Partially Occluded Object-Specific Segmentation in View-Based Recognition

Minsu Cho and Kyoung Mu Lee
School of EECS, ASRI, Seoul National University,151-742, Seoul, Korea
minsucho@diehard.snu.ac.kr, kyoungmu@snu.ac.kr

Abstract

We present a novel object-specific segmentation method which can be used in view-based object recognition systems. Previous object segmentation approaches generate inexact results especially in partially occluded and cluttered environment because their top-down strategies fail to explain the details of various specific objects. On the contrary, our segmentation method efficiently exploits the information of the matched model views in view-based recognition because the aligned model view to the input image can serve as the best top-down cue for object segmentation. In this paper, we cast the problem of partially occluded object segmentation as that of labelling displacement and foreground status simultaneously for each pixel between the aligned model view and an input image. The problem is formulated by a maximum a posteriori Markov random field (MAP-MRF) model which minimizes a particular energy function. Our method overcomes complex occlusion and clutter and provides accurate segmentation boundaries by combining a bottom-up segmentation cue together. We demonstrate the efficiency and robustness of it by experimental results on various objects under occluded and cluttered environments.

1. Introduction

During recent years, image segmentation has been interleaved with object recognition by various approaches [15, 1, 14, 6, 7]. The main reason of the interaction is that without specific constraints to guide the attention, image segmentation problem is intrinsically ambiguous. Images in real world are fundamentally ambiguous and our perception of an image depends on a lot of complex factors. Therefore, segmentation by itself is an ill posed problem without means of guiding and judging the result. In that sense, combining segmentation and recognition is expected to make a well posed approach and to be of practical use because object recognition can provide a meaningful constraint for segmentation. Also, the approach can improve the performance of both recognition and segmentation since there is intertwined correlation between them.

1.1. Related Works

Previous object specific segmentation [15, 5] and object class category segmentation [1, 6, 7] approaches combine various top-down strategies with bottom-up segmentation approaches. Deformable Templates [16] and Active Appearance Models [3] can be used to object segmentation. But, they need a prior shape and a trained statistical model of the target object that requires laborious supervision. Moreover, they are very sensitive to partial occlusion. The recent works of object-specific segmentation were proposed by Yu and Xhi [15] and Ferrari *et al*. [5]. Yu and Xhi demonstrate that using graph theoretic framework a specific object can be detected and segregated from background. However, their results generate coarse boundaries in segmentation and cannot manipulate severe occlusion especially when occlusion splits the object. Ferrari *et al*. give a simultaneous object recognition and segmentation method which explores around initial local feature correspondences, resulting in dense correspondences over target objects. Although their method is robust and applicable even to deformable objects, it only gives rough segmentation by a set of blobs.

The latest researches of object segmentation are focused on object category or class rather than specific objects. Borenstein *et al*. [1] use a discrete set of possible low-level segmentations in order to minimize a cost function that includes a bias towards the holistic shape. Kumar *et al*. [6] in their OBJ CUT algorithm make use of a trained layered pictorial structure as top-down information for a graph cut energy minimization. Levin *et al*. [7] propose a more efficient learning framework that simultaneously takes into account low-level and high-level cues using Conditional Random Field formulations. And, Tu *et al*. [14] in their image parsing framework adopt AdaBoost object detection as a proposal distribution over possible segmentation for a data-driven Monte-Carlo sampling. They try to construct an image hierarchy including categorical level. All these ob-

1-4244-1180-7/07/$25.00 ©2007 IEEE

ject category segmentation approaches indeed improve the quality of achieved segmentations. But, they still generate coarse segmentation in occluded and cluttered environments.

1.2. View-Based Recognition as Top-Down Cue

The main problem of previous object segmentation approaches is that their top-down strategies fail to explain the details of various specific objects. However, if we have both the matched model view of the target object and the approximated pose of it in the input image, we can expect more exact segmentation by exploiting the very specific detailed information of the model view aligned to the input image. When it comes to a perspective of combining top-down and bottom-up framework, we can simply say that the model view aligned to the input image serves as the best top-down cue for object segmentation.

