# Propagating Multi-class Pixel Labels throughout Video Frames

Albert Y. C. Chen and Jason J. Corso
Computer Science and Engineering
SUNY at Buffalo
{aychen,jcorso}@buffalo.edu

## Abstract

*The effective propagation of pixel labels through the spatial and temporal domains is vital to many computer vision and multimedia problems, yet little attention have been paid to the temporal/video domain propagation in the past. Previous video label propagation algorithms largely avoided the use of dense optical flow estimation due to their computational costs and inaccuracies, and relied heavily on complex (and slower) appearance models. We show in this paper the limitations of pure motion and appearance based propagation methods alone, especially the fact that their performances vary on different type of videos. We propose a probabilistic framework that estimates the reliability of the sources and automatically adjusts the weights between them. Our experiments show that the "dragging effect" of pure optical-flow-based methods are effectively avoided, while the problems of pure appearance-based methods such the large intra-class variance is also effectively handled.*

## 1. Introduction

Pixel labels have a great number of uses in the computer vision and multimedia community. For example, the labels are the disparity values in stereo vision, grayscale or color values in image denoising [11], and $\alpha$ values in interactive segmentation problems [7]. Since manually labeling every pixel in an image is highly impractical, many research have been conducted in propagating pixel labels throughout both the spatial and temporal domain [13, 7, 12, 2]. In interactive image segmentation tasks, a small number of manually annotated pixel labels are propagated to the remaining pixels to produce a foreground/background map for the whole image. Attempts are also made to propagate labels in the temporal domain for labeling video objects efficiently.

The problem of propagating pixel labels throughout video frames seems deceivingly easy at first glance: for any pixel $z^{t+1}_{\mathbf{n}'}$ in frame $t+1$, find the optical flow $\mathbf{m}^t$ from $z^t_{\mathbf{n}}$ to $z^{t+1}_{\mathbf{n}'}$ ($\mathbf{n}' = \mathbf{n} + \mathbf{m}^t$), and let $z^{t+1}_{\mathbf{n}'}$ take the same label as $z^t_{\mathbf{n}}$. A simple experiment on the commonly used *garden* sequence shows otherwise (Fig. 1). *Holes* (pixels with undetermined labels) form because the correspondence established by the forward flow between $z^t$ and $z^{t+1}$ is neither *one-to-one* (injective) nor *onto* (surjective) (Fig. 2). With reverse-flow-based propagation, the *dragging* effect occur because a reappearing (previously occluded) pixel $z^{t+1}_{\mathbf{n}}$ is forced to take some un-correlated $z^t_{\mathbf{n}'}$'s label by the reverse flow, while in theory it has no corresponding $z^t_{\mathbf{n}'}$.

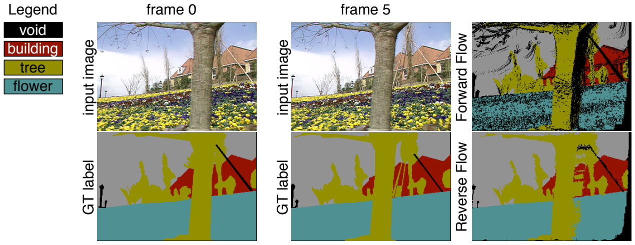

Figure 1. An example of why optical flow alone can't solve the video pixel label propagation problem: *holes* form with forward flow and the *dragging* effect plagues the reverse flow.

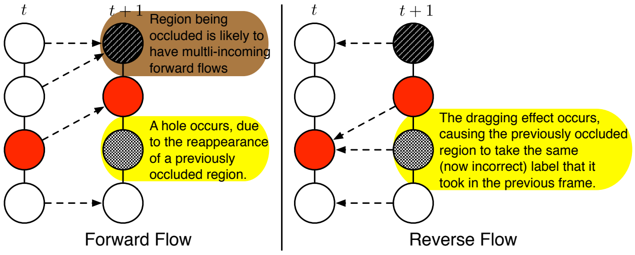
$t$
$t+1$
Region being occluded is likely to have multi-incoming forward flows
A hole occurs, due to the reappearance of a previously occluded region.
Forward Flow
$t$
$t+1$
The dragging effect occurs, causing the previously occluded region to take the same (now incorrect) label that it took in the previous frame.
Reverse Flow

Figure 2. The *holes* in forward flows and the *dragging* effect in reverse flows. When the colored ball moves, the (lightly shaded) region it left behind have no incoming forward flows or a incorrect outgoing reverse flow. The (heavily shaded) region being occluded by this motion frequently have multiple incoming forward flows.

