Robust Model-Free Tracking of Non-Rigid Shape

Lorenzo Torresani
Stanford University
ltorresa@cs.stanford.edu

Aaron Hertzmann
University of Toronto
hertzman@dgp.toronto.edu

Christoph Bregler
New York University
chris.bregler@nyu.edu

New York University CS TR2003-840

June 9, 2003

Abstract

We present a robust algorithm for estimating non-rigid motion in video sequences. We build on recent methods for tracking video by enforcing global structure (such as rank constraints) on the tracking. These methods assume color constancy in the neighborhood of each tracked feature, an assumption that is violated by occlusions, deformations, lighting changes, and other effects. Our method identifies outliers while solving for flow. This allows us to obtain high-quality tracking from difficult sequences, even when there is no single "reference frame" in which all tracks are visible.

1. Introduction

Shape and motion reconstruction from uncalibrated video sequences promises to produce high-quality 3D models for video analysis and computer graphics applications. Recent work has shown that, by viewing tracking and reconstruction as a global optimization problem, high-quality 3D shape and motion models may be obtained from video, even for non-rigid shapes [5, 11, 2]. However, these methods make the restrictive assumption of *color constancy*, that object features appear the same in all views. Almost all sequences of interest violate this assumption at times, such as with occlusions, lighting changes, motion blur, and many other common effects. Although it may be possible to explicitly model all sources of variability, in practice, this will yield an extremely difficult modeling and optimization problem.

In this paper, we propose a global optimization framework for tracking based on robust statistics: all violations of color constancy are modeled as outliers. This allows us to tackle more challenging tracking problems, without requiring us to explicitly model all errors. We demonstrate sequences for which previous global reconstruction methods fail. For example, previous methods require that all feature points be visible in all video frames, i.e. all features are visible in a single "reference frame;" our method relaxes this assumption and allows sequences for which no single "reference frame" exists. We also show examples where existing

techniques fail due to local changes in lighting and shape. Our method is based on the EM algorithm for robust outlier detection.

1.1. Relation to Previous Work

We build on recent techniques for exploiting rank constraints on optical flow. Conventional flow and tracking algorithms use only local information; namely, every track in every frame is estimated separately [4, 7]. Irani [5] treated optical flow as a global problem, combining information from the entire sequence — along with rank constraints on motion — to yield better tracking. Extending these ideas, Bregler et al. [3], Torresani et al. [11, 10] and Brand [2] describe tracking and reconstruction algorithms that solve for 3D shape and motion from sequences, even for non-rigid scenes. In this paper, we show how to embed all of these previous approaches in a robust framework, in order to handle cases for which they would otherwise fail, such as occlusions and lighting variation.

Robust algorithms for tracking have been widely explored in local tracking algorithms [1]. We formulate a probabilistic outlier model and solve globally over the entire sequence, using the EM algorithm. Unlike local robust methods, our method can handle "track" features that are completely occluded, by making use of global constraints on motion. Our outlier model is closely related to layer-based motion segmentation algorithms [12, 13, 6], which also often apply EM globally to a sequence. We use the outlier model to handle general violations of color constancy, rather than to specifically model multiple layers.

2. Robust flow

We now describe our basic robust optical flow algorithm. We first define a generative model for video sequences, and then a generalized-EM algorithm for estimating the parameters of this model from a sequence.

Image formation model. We assume that 3D motion can be described in terms of J points that move in 3D. At a given time t, point j projects to a 2D position $\mathbf{p}_{j,t} = (x_{j,t}, y_{j,t})$; these 2D projections are called *point tracks*. These point tracks are restricted by a 3D motion model that depends on the specific application. For example, the point tracks may be produced by orthographic projection of a rigid object, in which case, the matrix of motion vectors $\mathbf{p}_{j,t} - \mathbf{p}_{j,0}$ is restricted to be rank 3 [5]. We denote the complete tracking data over all tracks and all frames by a variable $\mathbf{P} = \{..., \mathbf{p}_{j,t}, ...\}$.

