Efficient Extraction of Human Motion Volumes by Tracking

Juan Carlos Niebles
Princeton University, USA
Universidad del Norte, Colombia
jniebles@princeton.edu

Bohyung Han
Electrical and Computer Engineering
UNIST, Korea
bhhan@unist.ac.kr

Li Fei-Fei
Computer Science Department
Stanford University, USA
feifeili@cs.stanford.edu

Abstract

We present an automatic and efficient method to extract spatio-temporal human volumes from video, which combines top-down model-based and bottom-up appearance-based approaches. From the top-down perspective, our algorithm applies shape priors probabilistically to candidate image regions obtained by pedestrian detection, and provides accurate estimates of the human body areas which serve as important constraints for bottom-up processing. Temporal propagation of the identified region is performed with bottom-up cues in an efficient level-set framework, which takes advantage of the sparse top-down information that is available. Our formulation also optimizes the extracted human volume across frames through belief propagation and provides temporally coherent human regions. We demonstrate the ability of our method to extract human body regions efficiently and automatically from a large, challenging dataset collected from YouTube.

1. Introduction

We propose an automatic and efficient algorithm to extract humans with arbitrary motions and poses from videos of unknown settings. This problem is critical in many real-world applications that require accurate and efficient human motion estimation. For example, a mobile agent that navigates the world by interacting with humans in real-time needs to identify and track people in its surroundings. Also, tasks such as video indexing, search, and intelligent surveillance would benefit greatly by accurate human behavior understanding. Traditionally, research in this area has been done mostly from a tracking perspective [8]. However, tracking humans in natural videos is notoriously challenging because of background clutter, variety of body poses and motion, unknown number of subjects, occlusions, illumination changes, unconstrained camera motion, etc.

There are two classes of approaches in human motion estimation. Model-based methods, or top-down models, encode the poses and movements of the human body with an *a priori* structure model [5, 9, 10, 12, 16, 17, 20, 19, 22, 21]. Learning and inference procedures attempt to fit the image evidence to the best configuration of the model, but typically involve many degrees of freedom, large search space, and complex observations; it is painfully slow in general, due to the large amount of computation.

On the other hand, contour-based representations of deformable shape are often used to describe human body structure and motion efficiently within a level-set framework [2, 3, 25]. In [18], the boundary of a human is identified through a region-based foreground/background segmentation based on multiple low-level cues. However, these methods ignore the structure of the human body and/or impose very strong priors, which may lead to critical limitations when estimating articulated and flexible human poses.

In this paper, we present a method that achieves a balance between efficiency and accuracy for extracting human motion volumes from uncontrolled videos. We observe that a combination of top-down and bottom-up modeling can extract accurate motion volumes with only a relatively small computational load. Our idea is simple: given a video sequence, we apply top-down human models in a very sparse set of key frames. The bottom-up algorithm then bootstraps this detailed human information to complete the rest of the extraction through a temporal propagation and a global optimization procedure. Our experiments show that the proposed method achieves a near real-time human tracking in natural videos. In this study, our contributions can be summarized as below:

- A system is designed to automatically extract human motion volume from challenging videos by combining the top-down and the bottom-up method.
- We propose a novel top-down modeling technique to obtain a probabilistic human body contour.
- A global optimization procedure based on belief propagation is proposed to improve the quality of results.

The rest of the paper is organized as follows. In Section 2, we give an overview of our algorithm and discuss