The aforementioned issues have been commonly treated as the results of inaccurate optical flows, and many algorithms following [7] have shunned optical flows for label propagation. Wang and Cohen [12] propagate the labels of a small subset of static pixels from $t$ to $t+1$, then perform spatial propagation on frame $t+1$ with BP [13]. Bai and Sapiro [2] treat the video as a space-time volume and propagate labels via the shortest geodesic distance, which is defined on local color gradients. These methods, although

The authors are grateful for the financial support provided in part by NSF CAREER IIS-0845282, NSF CNS-0855220, DARPA HR0011-09-1-0022, and DARPA/ARL Mind's Eye W911NF-10-2-0062

978-1-4244-9300-5/10/$26.00 ©2010 IEEE

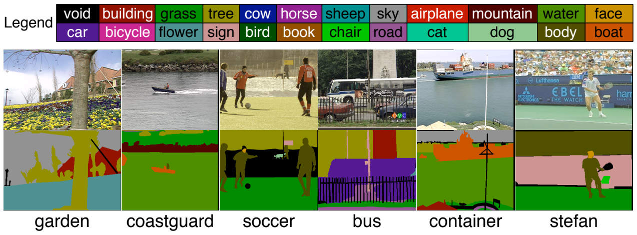

Figure 3. A snapshot of our multi-class pixel-wise annotation of the commonly used sequences collected at xiph.org. We follow the 24-class MSRC semantic class labels [9] for the annotation.

effective for interactive segmentation tasks, are not capable of handling occlusion and reappearance of objects at all. To address this issue, Criminisi et. al. [8] utilized a CRF with 2nd order HMMs to facilitate the correct labeling of reappearing objects in a foreground/background segmentation problem; however, its generalizability to multi-class labeling remains unknown. Recent approaches favor using complex appearance models for propagation, such as the local shape models in [3] and coupled HMM in [1].

These counterintuitive findings made us question: is optical flow-based propagation that bad? If not, when are they reliable for label propagation, and when would we require additional help? Upon close examination, we also noticed that previous conclusions are drawn from experiments performed on a small biased set of videos—videos with large foreground objects for two-class propagation and driving videos for multi-class propagation. In order to fairly compare the results, we build a larger and less-biased multi-class pixel-wise labeled dataset, which we will discuss in Sec. 2. We experiment and discuss the results of pure motion and appearance based methods in Sec. 3, followed by our analysis and design of the optical-flow trustworthiness metric and our label propagation framework in Sec. 4.3. We discuss the comparative results in Sec. 4.4 and conclude in Sec. 5.

## 2. Our new multi-class pixel label dataset

Previous datasets used for evaluating video pixel label propagation are small and biased. The interactive segmentation society focused only on 2-class (foreground, background) propagation [7, 12, 2, 8], where the foreground object tend to occupy a larger area of the scene. The only multi-class pixel-wise labeled ground truth dataset available as of now are all, by coincidence, driving sequences[6, 1]. This type of video consists mostly of objects moving from the vanishing point of the road towards the sides.

Instead of creating a dataset ourselves, which might be biased as well, we decide to adapt videos commonly used by the community collected at xiph.org. Samples are shown in Fig. 3, which includes well known sequences such as *garden* and *coastguard*. These videos cover a wider spectrum of possible camera and object movements, and is vital to fully inspect when and where optical flow alone is sufficient. For example, the camera is fixed and only the objects are moving in *container*, while the camera is moving and most objects are static in *garden*. In a few other sequences, such as *coastguard* and*stephan*, not only is the camera moving but also multiple objects in the scene are moving as well.

## 3. Motion v.s. Appearance based Propagation

### 3.1. Motion alone

In general, there are two ways of using optical flows to assign a pixel $z_{\mathbf{n}}^{t+1}$ in frame $t+1$ with a label from frame $t$: forward flow from $z_{\mathbf{n}'}^{t}$ to $z_{\mathbf{n}}^{t+1}$ represented as $f_{\text{fwd}}(z_{\mathbf{n}'}^{t}) = z_{\mathbf{n}}^{t+1}$, or reverse flow from $z_{\mathbf{n}}^{t+1}$ to $z_{\mathbf{n}''}^{t}$ represented as $f_{\text{rvs}}(z_{\mathbf{n}}^{t+1}) = z_{\mathbf{n}''}^{t}$. Since even the latest optical flow estimation methods are not guaranteed to solve occlusion and reappearance situations perfectly (as shown in Fig. 2), we use the classical Black and Anandon method [5] due to its efficiency and relative effectiveness [4].