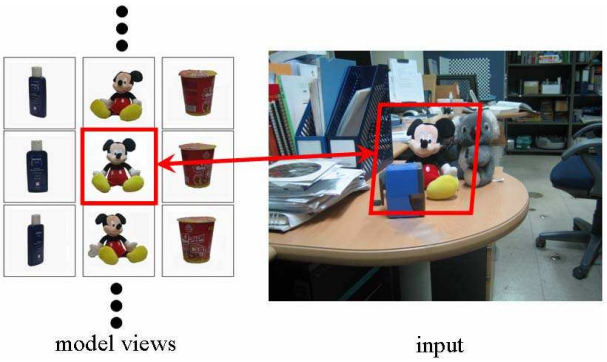
Our approach is supported by the fact that a variety of local invariant features bring about remarkable progress of view-based object recognition. View-based approaches using local features such as SIFT [8], Harris-Affine & Hessian-Affine [10], and MSER [9] now give a practical and robust solutions to automatic modelling and recognizing a wide range of 3D objects in cluttered and occluded environments at about real-time speed. In this paper, we propose the object segmentation method in thee view-based recognition system. The problem of segmentation is cast into that of labelling displacement vector and foreground status simultaneously for all pixels in input image. We formulate it by a MAP-MRF which minimizes a particular energy function. The inference is performed by loopy belief propagation algorithm in this paper.

This kind of method is primarily for object-specific segmentation rather than object 'category' segmentation. Even though object category segmentation is challenging problem and plays an important role, efficient object-specific segmentation still have significance and also can provide a cornerstone to object category segmentation.

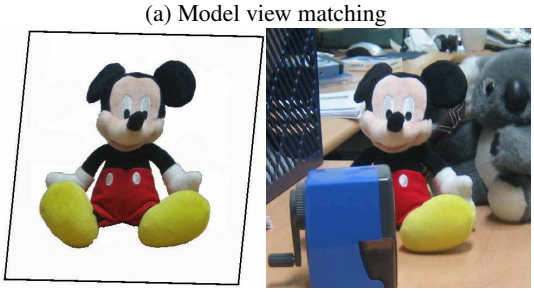
2. Problems of Segregation After Alignment

One could think that if the matched model view and the estimated pose are obtained by view-based object recognition, the segmentation of the target object is easily performed by simply measuring pixel-wise similarities after aligning the matched model view to an input image. However, in practice, it is still difficult to segregate the recognized object finely from the image due to several reasons.

For example, we can try to segment the target object by thresholding the color difference between corresponding pixels. First of all, given a set of model views, view-based object recognition is performed on an input image. In our experiment, we used Lowe's method with SIFT [8].



(a) Model view matching



(b) Aligned image pair

Figure 1. Object recognition result and the aligned image pair. (a) a red outlined region in the input is estimated pose of the matched model view. (b) the aligned image pair consists of the matched model view aligned to target object in the input (left) and the corresponding object part cropped from the input image (right).

Fig.1(a) illustrates a recognition result with given various model views and an input image in which the target object is cluttered and partially occluded. Fig.1(b) shows the aligned image pair obtained by transforming the matched model view to the input image.[1] Then, we calculate color dissimilarities for the corresponding pixels in the two images by measuring the distance in RGB color space. And, it is binarized with appropriate thresholds to obtain the object segmentation. The results are shown in Fig.2. In these results, we ignore the pixels falling into background of the aligned model view because there is no valid color information for segregation.

