number of templates. Despite a large configuration space, the model enables fast query and detection of figure-ground patterns having a high prior, as well as fast inference of an optimal compromise between a top-down suggested shape and the bottom-up prior. In contrast to other probabilistic models addressing this problem, our model provides a low computational framework – there are no normalization issues (partition functions), and computations are non-iterative and exact, thus avoiding convergence problems. The number of parameters is relatively small (2 for the bottom-up part and 2 for the top-down) and we use a fixed setting for all our experiments. The 2 bottom-up parameters, which are discussed in [4], control the relative contribution of texture vs. average intensity criteria for determining segment homogeneity; and the top-down and combination parameters define the Bernoulli (8) and exponential (13) distributions. Remaining difficulties include: addressing the estimation of object scale; and incomplete figures, occurring in conditions where both segmentation processes are challenged in the same region – such as the top-down missing a body part while the bottom-up merges it with its background.

The similarity measure derived in (10) may also be applied in addressing multiple classes by using template detections for classification. In addition, it could be generalized to templates consisting of multiple regions (e.g. an eye template may consist of the pupil, iris and their surrounding regions). We used only top-down shape information but the model could learn and apply other top-down knowledge. For instance, it could be used to "supervise" the learning of colors or textures characteristic of a specific shape (or parts thereof): Using horse images, for example, it can learn that regions grouped to form a horse head are more likely to be brown than green. This learned knowledge could then be reflected in the probability defined in (7).

Appendix A

The set of shape templates is extracted automatically from training examples. The first step is to collect a large random set of candidate templates from a set of binary (figure/ground) training images representing class silhouettes. Each candidate can be used to cover a region in one of these images if it successfully classifies more than 90% of the covered pixels as figure/background. The goal is to select a smaller subset to optimally cover (reconstruct) entire objects in the training images. We use a greedy algorithm to select such a subset. Candidates S_i are added one by one to maximize the gain in cover area:

$$S_i = \arg\max_{S_j} c(S^n \cup S_j) - c(S^n) \tag{17}$$

Where $c(S)$ denotes the total number of pixels covered by a set of shape templates S and S^n denotes the selected set at time n. The first selected candidate is therefore the one giving a maximal cover area $c(S_j)$ for the training shapes and the candidate selected at the n-th step is the one that mostly covers regions not covered by the previous $n-1$ selected templates. We stop the process when candidates no longer increase the gain in covered area (17). Note that this greedy algorithm is somewhat similar to the max-min selection suggested in [14]. However, their selection is based on mutual information rather than cover area gain.

Acknowledgement. This research is supported by MSRI and grant ONR MURI N00014-01-1-0890. We wish to thank Matthew Harrison for helpful discussions.

References

[1] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *CVPR, POCV*, Washington DC, 2004.

[2] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV (2)*, pages 109–124, 2002.

[3] X. Chen, Z. Tu, A. Yuille, and S. Zhu. Image parsing: Segmentation, detection and recognition. In *Proc. ICCV*, pages 18–25, Nice, France, 2003.

[4] M. Galun, E. Sharon, R. Basri, and A. Brandt. Texture segmentation by multiscale aggregation of filter responses and shape elements. In *ICCV03*, pages 716–723, 2003.

[5] F. Kschischang, B. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory*, 47:498–519, Feb 2001.

[6] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. In *CVPR (1)*, pages 18–25, 2005.

[7] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR (1)*, pages 878–885, 2005.

[8] L. Liu and S. Sclaroff. Region segmentation via deformable model-guided split and merge. In *ICCV (1)*, pages 98 – 104, 2001.

[9] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, Washington DC, 2004.

[10] A. Needham and R. Baillargeon. Effects of prior experience in 4.5-month-old infants' object segregation. *Infant Behaviour and Development*, 21:1–24, 1998.

[11] M. Peterson and B. Gibson. Shape recognition contributions to figure-ground organization in three-dimensional displays. *Cognitive Psychology*, 25:383–429, 1993.

[12] X. Ren, C. Fowlkes, and J. Malik. Cue integration for figure/ground labeling. In *NIPS*, 2005.

[13] X. Ren, C. Fowlkes, and J. Malik. Scale-invariant contour completion using conditional random fields. In *ICCV*, Oct 2005.

[14] S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7):682–687, Jul 2002.

[15] J. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *ICCV*, volume 1, pages 756–763, 2005.

[16] S. X. Yu, R. Gross, and J. Shi. Concurrent object recognition and segmentation by graph partitioning. In *NIPS*, 2002.