Individual images in a video sequence are created from the point tracks. Ideally, the window of pixels around each point track should remain constant over time; however, this window may be corrupted by noise and outliers. Let w be an index over a pixel window, so that $I(\mathbf{p}_{w,j,t})$ is a pixel in the window of track j in frame t. This pixel intensity should be a constant $\mu_{w,j}$; however, it will be corrupted by Gaussian noise with variance σ^2. Moreover, it may be replaced by an outlier, with probability $1 - \tau$. We define a hidden variable $W_{w,j,t}$ so that $W_{w,j,t} = 0$ if the pixel is replaced by an outlier, and $W_{w,j,t} = 1$ if it is valid. The complete PDF over individual pixels in a window is given by:

$$p(I(\mathbf{p}_{w,j,t})|W_{w,j,t} = 1, \mathbf{P}) = \mathcal{N}(I(\mathbf{p}_{w,j,t})|\mu_{w,j}; \sigma^2) \quad (1)$$

$$p(I(\mathbf{p}_{w,j,t})|W_{w,j,t} = 0, \mathbf{P}) = c \quad (2)$$

$$p(W_{w,j,t} = 1) = \tau \quad (3)$$

where $\mathcal{N}(I(\mathbf{p}_{w,j,t})|\mu_{w,j};\sigma^2)$ denotes a 1D Gaussian distribution with mean $\mu_{w,j}$ and variance σ^2, and c is a constant corresponding to the uniform distribution over pixel intensities.

For convenience, we do not model the appearance of video pixels that do not appear near point tracks, or correlations between pixels when windows overlap.

Problem statement. Given a video sequence I and track positions specified in some reference frames, we would like to estimate the positions of the tracks in all other frames, as well as which pixels are valid or outliers. The values $\mu_{w,j}$ are determined from the corresponding reference frame, in which the window is assumed to be completely uncorrupted. It is *not* required that the same reference frame be used for all tracks.

We pose this as a problem of estimating the point tracks $\mathbf{P}$ by maximizing

$$p(\mathbf{P}|I,\mu,\sigma,\tau) \tag{4}$$

This density can be expanded in terms of the hidden variables $W_{w,j,t}$:

$$\begin{aligned} p(\mathbf{P}|I,\theta) &= p(\mathbf{P},W_{w,j,t}=1|I,\theta)+p(\mathbf{P},W_{w,j,t}=0|I,\theta) &(5)\\ &= p(\mathbf{P}|W_{w,j,t}=1,I,\theta)p(W_{w,j,t}=1|I,\theta)+p(\mathbf{P}|W_{w,j,t}=0,I,\theta)p(W_{w,j,t}=0|I,\theta) &(6)\end{aligned}$$

where θ encapsulates the terms μ, τ, and σ. Denoting $\gamma_{w,j,t}=p(W_{w,j,t}=1)$, we can view this as a problem of jointly optimizing $\mathbf{P}$ and $\gamma_{w,j,t}$. The value $\gamma_{w,j,t}$ indicates the likelihood that pixel (w,j,t) is an outlier. In addition, we would like to jointly learn the maximum likelihood values of τ and σ^2.

Generalized EM algorithm. The above problem can be optimized using a generalized EM algorithm. In the E-step, we estimate the distribution $\gamma_{w,j,t}$, given our current estimate of the motion $\mathbf{P}$:

$$\gamma_{w,j,t} \equiv p(W_{w,j,t}=1|\mathbf{P},I,\theta) \tag{7}$$

Intuitively, this is estimated by warping the reference frame to the position $(\mathbf{p}_{w,j,t})$, and measuring reconstruction error. More formally, let

$$\begin{aligned} \alpha_0 &= p(I(\mathbf{p}_{w,j,t}),W_{w,j,t}=0|\mathbf{P},\theta) &(8)\\ &= p(I(\mathbf{p}_{w,j,t})|W_{w,j,t}=0,\mathbf{P},\theta)p(W_{w,j,t}=0|\mathbf{P},\theta) &(9)\\ &= (1-\tau)c &(10)\\ \alpha_1 &= p(I(\mathbf{p}_{w,j,t}),W_{w,j,t}=1|\mathbf{P},\theta) &(11)\\ &= p(I(\mathbf{p}_{w,j,t})|W_{w,j,t}=1,\mathbf{P},\theta)p(W_{w,j,t}=1|\mathbf{P},\theta) &(12)\\ &= \frac{\tau}{\sqrt{2\pi\sigma^2}}e^{-||I(\mathbf{p}_{w,j,t})-\mu_{w,j}||^2/(2\sigma^2)} &(13)\end{aligned}$$

Then, using Bayes' Rule and $p(I|\mathbf{P},\theta)=\alpha_0+\alpha_1$, we have

$$\begin{aligned} p(W_{w,j,t}=1|\mathbf{P},I,\theta) &= p(I|W_{w,j,t}=1,\mathbf{P})p(W_{w,j,t}=1|\mathbf{P},\theta)/p(I|\mathbf{P},\theta) &(14)\\ &= \alpha_1/(\alpha_0+\alpha_1) &(15)\end{aligned}$$