The results in Fig.2 show poor segregation between foreground and background. Partially, the illumination change and noise might have caused the simple color difference to be ineffective. However, those are not the main causes because illumination conditions are similar and the noise is negligible in this experiment. Considering the fact that most of errors occurred on the region around the boundaries of different colors, the poor results must be mainly due to alignment error. This alignment error partly comes

[1] In principle, our approach is compatible with any object recognition method which provides sufficient pose information for alignment of model view. We go on with this kind of aligned image pairs for proposed algorithm in the following sections

Figure 2. Segmentation results by binarization of RGB color difference. 4 results are selected among lots of results generated by various thresholds. In each pair, the left image describes pixels on the target object by red color and the segmented result is shown in the right. Lower right image shows zoomed part of bottom left which is the best one among them.

from numerical error and insufficient information for alignment (e.g. insufficient correspondences of local features). But, the more fundamental problem on the alignment is due to the difference of viewing direction to 3D object between the matched model view and the input image. In general, images of non-planar 3D object cannot be aligned exactly by any linear transformation unless the two images are obtained exactly from the same viewing angle.[2] In-depth rotation of non-planar 3D object makes non-linear transformation and some self-occlusions between corresponding pixels. The results in Fig.2 explain these intrinsic alignment problems of non-planar object.

3. Displacement-Foreground Labelling

The foregoing discussion suggests a conclusion that for finer object segmentation it is needed to estimate some non-linear transformation compensating the alignment error due to the difference of viewing direction and the like. And, in order to obtain good estimation, it is important that foreground labelling should be inferred simultaneously with the estimation of non-linear transformation in the process because they are innately correlated: if the foreground pixels are known exactly, the displacement between corresponding pixels can be solved with relative ease, and vise versa.

We consider estimating non-linear transformation between aligned image pair of non-planar 3D object as a kind of optical flow problem, in which foreground status of each pixel is estimated together for object segmentation. In this manner, we cast object segmentation problem into a labelling problem. This kind of formulation has been used in stereo correspondence problem. But, our method labels the displacement vector and foreground status simultaneously at each pixel of the input view with respect to the model view in the aligned image pair. This method is illustrated in Fig.3. Displacement denotes the vector from pixel position in the aligned model view to the corresponding position in the input image. Foreground status has a binary value representing whether the pixel is on foreground or not in the input image. Once we label correct displacement vector and foreground status on all the pixels in the input image, optimal object segmentation can be obtained by simply taking all the foreground pixels.

[2] In more detail, especially in case of Lowe's method that we used, it only estimates an affine transformation, not a perspective one, so even a planar object wouldn't be accurately aligned at all angles.

4. Proposed Algorithm

We formulate Bayesian displacement-foreground labelling in a MAP-MRF framework. MAP-MRF modelling and its variants have been successfully applied to various vision problems. Our formulation was inspired by MAP-MRF model for stereo matching problem in [13], and we extended and modeified it to the displacement-foreground labelling problem for object segmentation.

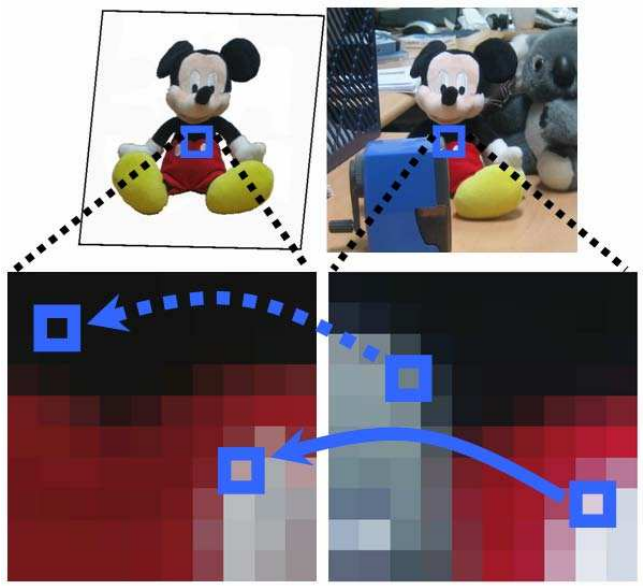


Figure 3. Displacement-foreground labelling. Given aligned image pair, we label displacement vector and foreground status for each pixel on the input image (upper right). In the bottom figures, some corresponding regions are zoomed for illustration. The arrows show the matched pixels by labelling the displacement of the pixels in the input image (lower right). The dotted arrow represents that the pixel in the input is matched to a background pixel or invisible pixel in the model view. While, the solid arrow means a foreground or visible corresponding pixel.