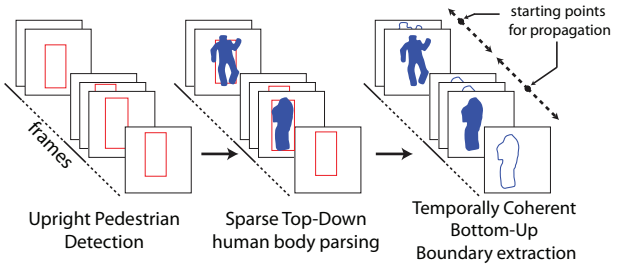


Figure 1. Algorithm overview. For each identified human obtained from pedestrian detection, a probabilistic human body shape on a sparse set of frames is computed by integrating top-down segmentations of detected pedestrian windows, which are driven by upright human pose templates (Section 3.1). A bottom-up boundary extraction based on the level-set formulation is employed to automatically refine and propagate the extracted contours to all frames (Section 3.2). The final human contours are obtained at all frames simultaneously, by jointly optimizing the level-set functions at all frames (Section 3.3).

several key properties. Each step of our algorithm is elaborated in Section 3. Experimental results and discussions are presented in Sections 4 and 5, respectively.

2. Algorithm Overview

Given an input video sequence, the goal of our algorithm is to carve out a spatio-temporal volume for each person in the video. The key strategy of our approach is the sparse introduction of top-down constraints, which are propagated in time in a bottom-up fashion. In this section, we describe the overall architecture of our method, which is also depicted in Fig. 1.

Person Detection and Clustering We first use an upright human detector [11] to generate potential human regions. The appearance similarity and the spatio-temporal coherence of the detections are employed to cluster detections in a similar fashion to [14]. Each resulting cluster is then associated to a unique individual, for which the spatio-temporal volume will be carved out. In practice, each cluster contains the bounding boxes for a person but there are many missing frames due to detection errors and pose variations.

Top-down Pose Estimation For each identified person, our algorithm performs a top-down extraction of the human region for a small subset of the frames. At each of these frames, the level-set function for the detected pedestrian is initialized based on the probabilistic integration of the upright human pose templates [24]. Such top-down driven extraction is utilized as an important constraint for the later bottom-up process. That is, the top-down information is delivered by the use of a set of fixed templates instead of more expensive part-based articulated models such as pictorial structures [6, 7, 16, 14]. However, it is only applicable to the frames with pedestrian detection, and the accurate estimation of an initial level-set function is still challenging due to the limited variety in the upright human body template database and the lack of discriminative features.

Bottom-up Contour Extraction and Propagation In the previous step, we obtained the level-set functions for a small subset of the detected pedestrians. The level-set functions for the rest of the frames are initialized by propagating existing ones to adjacent frames bidirectionally using low-level feature observations, with a procedure based on an extension of [2]. The bottom-up level-set approach can handle the arbitrary shape of an object efficiently, but is inherently susceptible to fall in local optima. The combination of the top-down and the bottom-up approaches reduces the drawbacks of both methods significantly. In addition, we jointly optimize the level-set functions at all frames simultaneously, which provides accurate and temporally coherent boundaries of the human body.

3. Efficient Extraction of Human Volumes

In this section, we describe three main components of our algorithm in detail, which include a top-down model-based probabilistic level-set initialization, a bottom-up feature-based propagation of level-set functions, and a global optimization process for contour extraction.

3.1. Top-Down Estimation of Human Body Region

For each identified person in the detection and clustering stage, a number of frames with pedestrian detections are available. Top-down shape priors are applied to the detected pedestrians, where the human poses in the candidate windows are constrained to be upright and the shape priors are in the form of pedestrian silhouette templates. Suppose that $B = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}$ is a set of the upright human body templates in the database obtained from [24]. At the j-th pedestrian detection window selected for top-down processing, we generate a set of multiple segmentations, each of which is driven by a different template in the database. In practice, we obtain each segmentation $\mathbf{h}_i$ within a level-set framework [4, 15], where the initial level-set function is given by the template $\mathbf{t}_i$. Each segmentation corresponds to an estimate of the human region. We integrate resulting segmentations probabilistically to obtain the probabilistic template $\mathbf{p}_j$ as

$$\mathbf{p}_j = \sum_{i=1}^{n} \omega_i \mathbf{h}_i, \tag{1}$$

where ω_i is estimated by matching the template and the resulting segmentation using multiple features—shape, color and edge—as follows:

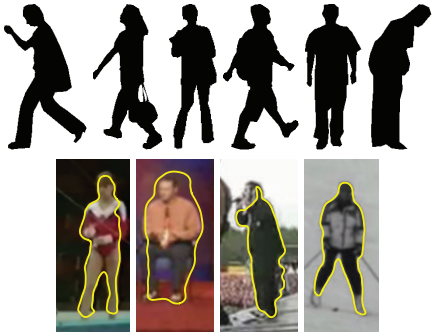


Figure 2. Top-down estimation of human region with upright human templates. (Top) Example templates in the pedestrian database from [24]. (Bottom) Template-driven segmentation results.

Shape We measure the shape distance r_s between the original contour of a template in the database and its induced level-set segmentation by

$$r_s = D_{\chi^2}(S(\mathbf{t}_i), S(\mathbf{h}_i)), \tag{2}$$

where $S(\cdot)$ are shape descriptors and D_{χ^2} is the χ^2 distance operator. In practice, we describe shapes using a histogram computed from the shape of the estimated region, which is a simplified version of the descriptor in [23].

Color We favor contours that induce the most distinct foreground and background color distributions. We estimate both distributions using the pixel assignments based on the estimated region $\mathbf{h}_i$. The distance between the foreground and the background color model—M_f^{color} and M_b^{color}, respectively—is defined by

$$r_c = D_{\chi^2}(M_f^{color}, M_b^{color}). \tag{3}$$

In practice, each color model is a multinomial distribution over the quantized color space.

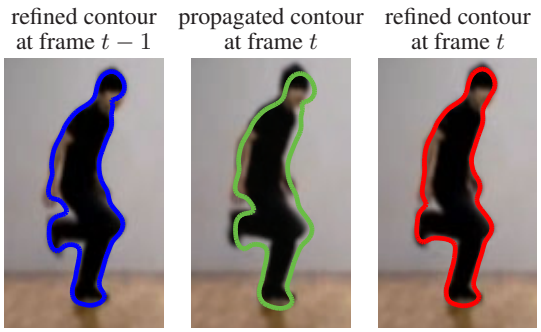
Edges The dissimilarity between the edge map in the pedestrian window and the estimated region is measured by

$$r_e = \frac{1}{N} \sum_{c_i \in C} \min_{e_j \in \mathrm{E}} D(c_i, e_j), \tag{4}$$

where C is the set of N points in the contour of $\mathbf{h}_i$ and E is the set of edge pixels in the edge map. D measures the Euclidean distance between two pixels. This is equivalent to the average distance from the points in the contour to the edge points in the image. The computation is done via Distance Transform.

We combine the multiple cues to obtain the weight for each template by

$$\omega_i = \frac{\exp\left(\kappa_0 + \sum_{j\in\{s,c,e\}} \kappa_j r_j\right)}{1 + \exp\left(\kappa_0 + \sum_{j\in\{s,c,e\}} \kappa_j r_j\right)}, \tag{5}$$



(a) (b) (c)

Figure 3. Bottom-up boundary extraction. An efficient level-set method is proposed to extract the human region boundary. Initial boundaries from the top-down procedure (a) are propagated across time (b) and refined by evolving the implicit level-set function (c). The final boundary is generally more accurate after a few iterations. This figure is best viewed in color.

where parameters κ_j $(j = 0, s, c, e)$ are learned from examples using logistic regression and $\sum_i \omega_i = 1$.

In our algorithm, the top-down constraints are applied to a small number of frames since top-down processing is more computationally expensive than bottom-up processing. Because, we would not gain much benefit from top-down processing many consecutive frames, it is applied to some frames that are temporally far apart. We first select a frame randomly and add more frames that are most distant from the current set of selected frames. A trade-off is observed here; the more frames selected for top-down processing, the more accurate the constraints for bottom-up processing will be, but at the same time the computational cost is increased.