The task of propagating labels with forward flows alone is to determine the proper label $L(\cdot)$ for all $z_{\mathbf{n}}^{t+1}$ (collectively represented as $\mathbf{z}^{t+1}$) by using:

$$L(z_{\mathbf{n}}^{t+1}) := L(z_{\mathbf{n}'}^{t}) \quad \text{where} \quad f_{\text{fwd}}(z_{\mathbf{n}'}^{t}) = z_{\mathbf{n}}^{t+1} \tag{1}$$

For the task of propagating labels with the reverse flows,

$$L(z_{\mathbf{n}}^{t+1}) := L(z_{\mathbf{n}'}^{t}) \quad \text{where} \quad f_{\text{rvs}}(z_{\mathbf{n}}^{t+1}) = z_{\mathbf{n}'}^{t} \ . \tag{2}$$

The forward and inverse flow functions $f_{\text{fwd}}(\cdot)$, $f_{\text{rvs}}(\cdot)$ are both non-injective and non-surjective. Deciding $L_{\mathbf{n}}^{t+1}$ (short-hand notation for $L(z_{\mathbf{n}}^{t+1})$) with $f_{\text{rvs}}(\cdot)$ is straightforward since $z_{\mathbf{n}}^{t+1}$ is in the *domain* and is guaranteed to have a corresponding $L_{\mathbf{n}'}^{t}$. Determining $L_{\mathbf{n}}^{t+1}$ with $f_{\text{fwd}}(\cdot)$ is trickier, since $z_{\mathbf{n}}^{t+1}$ is in the *codomain* of a non-injective/surjective function; additional information is needed to determine the appropriate $L_{\mathbf{n}}^{t+1}$ when there are zero or multiple corresponding $Lz_{\mathbf{n}'}^{t}$.

### 3.2. Appearance Model alone

A simple CIE-Lab color space based non-parametric appearance model is learned for every label $\mathcal{L} = \{l_a, l_b, ...\}$ we wish to propagate in the first frame. For the following frames, we extract the color distribution of the subimage $s(\cdot)$ centered at $z_{\mathbf{n}}^{t+1}$ to determine the most likely label $L(\cdot)$:

$$P\left(s(z_{\mathbf{n}}^{t+1}) \mid L(z_{\mathbf{n}}^{t+1})\right) = 1/d(H_s, H_l) \ , \tag{3}$$

where $L(z_{\mathbf{n}}^{t+1}) \in \mathcal{L}$ and a simple *Intersection* measure [10] is used to compute the distance between $H_s$ and $H_l$.

### 3.3. Experiments, Results and Discussion

Experiment results on using the motion or appearance model alone are quite conflicting: instead of having one constantly outperform the other, the numbers varied widely from video to video. Videos with large regions of frequent occlusion/reappearance result in extremely poor performance for optical-flow-based propagation methods, such

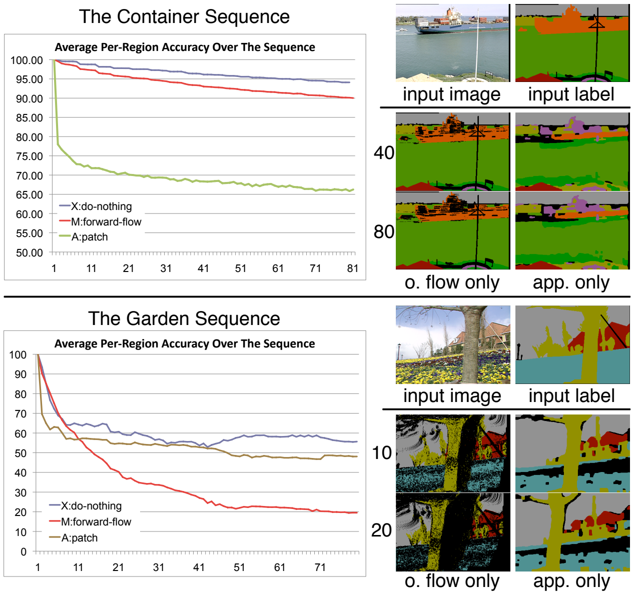

Figure 4. Results from using only motion or appearance model.

as the *garden* sequence shown in Fig. 4. Videos with less occlusion/reappearance and where the appearance of objects are multi-modal cause appearance model to perform worse, such as the *container* sequence shown in Fig. 4.