We model the displacement-foreground labelling by a coupled MRF: the observation nodes of our MRF model correspond to the image pixel lattice of the input image in the given aligned image pair. One observation node has the some color information from the given aligned image pair I, which is needed for estimating the displacement of the pixel at the node position. D is the displacement vector nodes defined on the observation nodes, and F is the foreground status nodes also defined on them. Fig.4 illustrates this coupled MRF. D consists of 2-dimensional vector values for each pixel. In F, each foreground status has the value 1 for foreground, 0 for background.

Using Bayes' rule, the joint posterior probability over D and F given an aligned image pair I can be factorized as:

$$P(D, F|I) = \frac{P(I|D, F)P(D, F)}{P(I)} \tag{1}$$

4.1. Likelihood Term

Assuming that the image noise follows an independent identical distribution and S denotes the set of pixels in the input image of the aligned image pair and, we can define the likelihood as follows:

$$P(I|D, F) \propto \prod_{s \in S} \exp\left(-\phi(s, d_s, f_s, I)\right). \tag{2}$$

where ϕ is an evidence function which relates how compatible a displacement vector d_s and foreground status f_s at pixel s is with the color information observed in the aligned image pair I. Our likelihood estimates color dissimilarity of corresponding pixels considering not only displacement but also foreground status of the pixel s because the likelihood of background pixels should not be determined in the same way of foreground pixels.

We define, the evidence function ϕ as follows.

$$\phi(s, d_s, f_s, I) = \sigma_\phi L(s, d_s, I) f_s + K_\phi(1 - f_s), \tag{3}$$

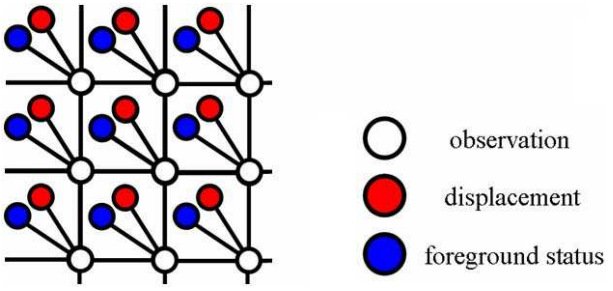


Figure 4. MRF for displacement-foreground labelling. Two latent nodes, displacement and foreground status are connected to each observation node. Observation nodes correspond to the input image pixels in the given aligned image pair.

where $L(s, d_s, I)$ is the matching cost function of the pixel s with displacement d_s given the aligned image pair I. If the pixel is on background (f_s = 0), ϕ ignores the matching cost and adopts a constant penalty energy K_ϕ which prevents the background regions from prevailing.

4.2. Prior Term

The Markov random field has the property that the conditional probability of a site in the field depends only on its neighboring sites. Assuming N are the edges in the four-connected image grid graph, the prior probability $P(D, F)$ can be expanded as:

$$P(D, F) = \prod_{(s,t) \in N} \exp\left(-\psi(d_s, f_s, d_t, f_t)\right), \tag{4}$$

where ψ is the compatibility function of pixels s and t , d_s and d_t are the displacement vectors, and f_s and f_t is the binary foreground status of the neighboring pixels s and t. To enforce spatial interactions between s and t, we define ψ as follows:

$$\psi(d_s, f_s, d_t, f_t) = \gamma(d_s, d_t) + K_\psi \delta(f_s - f_t) + \rho(f_s, f_t), \tag{5}$$

where the first term γ penalizes the difference of displacement assignments of neighboring pixels. The penalty means the assumption that the neighboring pixels have similar displacements. In other words, the object is assumed to be smooth. The second term imposes constant penalty K_ψ for the change of foreground status so that foreground regions and background regions tend to be clustered. The third term ρ encodes region similarities incorporating bottom-up segmentation result of the input image into our model. It penalizes the occurrence of boundaries in homogeneous regions of the bottom-up segmentation. This bottom-up segmentation cue encourages the boundaries to be placed along bottom-up segmentation boundaries as possible.