The accurately estimated contours of the human body in this step are used as constraints for the bottom-up propagation. Some examples of the templates and the extracted contours by the top-down process are presented in Figure 2.

3.2. Bottom-Up Propagation of the Human Volume

After obtaining a sparse set of contours with the top-down process, we propagate the contours efficiently to other frames. The problem is formulated within a level-set contour tracking framework based on bottom-up cues. Since level-set based segmentations frequently converge to local optima, we alleviate the problem by accurate initialization of the level-set functions with the top-down constraints.

In the level-set framework [4, 15], a region of interest R in image I is implicitly represented by a non-parametric level-set function $\mathbf{\Phi}$:

$$R = \{\mathbf{x} \in I | \mathbf{\Phi}(\mathbf{x}) > 0\}, \tag{6}$$

where $\mathbf{x}$ is a pixel in the image, and the boundary is defined by the set of points such that $\mathbf{\Phi}(\mathbf{x}) = 0$. A foreground

segment R is obtained by an iterative procedure based on low-level features from an initial level-set function.

In our formulation, we propagate the level-set functions induced by the top-down templates in both directions, forward and backward. Let $\mathbf{\Phi}^{t-1}$ be the initialized level-set function at frame $t-1$, and T the length of the sequence of interest. We propagate $\mathbf{\Phi}^{t-1}$ to the temporally adjacent frames, by employing an image registration technique [1] that finds a rigid warping of $\mathbf{\Phi}^{t-1}$ to the new frame, e.g., frame t. We then apply a fixed small number of level-set iterations (typically, 5) to partially optimize the level-set function based on the observation in the image. Such bidirectional propagation terminates when the level-set functions in all frames are initialized.

Let us present how the level-set function $\mathbf{\Phi}^t$ is optimized in each frame iteratively based on low-level image features. [1] Fig. 3 shows an illustration of the within-frame level-set optimization. The goal is to evolve the initial level-set function by maximizing the conditional probability given by

$$p(\mathbf{\Phi}^t|\mathbf{x}^t,\mathbf{y}^t) = \prod_{i=1}^{N^t} p(\mathbf{\Phi}_i^t|\mathbf{x}_i^t,\mathbf{y}_i^t), \tag{7}$$

where N^t is the number of pixels, $\mathbf{y}^t$ is the observed image feature, and the pixel-wise level-set likelihood is given by

$$\begin{aligned} p(\mathbf{\Phi}_i^t|\mathbf{x}_i^t,\mathbf{y}_i^t) &\propto p(\mathbf{x}_i^t|\mathbf{\Phi}_i^t,\mathbf{y}_i^t)p(\mathbf{\Phi}_i^t) \\ &= p(\mathbf{\Phi}_i^t)\sum_M p(\mathbf{x}_i^t|\mathbf{\Phi}_i^t,M)p(M|\mathbf{y}_i^t), \end{aligned} \tag{8}$$

where M is the model parameter for foreground (M_f) or background (M_b).

In [2], only color distribution is utilized to model the foreground and background regions, but motion information based on optical flow is additionally employed in our method to provide foreground and background probabilities to each pixel, $p(M|\mathbf{y}_i^t)$. For example, the foreground probability of a pixel in the new frame is computed by transforming the foreground probability map in the previous frame using the motion vector. When more than one location in the old frame is transformed to the same location in the new frame, the average probability is assigned to the corresponding position in the new frame. If no pixel is transformed to a location in the new frame, we assign the median of its spatial neighborhood. We combine color and motion likelihoods for a final measurement map as the product of the two factors, which is given by

$$p(M|\mathbf{y}_i^t) = p(M^{color}|\mathbf{y}_i^t)\cdot p(M^{motion}|\mathbf{y}_i^t). \tag{9}$$

The integration of motion for the measurement process is particularly helpful to avoid distractions toward background objects visually similar to the target. In practice we use a simple optical flow estimation algorithm based on the Lucas-Kanade method [1].