These findings have urged us to develop measures for determining the reliability of individual optical flows for the label propagation task. We develop and adapt these measures into a probabilistic framework as discussed as follows.

# 4. Our Proposed Method

## 4.1. The Probabilistic Pixel Labeling Model

We use a probabilistic framework to jointly optimize the clues we obtain from the optical flows, appearance models, and prior knowledge such as the spatial smoothness constraint. The label propagation task becomes a problem of determining the optimum set of labels $\mathbf{L}^{t+1}$ for all the pixels $\mathbf{z}^{t+1}$ in frame $t+1$:

$$E\left(\mathbf{L}^{t+1}|\mathbf{L}^{t},\mathbf{z}^{t+1},\mathbf{z}^{t}\right) = U^{\mathrm{M}}(\mathbf{L}^{t+1},\mathbf{L}^{t},\mathbf{z}^{t+1},\mathbf{z}^{t}) + \lambda_1 U^{\mathrm{C}}(\mathbf{L}^{t+1},\mathbf{Z}) + \lambda_2 V^{\mathrm{S}}(\mathbf{L}^{t+1},\mathbf{z}^{t+1}) \ . \tag{4}$$

where $\mathbf{Z}$ is the collection of all frames we've seen so far.

The weights $\lambda_1$ and $\lambda_2$ are typically estimated during training and fixed afterwards; our proposed location-varying, flow reliability-dependent weights $\lambda_1$ are developed later in Sec. 4.3. We discuss the individual energy terms in the following sub-section.

## 4.2. The Individual Energy Terms

**The Motion Evidence Term** $U^{\mathrm{M}}(\cdot)$ generalizes Eq. 2 to deal with situations where multiple incoming flows are present. The idea is to measure the "confidence" level of the all flows, either intrinsically during the flow computation process via the error measure, or extrinsically by calculating the cross-correlation between the two regions where the flow originates and terminates. We choose the label $L^{t+1}_{\mathbf{n}} := L^{t}_{\mathbf{n}'}$ that maximizes the overall confidence over all flows, or equivalently, minimizes the energy function defined over the flows. We associate the Pott's model with spatially-varying per pairing weights $w(\cdot)$ to discount the penalty given to the more confident flows:

$$U^{\mathrm{M}}(\mathbf{L}^{t+1},\mathbf{L}^{t},\mathbf{z}^{t+1},\mathbf{z}^{t}) = \sum_{\mathbf{n}} \sum_{\mathbf{n}' \ | z^{t}_{\mathbf{n}'} \in f(z^{t+1}_{\mathbf{n}})} w(z^{t+1}_{\mathbf{n}}, z^{t}_{\mathbf{n}'}) \left(1-\delta\left(L^{t+1}_{\mathbf{n}}, L^{t}_{\mathbf{n}'}\right)\right) \ , \tag{5}$$

where $\delta$ is the Kronecker delta and $f(z^{t+1}_{\mathbf{n}})$ is the set of pixels in $\mathbf{z}^{t}$ that are associated with $z^{t+1}_{\mathbf{n}}$ by $f_{\mathrm{fwd}}(\cdot)$ and $f_{\mathrm{rvs}}(\cdot)$. The spatially-varying weights are defined such that

$$w(z^{t+1}_{\mathbf{n}}, z^{t}_{\mathbf{n}'}) \propto ||s(z^{t+1}_{\mathbf{n}}) - s(z^{t}_{\mathbf{n}'})|| \ , \tag{6}$$

where again $s(z^{t+1}_{\mathbf{n}})$ is the local sub-window centered at $z^{t+1}_{\mathbf{n}}$ and $||\cdot||$ is the distance (we use K-L divergence) between the histograms of the two patches $s(z^{t+1}_{\mathbf{n}})$ and $s(z^{t}_{\mathbf{n}'})$.

