Segmentation by Combining Parametric Optical Flow with a Color Model

Adrian Ulges
Department of Computer Science, Technical University Kaiserslautern
a_ulges@informatik.uni-kl.de

Thomas M. Breuel
Technical University Kaiserslautern and German Research Center for Artificial Intelligence (DFKI)
breuel@iupr.dfki.de

Abstract

We present a simple but efficient model for object segmentation in video scenes that integrates motion and color information in a joint probabilistic framework. Optical flow is modeled using parametric motion with Gaussian noise. The color distribution of foreground and background is described by histograms or Gaussian mixture models. Optimization is carried out using an efficient graph cut algorithm.

In quantitative experiments on a variety of video data, we demonstrate that the proposed approach leads to significant reductions in error rates compared to a state-of-the-art motion-only segmentation.

1. Introduction

Segmentation has been an intensively studied problem in computer vision for decades, but yet continues to be a challenge [8]. For still images, pixels are usually grouped into regions of coherent appearance [9, 11] that cannot be assumed to correspond to meaningful objects in general.

For video, an attractive alternative is to segment frames into *layers* of coherent motion. The resulting regions can be associated with rigid objects (or rigid components of articulated objects). However, the estimation and grouping of motion is based on assumptions that are often violated in practice [1]: (1) the color of moving pixels is assumed to be constant, (2) the reliable estimation of motion demands the presence of texture, (3) spatial coherence of motion is assumed, which is violated at motion boundaries, and (4) the foreground motion must be discriminable from the background motion. Though workarounds for some problems exist, motion segmentation remains a challenging problem.

While motion and color clues individually do not provide reliable segmentations, we present a simple and efficient combination of both. Our approach decomposes a scene into a foreground and background layer. For each layer, a parametric motion model and a Gaussian error model are assumed, and the color distribution is modeled by histograms or Gaussian mixture models. Cost terms are formulated for assigning each pixel to foreground or background depending on how well it fits the associated motion and color. The pixel costs are integrated with a smoothness prior, and optimization is carried out using a graph cut algorithm.

2. Related Work

Starting with the work by Horn and Schunck [4], motion estimation and motion segmentation have been in the focus of computer vision research since the 1980's. Though problems like lack of texture and illumination changes still pose challenges, the robustness of motion segmentation could be increased by explicit models for motion discontinuities [13], robust error measures [1], and joint flow estimation and segmentation [3]. From the view of motion segmentation, our approach can be seen as an enhancement of a state-of-the-art method [3] with color clues.

Similar to our work are also recently introduced sprite-based methods [6, 7]. These decompose a scene into layers associated with template images ("sprites"). Each frame is explained by mapping sprites into the image domain while handling their occlusion. Sprites can be seen as precise pixel-wise color models. In contrast to them, the simpler models presented here do not provide pixel accuracy, but higher robustness to 3D changes. Also, since our approach estimates parametric motion by a simple least-squares estimation over flow error statistics, it runs in near real-time.

Figure 1. In contrast to motion only (left), the combination with a color model (right) gives a better segmentation.

3. Approach

Given image data I with pixel positions x, a binary mask m is estimated such that $m(x) = 1$ whenever x belongs to the foreground. Further, a parameter vector θ describes the background and foreground motion and appearance.

Using a MAP parameter estimation approach and assuming a uniform prior over θ, we obtain the optimization criterion:

$$\hat{m}, \hat{\theta} = \arg\max_{m,\theta} \quad p(I|m,\theta) \cdot P(m) \tag{1}$$

For the likelihood term $p(I|m,\theta)$, we assume that (1) pixels are independent, and (2) each pixel carries independent color and motion information: $I(x) = [I_v(x), I_c(x)]$.

$$p(I|m,\theta) \propto \left[\prod_x p(I_c(x)|m,\theta_c)\right]^{\alpha} \cdot \left[\prod_x p(I_v(x)|m,\theta_v)\right]^{1-\alpha}, \tag{2}$$

where the parameter $\alpha \in [0,1]$ determines how strongly unlikely color and motion values are penalized and thus balances the influence of color and motion information.