We also introduce a new geometric prior $p(\mathbf{\Phi}_i^t)$. The new prior favors level-set functions which are close to a signed distance function. In addition to the standard constraint in the size of the gradient [2, 13], we also constrain its direction. The geometric prior $p(\mathbf{\Phi}_i)$ is defined as

$$p(\mathbf{\Phi}_i) \equiv p_m(\mathbf{\Phi}_i)p_d(\mathbf{\Phi}_i), \tag{10}$$

where p_m and p_d are the magnitude and direction term, respectively. Each term is given by

$$p_m(\mathbf{\Phi}_i) = \frac{1}{\sigma_{m,i}\sqrt{2\pi}}\exp\left(-\frac{(|\nabla\mathbf{\Phi}_i|-1)^2}{2\sigma_{m,i}^2}\right) \tag{11}$$

$$p_d(\mathbf{\Phi}_i) = \frac{1}{\sigma_{d,i}\sqrt{2\pi}}\exp\left(-\frac{\left(\alpha_i^\top\nabla\tilde{\mathbf{\Phi}}_i-1\right)^2}{2\sigma_{d,i}^2}\right), \tag{12}$$

where α_i is the direction of local center of mass around $\mathbf{x}_i$, $\nabla\tilde{\mathbf{\Phi}}_i$ is normalized gradient of $\mathbf{\Phi}_i$ and $\sigma_{m,i}$ and $\sigma_{d,i}$ describe uncertainty of each pixel. We favor gradient directions of level-set function that coincide with inward direction to the human body. Such prior tends to yield smoother level-set functions and human boundaries.

We can now proceed to optimize the objective function with respect to the level-set function $\mathbf{\Phi}$. The optimization problem is equivalent to maximizing the log-likelihood:

$$\log\left(p(\mathbf{\Phi}|\mathbf{x},\mathbf{y})\right) \propto \sum_{i=1}^{N}\log\left(p(\mathbf{x}_i|\mathbf{\Phi}_i,\mathbf{y}_i)\right) - \log p(\mathbf{\Phi}_i), \tag{13}$$

and the optimization is performed iteratively by gradient ascent method with the following update:

$$\begin{aligned} &\frac{\partial\log\left(p(\mathbf{\Phi}_i|\mathbf{x}_i,\mathbf{y}_i)\right)}{\partial\mathbf{\Phi}_i} = \\ &\frac{\delta_\epsilon(P_f-P_b)}{p(\mathbf{x}_i|\mathbf{\Phi}_i,\mathbf{y}_i)} - \left(\frac{\partial\log p_m(\mathbf{\Phi}_i)}{\partial\mathbf{\Phi}_i} + \frac{\partial\log p_d(\mathbf{\Phi}_i)}{\partial\mathbf{\Phi}_i}\right). \end{aligned} \tag{14}$$

3.3. Temporally Coherent Global Optimization

We have described the process of obtaining top-down human region estimates and its efficient propagation using bottom-up cues. However, the estimated human volumes in the previous step may not be reliable due to abrupt changes of a target object, falling in local optima, weak features, and so on. To overcome these issues, we employ a global optimization that integrates temporal information more tightly. This is achieved by introducing explicit dependencies between temporally adjacent frames and jointly optimizing the level-set functions at all frames simultaneously. Such dependencies favor the extraction of contours that are more

[1] Our level-set evolution is based on the algorithm in [2], where a more detailed presentation is available.