**The Appearance Likelihood Term** $U^{\mathrm{C}}(\cdot)$ in Eq. 4 determines how likely a pixel $z^{t}_{\mathbf{n}}$ was generated by one of the label classes. We use the appearance model defined in Sec. 3.2 (except that a model is now learned for every coherent segment) and the energy is defined as:

$$U^{\mathrm{C}}(\mathbf{L}^{t+1},\mathbf{z}^{\mathbf{t+1}}) = -\sum_{\mathbf{n}} \log P(s(z^{t+1}_{\mathbf{n}})|L^{t+1}_{\mathbf{n}}) \ . \tag{7}$$

$P(s(z^{t+1}_{\mathbf{n}})|L^{t+1}_{\mathbf{n}})$ is defined in (3). The appearance likelihood models are updated after every new frame is labeled.

**The Spatial Continuity Term** $V^{\mathrm{S}}(\mathbf{L}^{t},\mathbf{z}^{t})$ is defined as:

$$V^{\mathrm{S}}(\mathbf{L}^{t+1},\mathbf{z}^{t+1}) = \sum_{(\mathbf{n},\mathbf{n}')\in\mathbf{C}} w(z^{t+1}_{\mathbf{n}}, z^{t+1}_{\mathbf{n}'}) \left(1-\delta\left(L^{t+1}_{\mathbf{n}}, L^{t+1}_{\mathbf{n}'}\right)\right) \ , \tag{8}$$

where $\mathbf{C}$ is the set of all neighboring pairs of pixels, $\mu$ is the contrast parameter set to $\mu = (2\langle\, ||z^{t}_{\mathbf{n}} - z^{t}_{\mathbf{m}}||^2 \,\rangle)^{-1}$, where $\langle\cdot\rangle$ is the expectation over all neighbor pairs in an image.

## 4.3. Estimating the Reliability of Optical Flows and Defining our Reliability-driven Weights

Figure 5 (a-c) shows examples where a combination of forward and reverse flows can be used to estimate when occlusion and reappearance are occurring. If $z^{t+1}_{\mathbf{n}}$ belongs to a reappearing region, there would likely be no good match for it in $\mathbf{z}^{t}$, therefore $f_{\mathrm{rvs}}(z^{t+1}_{\mathbf{n}})$ is likely to be different from those $z^{t}_{\mathbf{n}'}$ where $f_{\mathrm{fwd}}(z^{t}_{\mathbf{n}'}) = z^{t+1}_{\mathbf{n}}$, as shown in (b) and (c). When $z^{t}_{\mathbf{n}'}$ is being occluded, there would likely be no $f_{\mathrm{rvs}}(\cdot) = z^{t}_{\mathbf{n}'}$, resulting in phenomena similar to (b) and

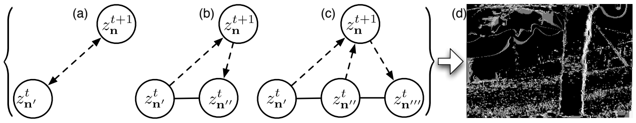

Figure 5. Flow reliability estimation

(c). Contrarily, when $f_{\text{rvs}}(z_{\mathbf{n}}^{t+1}) = z_{\mathbf{n}'}^{t}$ where $f_{\text{fwd}}(z_{\mathbf{n}'}^{t}) = z_{\mathbf{n}}^{t+1}$ as in (a), the flows are likely to be more reliable.

Based on these observations, we derive a simple yet effective measure for determining location-varying flow-depent $\lambda_1$'s for every pixel $z_{\mathbf{n}}^{t+1}$, represented as $\lambda_1(z_{\mathbf{n}}^{t+1})$:

$$\lambda_1(z_{\mathbf{n}}^{t+1}) = w' \sum_{\mathbf{n}' \; |z_{\mathbf{n}'}^{t} \in f(z_{\mathbf{n}}^{t+1})} ||f_{\text{rvs}}(z_{\mathbf{n}}^{t+1}) - z_{\mathbf{n}'}^{t}|| \qquad (9)$$

where $w'$ is the normalizing weight estimated from the average $\lambda_1(\cdot)$'s. The weight map formed by all $\lambda_1(z_{\mathbf{n}}^{t+1})$'s is shown in Fig. 5 (d). Brighter-colored regions represent larger $\lambda_1(\cdot)$. The area on the right of the tree where previously occluded regions are reappearing gives $U^C(\cdot)$ higher weight, since no reliable $f_{\text{fwd}}(\cdot)$, $f_{\text{rvs}}(\cdot)$ exist. Similarly, the area around the pole on the right and the boundaries of the tree branches on the left also rely more on $U^C(\cdot)$.