In the generalized M-step, we solve for optical flow given the outlier probabilities $\gamma_{w,j,t}$. The outlier probabilities provide a weighting function for tracking: pixels likely to be valid are given more weight in the tracking. Our goal is to minimize the following energy function:

$$\begin{aligned} Q(\mathbf{P},\theta) &= E_\gamma[-\log p(\mathbf{P}|\theta)] \quad (16)\\ &= -\sum_{w,j,t}\gamma_{w,j,t}\log p(I(\mathbf{p}_{w,j,t}),W=1|\theta) \\ &\quad -\sum_{w,j,t}(1-\gamma_{w,j,t})\log p(I(\mathbf{p}_{w,j,t}),W=0|\theta) \\ &= \sum_{w,j,t}\gamma_{w,j,t}(I(\mathbf{p}_{w,j,t})-\mu_{w,j,t})^2/(2\sigma^2) \\ &\quad +\sum_{w,j,t}\gamma_{w,j,t}\log\sqrt{2\pi\sigma^2}-\sum_{w,j,t}\gamma_{w,j,t}\log\tau \\ &\quad -\sum_{w,j,t}(1-\gamma_{w,j,t})\log c(1-\tau) \quad (17)\end{aligned}$$

To solve for the motion, we linearize the target image around the current estimate $\mathbf{p}^0_{w,j,t}$:

$$I(\mathbf{p}_{w,j,t}) \approx I(\mathbf{p}^0_{w,j,t})+\nabla I^T(\mathbf{p}_{w,j,t}-\mathbf{p}^0_{w,j,t}) \quad (18)$$

where ∇I denotes a 2D vector of partial derivatives of I evaluated at $\mathbf{p}^0_{w,j,t}$. One such linearization applied for every pixel w in every window j for every frame t. Substituting Equation 18 into 17 yields the following quadratic energy function for the motion $\mathbf{P}$:

$$\sum_{t,j,w}\gamma_{w,j,t}(I(\mathbf{p}^0_{w,j,t})+\nabla I^T(\mathbf{p}_{w,j,t}-\mathbf{p}^0_{w,j,t})-\mu_{w,j,t})^2/(2\sigma^2) \quad (19)$$

plus terms that are constant with respect to $\mathbf{P}$. The optimization of this function for $\mathbf{P}$ depends on the specific motion model. For example, covariance-weighted factorization may be used for orthographic projection of rigid motion [8]; bundle adjustment may be used for perspective projection of rigid scenes.

The noise variance and outlier probabilities are also updated in the M-step, by optimizing $Q(\mathbf{P},\theta)$ for τ and σ:

$$\tau \leftarrow \sum_{w,j,t}\gamma_{w,j,t}/(JNT) \quad (20)$$

$$\sigma^2 \leftarrow \sum_{w,j,t}\gamma_{w,j,t}(I(\mathbf{p}_{w,j,t})-\mu_{w,j,t})^2/\sum_{w,j,t}\gamma_{w,j,t} \quad (21)$$

where T is the number of images and N is the number of pixels in a window.

Implementation details. We initialize our algorithm using conventional coarse-to-fine Lucas-Kanade tracking [7]. Since the conventional tracker will diverge if applied to the entire sequence at once, we correct the motion every few frames by applying our global robust EM algorithm over the subsequence thus far initialized. This process is repeated until we reach the end of the sequence. We refine this estimate by additional EM iterations. The values of σ and τ are initially held fixed at 10 and 0.3, respectively. They are then updated in every M-step after the first few iterations.

Experiments. We applied the robust tracking algorithm technique to two video sequences of human motion. The first is a video consisting of 660 frames recorded in our lab with a consumer digital camera. The video contains non-rigid deformations of a human torso. Although most of the features tracked are characterized by distinctive 2D texture, their local appearance changes considerably during the sequence, due to occlusions, shape deformations, varying illumination in patches, and motion blur. More than 25% of the frames contain occluded features, due to arm motion and large torso rotations. 77 features were selected automatically in the first frame using the criterion described by Shi and Tomasi [9]. Figure 1(a) shows their initial locations in the reference frame. The sequence was initially processed assuming rank 5 motion,

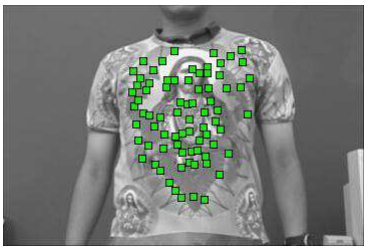
(a)  (b) 