The prior $P(m)$ enforces a smooth and short object boundary. This is done using a Gibbs distribution (where $\mathcal{C}$ is the set of all neighbor pixel pairs and $\beta > 0$ is a parameter that weights the importance of the prior):

$$P(m) \propto \prod_{(x,y)\in\mathcal{C}, m(x)\neq m(y)} e^{-\beta} \tag{3}$$

By taking the logarithm, we obtain an energy function, minimizing which is equivalent to maximizing (1):

$$\begin{aligned} E_1(m,\theta;I) \quad = \quad & \underbrace{\alpha \cdot \sum_x -\log p(I_c(x)|m,\theta_c)}_{color\ cost} \\ & + \underbrace{(1-\alpha)\sum_x -\log p(I_v(x)|m,\theta_v)}_{motion\ cost} \\ & + \underbrace{\sum_{(x,y)\in\mathcal{C}, m(x)\neq m(y)} \beta}_{smoothness\ cost} \end{aligned} \tag{4}$$

E_1 consists of three terms: the first two regulate the fit of the foreground and background pixels to regions of coherent parametric color and motion, while the last one forces the segmentation boundary to be smooth by penalizing its length.

3.1 Color Information

We assume that background and foreground color are modeled by parametric densities with parameters $\theta_c = (\theta_c^b, \theta_c^f)$. The color likelihood is:

$$\begin{aligned} p(I_c(x)|m,\theta_c) = m(x) \cdot p_c(I_{RGB}(x)|\theta_c^f) \\ +[1-m(x)] \cdot p_c(I_{RGB}(x)|\theta_c^b). \end{aligned} \tag{5}$$

For the color distributions $p_c(I_{RGB}(x)|\theta_c^f)$ and $p_c(I_{RGB}(x)|\theta_c^b)$, we use Gaussian mixture models (where $\theta_c^{f/g}$ are component priors, means, and covariances) or color histograms with entries $\theta_c^{f/g}$.

3.2 Motion Information

For $p(I_v(x)|m,\theta_v)$, we use the *motion competition* model based on optical flow error [3]: two parametric (e.g., constant of affine) motions $\theta_v = (\theta_v^b, \theta_v^f)$ for background and foreground with variances σ_b^2, σ_f^2 are assumed. For each pixel position x, the model predicts the associated motion vectors for foreground, $v_f(x)$, and background, $v_b(x)$.

Like [3], we measure the quality of these motion predictions using the optical flow error $e_{f/g}(x) = \nabla I(x) \cdot v_{f/g}(x) + I_t(x)$, which is assumed to be normally distributed with $e_{f/g}(x) \sim \mathcal{N}(0, \sigma_{f/g}^2 \cdot ||\nabla I(x)||^2)$ (∇I is the image gradient and I_t is the temporal derivative).

The motion model can thus be rewritten as:

$$\begin{aligned} p(I_v(x)|m,\theta_v) = m(x) \cdot \mathcal{N}(e_f(x); 0, \sigma_f^2||\nabla I(x)||^2) \\ +[1-m(x)] \cdot \mathcal{N}(e_b(x); 0, \sigma_b^2||\nabla I(x)||^2) \end{aligned} \tag{6}$$

3.3 Extensions: Shape and Contrast

The model from Equation (4) can be improved using two observations. The first one is that motion boundaries tend to coincide with image edges. Therefore, a **contrast term** is used similar to [7]:

$$\begin{aligned} E_3(m_t,\theta^t;I,m_{t-1}) = E_2(m_t,\theta^t;I,m_{t-1}) \\ -\beta\cdot\eta\sum_{(x,y)\in\mathcal{C},m(x)\neq m(y)}\left[1-exp(-\tfrac{(I(x)-I(y))^2}{2\sigma^2})\right] \end{aligned}$$

i.e., the smoothness cost from Equation (4) is reduced by a factor η depending on the pixel difference (σ^2 is

Figure 2. In static scenes, background subtraction can provide accurate ground truth for testing motion segmentation.

estimated as two times the mean squared pixel difference). Consequently, region boundaries are favored to coincide with image edge locations.