We define γ function as a simple version of robust function used in [13]. Thus, we define γ by

$$\gamma(d_s, d_t) = \min(\sigma_\gamma |d_s - d_t|, K_\gamma), \tag{6}$$

where we model displacement discontinuity implicitly by adopting the upper bound of penalty K_γ because abrupt change of displacement can occur at object edges in non-planar object. Therefore, if the difference of neighboring displacements exceeds K_γ, γ regards it as displace discontinuity implicitly and imposes the constant penalty which controls the excess of the discontinuities.

Function ρ incorporating the bottom-up segmentation cue is expressed as

$$\rho(f_s, f_t) = \begin{cases} K_\rho & if\ f_s \neq f_t\ and\ seg(s) = seg(t) \\ 0 & otherwise \end{cases} \tag{7}$$

where $seg(s)$ denotes the label of the bottom-up segmentation result at pixel s. This function penalizes the change of foreground status between two neighboring pixels when the pixels are in the same segmented region. For this purpose, segmentation cue needs to be highly over-segmented enough not to miss the boundary of the target object. But, any bottom-up method can be used for this cue.

4.3. Inference Algorithm

Integrating all the detailed terms in eqs. (3), (5)-(7) into the likelihood in eq.(2) and the prior in eq.(4), our Bayesian displacement-foreground labelling model in eq.(1) is established on MAP-MRF. There are several inference algorithms to solve this MAP-MRF such as belief propagation, graph cut, MCMC sampling and so on. In this paper, we solve it efficiently by loopy belief propagation algorithm. Belief propagation is an iterative inference algorithm that propagates messages in Bayesian networks, which can obtain an exact solution of directed acyclic graph. To find the approximation of MAP solution of MRF with network loops, the loopy belief propagation [13] ignores the existence of loops in the networks. It has been applied successfully for the solutions of various MAP-MRF problems [11]. Moreover, the efficient method [4] for belief propagation substantially improves running time for inference. We use the efficient loopy belief propagation algorithm to search for the optimal labels of displacement and foreground status in our implementation. [3]

The inference process assigns individual energy for each label in all the nodes in the proposed combined MRF model, which can be actually interpreted as a scaled probability. In that sense, our final decision for foreground status can be interpreted as a probabilistic approach integrating bottom-up and top-down cues, and consider contextual plausibility in foreground segregation rather than the method used in section 2.

5. Experimental Results

The proposed algorithm has been tested on several aligned image pairs obtained by Lowe's object recognition method [8]. For the pixel matching cost function L of eq.(3), we simply use the Euclidean distance in RGB color space. Maximum length of displacement vector was set to 5 pixels long as we assumed that the initial alignment error is not so big. The highly over-segmentation result of mean shift color segmentation method [2] was used as the bottom-up segmentation cue in function ρ of eq.(7). The parameters of the proposed algorithm in our experiments are: σ_ϕ=0.07, K_ϕ=3.65, K_ψ=6.75, σ_γ=0.72, K_γ=25, K_ρ=30.

[3] Note that in the respect of optimization rather than running time, graph cut might produce better results than loopy belief propagation as D. Scharstein and R. Szeliski have shown for stereo matching [12].

The results of our experiments are displayed with four ways: (i) the displacement field (ii) the foreground status field (iii) segmented object (iv) object image synthesized from the model view using optimal displacement field.