Figure 1: Reference frames. Regions of interest were selected manually, and individual tracks selected automatically using Shi and Tomasi's method [9]. Note that, in the second sequence, most tracks are clearly visible in only one reference frame. (Refer to the electronic version of this paper to view the tracks in color.)

and progressively increased during optimization to rank 10 (corresponding to 3 basis shapes under weak perspective projection). Estimated positions of features with and without robustness are shown in Figure 2 and in the accompanying videos. It is worth noting that tracking this sequence without robustness fails even *before* occlusions appear, due to deformations and variations in illumination. Figure 3 shows the 3D reconstruction obtained by the factorization. Our algorithm yields an accurate reconstruction despite the difficulty of the sequence, reconstructing even occluded regions.

The second video contains 100 frames long of mostly-rigid head/face motion. The sequence is challenging due to low resolution and low frame rate (15 fps). In this example, there is no single frame in which feature points from both sides of the face are clearly visible, so existing global techniques cannot be applied. To test our algorithm, we used 45 features automatically selected from two separate reference frames (Figure 1(b)). Tracking was assumed to be rank 4, corresponding to rigid motion under weak perspective projection. Points from the left side of the person's face are occluded for more than 50% of the sequence. Robust tracking is necessary for even short subsequences; without it, some of the features on the left side of the face are lost or incorrectly tracked after just four frames. Within 14 frames, all points from the left side are completely lost. With robust tracking, our algorithm successfully tracks, making use of geometric constraints to fill in missing tracks (Figure 4).

3. Discussion and future work

We have presented techniques for tracking and reconstruction from video sequences that contain occlusions and other common violations of color constancy. Most tracking of challenging footage and severe occlusion as we present here can only be achieved with very strong appearance and very restricted dynamical models. We have shown how to track such difficult sequences without any prior knowledge of appearance and

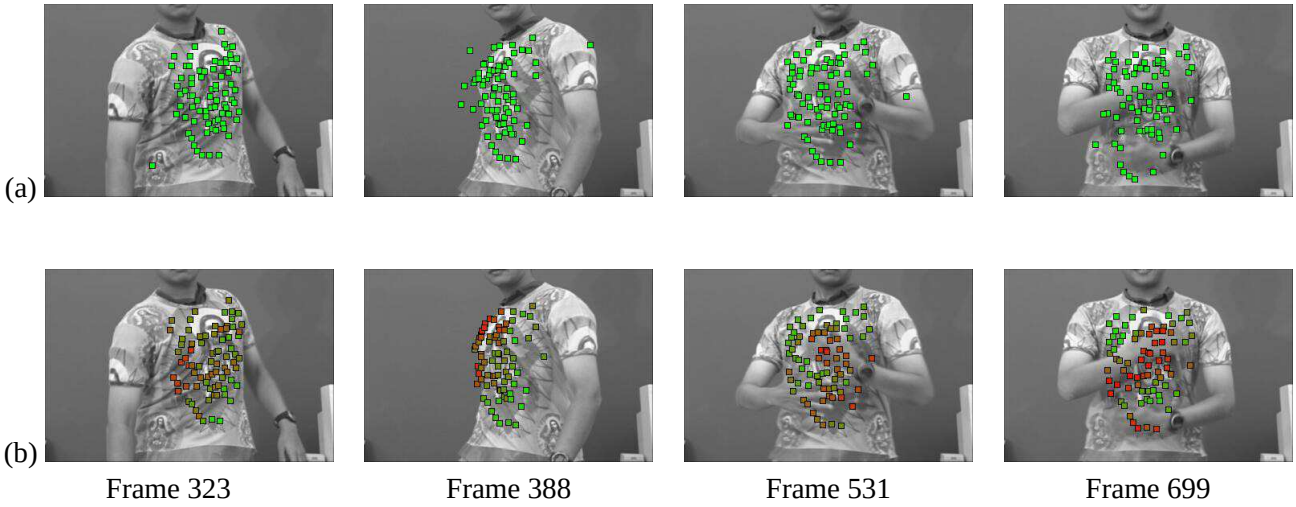


Figure 2: (a) Rank-constrained tracking of the first sequence without outlier detection, using the reference frames shown in Figure 1. Tracks on occluded portions of the face are consistently lost. (b) Robust, rank-constrained tracking applied to the same sequence. Tracks are colored according to the average value of $\gamma_{w,j,t}$ for the pixels in the track's window: green for completely valid pixels, and red for all outliers.