The second observation is that objects should move smoothly between successive frames. This can be formulated using an additional **shape consistency** term over the segmentation masks of the current (m_t) and previous frame (m_{t-1}):

$$E_2(m_t, \theta^t; I, m_{t-1}) = E_1(m_t, \theta^t; I) + \sum_{x, m_t(x) \neq m_{t-1}(x)} \gamma \quad (7)$$

where the Lagrangian multiplier γ regulates the influence of shape consistency. Note that this term helps in cases where motion is not discriminative (e.g., if the object stands still) such that the proposed approach relies on color and shape consistency clues instead.

3.4 Optimization

To estimate the segmentation mask m_t and parameters θ for color and motion in foreground and background, the energy E_3 is minimized given the segmentation mask m_{t-1} and weights $\alpha, \beta, \gamma, \eta$. The optimization is carried out in an iterative scheme similar to [3], where θ and m are alternately estimated.

1. Initialize the mask (e.g., $m_t^1 := m_{t-1}$). Set $k = 1$
2. Estimate θ_c and θ_v using statistics over foreground (in case of θ_c^f, θ_v^f) and background (θ_c^b, θ_v^b) regions in m_t^k
3. Estimate m_t^{k+1} by fixing θ and minimizing E_3 using a graph cut algorithm.
4. If $m_t^{k+1} \neq m_t^k$: set $k = k + 1$ and goto (2). Else return m_t^{k+1}

Boykov and Kolmogorov [2] give an efficient algorithms for solving the graph cut problem. Also, the motion competition model can be estimated very efficiently from statistics of flow error as has been reported in [3]. At a resolution of 160×120, our non-tuned prototype runs at 3 fps. on a 2.4 GHz machine.

4 Experiments

A key challenge for training and evaluating segmentation systems is that the acquisition of ground truth is time-consuming, such that often only qualitative evaluations on few test sequences are done. To provide a quantitative evaluation, we use two test sets:

Static Scenes: In case of a static background and fixed camera focal length, an almost perfect segmentation can be achieved by robust background subtraction techniques (e.g., [12]). This provides a simple way to generate ground truth automatically. For this purpose, we used an implementation of a state-of-the-art online background subtraction approach. The method weights a shadow model [5] and local histograms of gradient directions [10] and integrates them with a background attenuation and graph cut as in Sun's Background Cut [12]. From a dataset of 24 short video clips in which several objects are presented to a camera in front of several static scenes, 507 frames are obtained (frames without foreground objects in them were ignored). For an impression of the ground truth segmentations obtained, see Figure 2. Note, however, that this approach is restricted to static scenes and is therefore significantly limited compared to our framework.

The proposed approach was tested on this dataset at a resolution of 160×120 pixels using an affine motion model and $\beta = 6, \gamma = 2.5, \eta = 0.5$ (this setup was obtained by a grid search optimization of segmentation error). To evaluate the influence of color information on the system, the color weight α was varied. If $\alpha = 0$, the system uses motion only and is similar to the approach from [3]. With increasing α the influence of color on the segmentation increases.

A sample result is illustrated in Figure 1 - in this scene, motion alone is not sufficient to segment the object from the background (possible explanations are sudden illumination changes and a motion pattern that does not suit the parametrization). If combined with color information, however, the object can be segmented from the background almost perfectly.

Quantitative results are given in Figure 4, where the segmentation error is plotted against the color weight α. Four color models were used: a Gaussian mixture model, a histogram model, and as baselines Normal densities with full and diagonal covariance matrices. Our insights from this experiment are: (1) By choosing a proper weight ($\alpha \approx 0.05$), segmentation quality can be improved by about 2 % . According to a t-test (level 0.5%) this improvement is significant. (2) To achieve this, a color model of a certain complexity is necessary - while both normal densities fail to improve segmentation quality, the more complex mixture model (12 com-

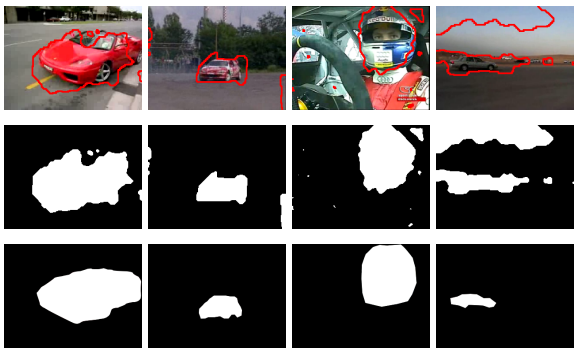

Figure 3. Segmentation results in dynamic scenes. Top row: input. Center: results. Bottom row: ground truth (manually acquired).

