Motion Coherent Tracking with Multi-label MRF optimization

David Tsai
caihsiaoster@gatech.edu
Matthew Flagg
mflagg@cc.gatech.edu
James M. Rehg
rehg@cc.gatech.edu

Computational Perception Laboratory
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA

We present a novel off-line algorithm for target segmentation and tracking in video. In our approach, video data is represented by a multi-label Markov Random Field model, and segmentation is accomplished by finding the minimum energy label assignment.

Our goal is to obtain higher-quality segmentations than existing on-line methods, without requiring significant user interaction. The primary novelty of our approach is our treatment of the inter-related tasks of segmenting the target and estimating its motion as a single global multi-label assignment problem. Energy functions enforce the *temporal coherence* of the solution, both spatially and across time. The result is a clean problem formulation based on global energy minimization. In contrast, on-line tracking methods can employ a diverse set of techniques to achieve good performance, including adaptive cue combination [1], spatially-varying appearance models [3], and shape priors [2]. We demonstrate experimentally that our approach can yield higher-quality segmentations than these previous methods, at the cost of greater computational requirements within a batch formulation.

Given a video sequence and manual initialization of a target of interest in the first frame, our goal is to carve the moving target out of the video volume, yielding a target segmentation at every frame. We adopt the volumetric MRF formulation, in which the video volume is represented as a multi-label MRF with hidden nodes corresponding to the unknown labels. The resulting optimization problem is to find the label assignment l to p that minimizes

$$E(l_p) = \sum_{p \in G} V_p(l_p) + \sum_{p \in G} \sum_{q \in N(p)} V_{pq}(l_p, l_q) \quad (1)$$

where $V_p(\cdot)$ are the unary potentials representing the data term, $V_{pq}(\cdot,\cdot)$ are the pairwise potentials representing the smoothness term,G represents the entire set of nodes, and N represents the neighborhood system of the nodes, both spatially and temporally.

In contrast to the standard approach to MRF-based segmentation, our label set *augments* the usual foreground/background binary attribute with a discrete representation of the flow between frames. Associated with each label is a quantized motion field $\{d^1, \cdots, d^i\}$, such that the label assignment l_p to pixel site p is associated with displacing that node by the corresponding vector d^{l_p}. Figure. 1 illustrates these combinations for a simple 1D example. Note that we take the cartesian product of attributes and flows (rather than their sum) because the interaction between these two components is a key element in enforcing temporal coherence between frames. In order to optimize the energy function in Equation 1, we adopt the Fast-PD method of Komodakis et. al. [4, 5].

The second goal of this work is to facilitate a deeper understanding of the trade-offs and issues involved in on-line and off-line formulations of video segmentation and tracking, via a standardized database of videos with ground-truth segmentations. There has been very little comparative work addressing the segmentation performance of tracking methods. Our starting point was to identify three properties of video sequences that pose challenges for segmentation quality: color overlap between target and background appearance, interframe motion, and change in target shape. We developed a quantitative measure for each these phenomena, and we systematically assembled an evaluation dataset, called *SegTrack*, which spans the space of challenges. We also provide direct comparison between our batch tracking method and two state-of-the-art on-line contour-based algorithms [1, 3].

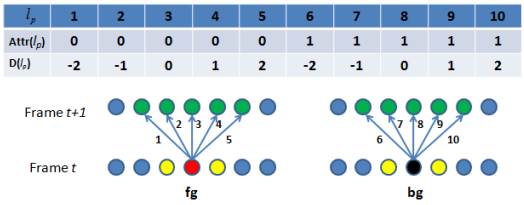

Figure 1: **Illustration of label definition**. We illustrate the label space for a center pixel in frame t. If the maximum displacement in the x direction is 2, then there are 5 possible displacements ranging from (-2,0) to (2,0). In each case, the pixel can also be labeled either foreground (red, lower left figure) or background (black, lower right figure), resulting in 10 possible labels per pixel.

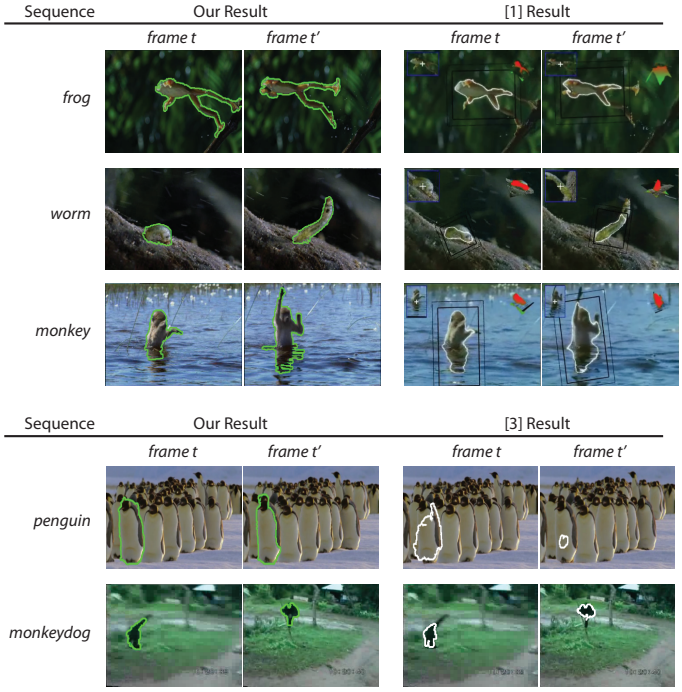

Figure 2: **Comparative results**: This figure compares selected output frames from our method with two competitors.

[1] Charles Bibby and Ian Reid. Robust real-time visual tracking using pixel-wise posteriors. In *ECCV*, 2008.

[2] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *IJCV*, 22(1):61–79, 1997.

[3] P. Chockalingam, N. Pradeep, and S. T. Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. In *ICCV*, 2009.

[4] N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *ICCV*, 2007.

[5] Nikos Komodakis and Georgios Tziritas. A new framework for approximate labeling via graph cuts. In *ICCV*, 2005.