### 4.4. Experiments, Results and Discussions

We use one of the standard energy minimization methods, the Graph-Cuts Expansion as in [11], to obtain $\mathbf{L}^{t+1}$ at each iteration. Figure 6 and 7 shows that our proposed method properly weighs between multiple sources of information and constraints and achieves quite a significant improvement. In the *garden* sequence, the rapidly moving tree trunk with large regions of occlusion/reappearance causes optical flow based methods to drag on and propagate the error, while our proposed method properly fills in the gap with the appearance information. The pure appearance model is prone to intra-class variances, and the upper region of the flowers in the garden sequence being wrongly assigned the *void* label. In the container sequence, the large intra-class variance causes the appearance model to incorrectly assign assign *road* and *tree* to the upper part of the container; our method properly filled in the region with motion clues.

## 5. Conclusion

We showed the issues of pure motion and appearance based video pixel label propagation methods, and proposed a probabilistic framework that estimates the reliability of motion and appearance information then automatically weigh between them. Our experiments show that the "dragging effect" of pure optical-flow-based methods are effectively avoided, while the weakness of appearance-based methods such the as large intra-class-variance is also effectively handled.

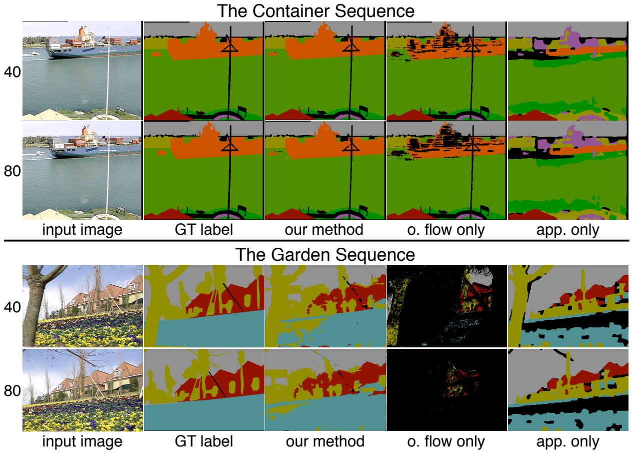

Figure 6. Qualitative comparison of the propagation results.

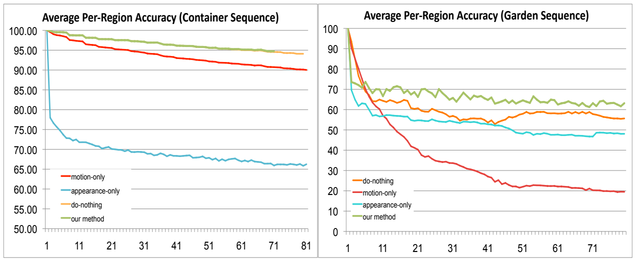

Figure 7. Quantitative comparison of the propagation results.

## References

[1] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *Proc. of CVPR*, 2010.

[2] X. Bai and G. Sapiro. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *IJCV*, 2009.

[3] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video SnapCut: robust video object cutout using localized classifiers. In *ACM SIGGRAPH*, 2009.

[4] S. Baker, S. Roth, D. Scharstein, M. Black, J. Lewis, and R. Szeliski. A database and evaluation methodology for optical flow. In *Proc. of ICCV*, 2007.

[5] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 1996.

[6] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *PR Letters*, 2009.

[7] Y. Chuang, A. Agarwala, B. Curless, D. Salesin, and R. Szeliski. Video matting of complex scenes. In *ACM SIGGRAPH*, 2002.

[8] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov. Bilayer Segmentation of Live Video. In *Proc. of CVPR*, 2006.

[9] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. of ECCV*, 2006.

[10] M. Swain and D. Ballard. Indexing via color histograms. In *Proc. of ICCV*, 1990.

[11] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. *IEEE PAMI*, pages 1068–1080, 2008.

[12] J. Wang and M. Cohen. An iterative optimization approach for unified image segmentation and matting. In *Proc. of ICCV*, 2005.

[13] J. Yedidia, W. Freeman, and Y. Weiss. Generalized Belief Propagation. In *NIPS*, volume 13, pages 689–695, 2000.