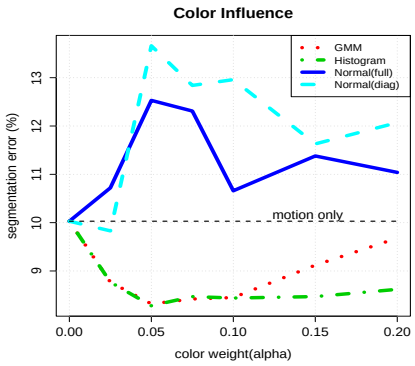

Figure 4. The segmentation error, plotted against the color weight α. For GMMs and histogram color models, segmentation error can be reduced by about 2 %.

ponents) and histogram (10^3 bins) lead to similar improvements.

General Scenes: For scenes in which the background is allowed to move, 29 pairs of frames were segmented manually. The data was sampled from videos downloaded from the video portal `revver.com` and shows cars, faces, and animals in motion.

For these scenes, larger images were used (240×180 pixels) and also smoothed with a Gaussian filter. We tested the system with a constant motion model and a histogram color model. Compared to the static case, our quantitative results show higher error rates (since the segmentation problem is more difficult for dynamic scenes), but also a reduction from 13.5 % ($\alpha = 0$) to 12.6 % ($\alpha = 0.3$) by using color information. A paired t-test was successful at a significance level of 25%; stronger statements might be possible with higher numbers of samples, but these are difficult to acquire. Some sample segmentations are illustrated in Figure 3.

5 Conclusions

In this paper, a segmentation approach was presented that combines color and optical flow in a joint probabilistic framework. Our quantitative experiments demonstrate that our way of integrating color clues with a state-of-the-art motion-only approach [3] improves segmentation results in both static and dynamic scenes.

Since our system is restricted to a single foreground region so far, a possible extension would be to integrate the approach with an α-expansion algorithm [7] to allow a segmentation of multiple foreground objects [1].

[1] This work was supported in part by the Stiftung Rheinland-Pfalz für Innovation, project InViRe (961-386261/791)

References

[1] M. Black and P. Anandan. The Robust Estimation of Multiple Motions: Parametric and Piecewise-smooth Flow Fields. *CVIU*, 1996.

[2] Y. Boykov and V. Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE PAMI*, 2004.

[3] D. Cremers and S. Soatto. Motion Competition: A Variational Approach to Piecewise Parametric Motion Segmentation. *IJCV*, 2005.

[4] B. Horn and B. Schunk. Determining Optical Flow. Technical report, M.I.T., 1980.

[5] T. Horprasert, D. Harwood, and L. Davis. A Statistical Approach for Realtime Robust Background Subtraction and Shadow Detection. In *Framerate Workshop*, 1999.

[6] A. Kannan, N. Jojic, and B. Frey. Generative Model for Layers of Appearance and Deformation. In *AISTATS Workshop 2005*, Barbados, 2005.

[7] M. Kumar, P. Torr, and A. Zisserman. Learning Layered Motion Segmentations of Video. *IJCV*, 2008.

[8] D. Martin, C. F. ad D. Tal, and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *ICCV'01*, Vancouver, Canada, 2001.

[9] F. Meyer and S. Beucher. Morphological Segmentation. *J. Vis. Com. Img. Repres.*, 1990.

[10] P. Noriega and O. Bernier. Real Time Illumination Invariant Background Subtraction Using Local Kernel Histograms. In *BMVC'06*, Edinburgh, UK, 2006.

[11] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE PAMI*, 2000.

[12] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum. Background Cut. In *ECCV'06*, Graz, Austria, 2006.

[13] Y. Weiss. Smoothness in Layers: Motion Segmentation using Nonparametric Mixture Estimation. In *CVPR'97*, San Juan, Puerto Rico, 1997.