Fig.5 presents the results obtained by our proposed algorithm to the $Mickey$ image appeared in section 2. The algorithm was implemented in C++ and the time needed for the segmentation was about 2 minutes on a machine with Pentium IV 2.4GHz CPU and 2.0GB RAM with 64 iterations of belief propagation. As shown in Fig.5 (f) the segmentation result is quite satisfactory despite clutter and occlusions. Our segmentation corrects the naive pixel correspondences of the previous binarization method in Fig.2 and makes accurate segmentation boundaries integrating a bottom-up segmentation cue. From the synthesized image (g), we can identify the displacements are well approximated because the synthesized object image using the labelled displacements has a strong resemblance to the object in the input image.

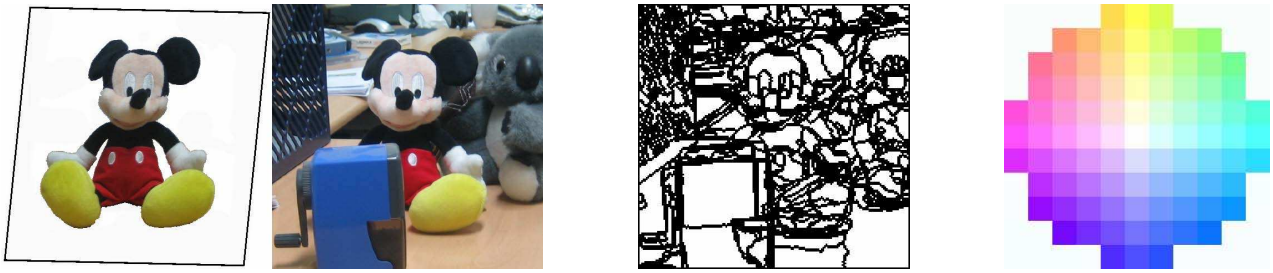
Note that our algorithm detects and includes small invisible parts due to self-occlusion into the final segmentation in several parts of our test images. For example, the self-occluded region around the $Mickey$'s nose is correctly segregated unlike the result in Fig.2. Actually, in our formulation, a local part that is invisible in the model view can be incorporated into the segmentation if it is close to visible part with similar color. This is because the pixels in the input are matched to pixels in the model according to the proximity of color, distance, and bottom-up segmentation boundaries.

In Fig.6, we demonstrate the performance on more target objects and input images. The results show the proposed algorithm is robust to severe occlusions which split the objects as shown in the $phone$ and $drill$ and $koala$ examples. It is because the detailed appearance and pose information from the aligned model view helps our segmentation algorithm to segregate topologically complex occlusion and clutter with estimated displacements.

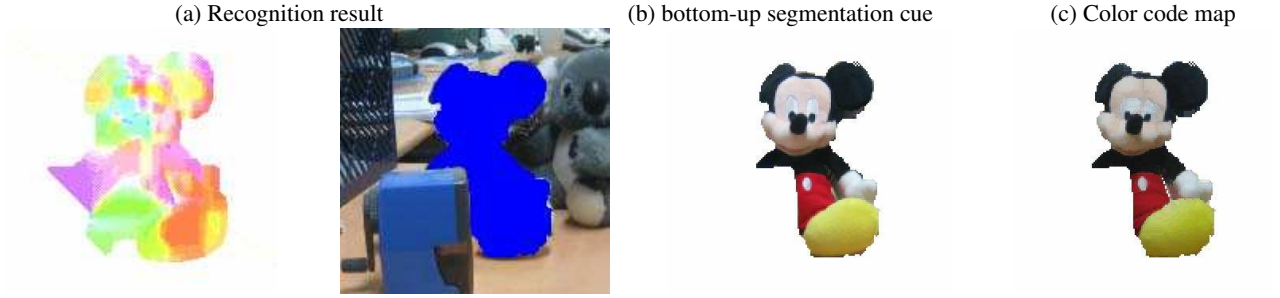
Some results show our method is relatively sensitive to color variation. When illumination changes, foreground part of the object can be labelled background as shown in the $ramenbox$ example. In cup example, some part in clutter is incorrectly labelled foreground because the part is very similar with the model in color and position. The shadow of cluttering objects in $drill$ example disturbs the boundaries of segmentation around it. However, considering the complexity of occlusion, clutter, and our simple color similarity measure, the segmentation results of these examples demonstrate the power of our formulation.