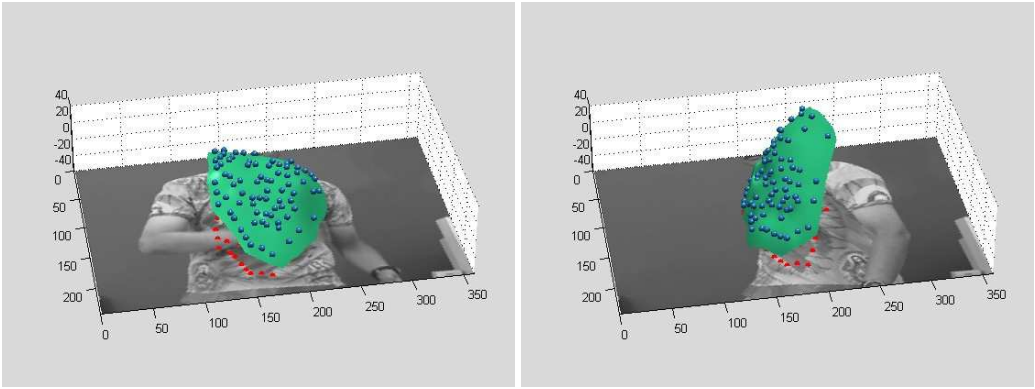


Figure 3: Reconstructions of frames from the first sequence, using robust tracking. Note that occluded areas are accurately reconstructed.

Figure 4: (a) Rank-constrained tracking of the second sequence without outlier detection, using the reference frames shown in Figure 1. Tracks on occluded portions of the face are consistently lost. (b) Robust, rank-constrained tracking applied to the same sequence. Tracks are colored according to the average value of $\gamma_{w,j,t}$ for the pixels in the track's window: green for completely valid pixels, and red for all outliers.

dynamics.

We expect that these techniques can provide a bridge to very practical tracking and reconstruction algorithms, by allowing one to model important variations in detail (such as a more sophisticated lighting or visibility model) without having to model all other sources of non-constancy.

It would be straightforward to handle true perspective projection for rigid scenes in our algorithm, by replacing our closed-form M-steps with bundle adjustment.

Acknowledgments

We are grateful to Aseem Agarwala for providing the second test sequence. Portions of this work were performed while LT was visiting New York University, AH was at University of Washington, and while CB was at Stanford University. LT and CB were supported by ONR grant N00014-01-1-0890 under the MURI program. AH was supported in part by UW Animation Research Labs and NSF grant IIS-0113007.

References

[1] Michael J. Black and P. Anandan. A Framework for the Robust Estimation of Optical Flow. In *Proc. ICCV 1993*, pages 231–236, May 1993.

[2] Matthew Brand. Morphable 3D models from video. In *Proc. CVPR 2001*, 2001.

[3] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering Non-Rigid 3D Shape from Image Streams. In *Proc. CVPR 2000*, 2000.

[4] B. K. P. Horn. *Robot Vision*. McGraw-Hill, New York, NY, 1986.

[5] Michal Irani. Multi-Frame Optical Flow Estimation Using Subspace Constraints. In *Proc. ICCV 99*, September 1999.

[6] Nebojsa Jojic and Brendan Frey. Learning Flexible Sprites in Video Layers. In *Proc. CVPR 2001*, 2001.

[7] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. 7th Int. Joint Conf. on Artifi cial Intelligence*, 1981.

[8] Daniel D. Morris and Takeo Kanade. A Unifi ed Factorization Algorithm for Points, Line Segments and Planes with Uncertainty Models. In *Proc. ICCV 98*, pages 696–702, January 1998.

[9] Jianbo Shi and Carlo Tomasi. Good Features to Track. In *Proc. CVPR '94*, pages 593–600, 1994.

[10] Lorenzo Torresani and Christoph Bregler. Space-Time Tracking. In *Proc. ECCV 2002*, 2002.

[11] Lorenzo Torresani, Danny Yang, Gene Alexander, and Christoph Bregler. Tracking and Modeling Non-Rigid Objects with Rank Constraints. In *Proc. CVPR 2001*, 2001.

[12] John Y. A. Wang and Edward H. Adelson. Representing moving images with layers. *IEEE Trans. Image Processing*, 3(5):625–638, 1994.

[13] Yair Weiss and Edward H. Adelson. Perceptually organized EM: A framework for motion segmentation that combines information about form and motion. Technical Report TR 315, MIT Media Lab Perceptual Computing Section, 1994.