6. Conclusion and Future Work

We presented a new approach to object-specific segmentation which exploits the information of the model view



(a) Recognition result (b) bottom-up segmentation cue (c) Color code map



(d) Displacement (e) Foreground (f) Segmentation (g) Synthesis

Figure 5. Segmentation of the $Mickey$ image. (a) aligned image pair. (b) over-segmented bottom-up segmentation cue. (c) color code map where the color represents the orientation and magnitude of the displacement vector from the input to the model view in the aligned image pair. (d) displacement field shown in color coded fashion. (e) foreground status field covered on the input where blue pixels represent foreground. (f) segmented object by the foreground status field (g) synthesized object image where encodes matched pixel colors of the model view using the displacement field.

aligned to an input image in view-based recognition. We demonstrated a difficult point in this kind of object segmentation, which cannot be solved using a naive approach of binarizing the color differences between the aligned images. Thus, the segmentation problem is reformulated into displacement-foreground labelling problem, in which both displacement and foreground status are simultaneously estimated for some non-linear transformation mainly caused by the difference of viewing direction in two views of non-planar 3D target object. We use the matched model view and pose as top-down cue for object-specific segmentation in our MAP-MRF framework. Then, Belief propagation was used for inference of the solution. Finally, the experiments demonstrate that our method is able to obtain good segmentation despite severe occlusion and clutter.

In contrast to previous methods, our method provides strong robustness to partial occlusion and clutter. But, it is relatively sensitive to color variation (e.g. illumination change or shadow) because our method uses simple pixel-based measure of color similarity. Also, severe view-point change and deformation can distract our method. Therefore, in our future work, we will extend our algorithm to have more robustness to color variation and shape deformation of the target object. For illumination change, we can employ an adaptive parameter setting scheme that exploit the color distributions around initially matched local features.

Acknowledgements

This work was supported in part by the ITRC program by Ministry of Information and Communication and in part by the Agency for Defense Development, through the Image Information Research Center, at KAIST, Korea.

References

[1] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *Workshop on Perceptual Organization in Computer Vision*, page 46, 2004.

[2] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell*, 24(5):603–619, 2002.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54, 2006.

[5] V. Ferrari, T. Tuytelaars, and L. J. V. Gool. Simultaneous object recognition and segmentation by image exploration. In *European Conference on Computer Vision*, pages Vol I: 40–54, 2004.

[6] M. P. Kumar, P. H. S. Torr, and A. Zisserman. OBJ CUT. In *CVPR*, pages 18–25. IEEE Computer Society, 2005.

[7] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In *European Conference on Computer Vision*, pages IV: 581–594, 2006.

Figure 6. Results for 6 examples from the first row; *book*, *phone*, *ramenbox*, *cup*, *koala*, and *drill*.

[8] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, Sept. 2004.

[10] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[11] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 467–475, July 1999.

[12] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[13] J. Sun, N.-N. Zheng, and H.-Y. Shum. Stereo matching using belief propagation. *IEEE Trans. Pattern Anal. Mach. Intell*, 25(7):787–800, 2003.

[14] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. In *ICCV*, pages 18–25. IEEE Computer Society, 2003.

[15] S. X. Yu and J. Shi. Object-specific figure-ground segregation. In *CVPR*, pages 39–45. IEEE Computer Society, 2003.

[16] A. L. Yuille, D. S. Cohen, and P. W. Hallinan. Feature extraction from faces using deformable templates. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 104–109, San Diego, June 4-8 1989.