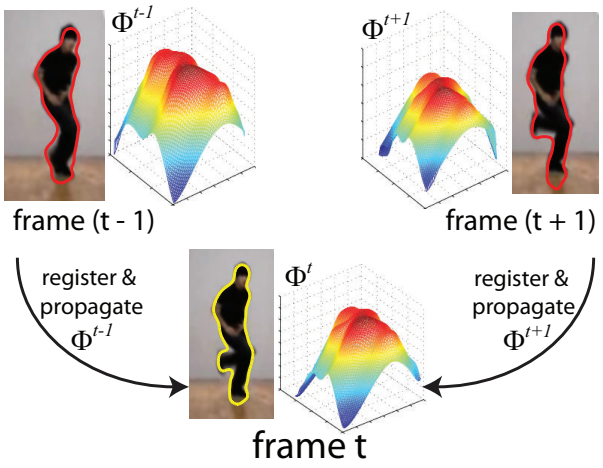


Figure 4. Temporally coherent contour extraction. Our formulation globally optimizes the level-set functions at all frames simultaneously. The level-set function that represents the human boundary is propagated in both directions, forward and backward, which yields a temporally coherent and accurate boundary.

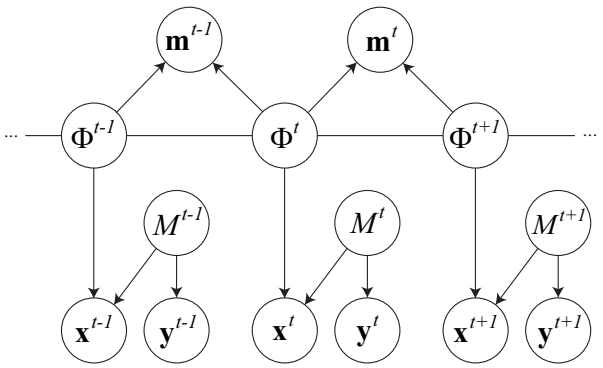


Figure 5. Graphical model for the temporally coherent global optimization. Circles indicate random variables, arrows indicate conditional dependencies and undirected links express mutual dependencies. Image locations are represented by $\mathbf{x}$, image observations by $\mathbf{y}$ and appearance models by M. The level-set function that implicitly defines the human contour is represented by $\mathbf{\Phi}$. The motion information that registers the level-set function across frames is indicated by $\mathbf{m}$. Superscripts indicate time step. By introducing dependencies between the level-set functions from adjacent frames, we can jointly estimate the optimal $\mathbf{\Phi}^{1:T}$ that exhibits temporal consistency and better extracts the the human boundary.

accurate and temporally coherent. As illustrated in Fig. 4, the additional dependencies lead to a bidirectional propagation of the extracted contours among adjacent frames.

The set of $\mathbf{\Phi}^{1:T}$ resulting from the process described in previous sections provides initial level set estimates for the following optimization procedure. We introduce a graphical model to encode the dependencies between $\mathbf{\Phi}^t$ and ($\mathbf{\Phi}^{t+1}$, $\mathbf{\Phi}^{t-1}$), which is shown in Fig. 5. We obtain the globally optimal $\mathbf{\Phi}^{1:T}$ by temporal belief propagation in an iterative message passing procedure.

A new objective function given by Eq. 15, defines the optimization problem. The first factor in Eq. 15 specifies the estimations from individual frames, and the second factor defines the relationship between adjacent frames that contributes to the temporal coherence of the extracted volume.

The message for temporal propagation between the frame t and $t+1$ is defined as

$$\Psi(\mathbf{\Phi}_i^t, \mathbf{\Phi}_i^{t+1}) = \exp\left(-\frac{(\mathbf{\Phi}_i^t - \mathbf{\Phi}_i^{t+1})^2}{\sigma_m^2}\right), \qquad (16)$$

and the update message is

$$\frac{\partial \log\left(\Psi(\mathbf{\Phi}_i^t, \mathbf{\Phi}_i^{t+1})\right)}{\partial \mathbf{\Phi}_i^t} = -\frac{2\left(\mathbf{\Phi}_i^t - \mathbf{\Phi}_i^{t+1}\right)}{\sigma_m^2}, \qquad (17)$$

which favors the temporal consistency of the human motion boundaries. Note that messages are received at frame t from both directions, from frame $t-1$ and $t+1$. The gradient ascent update including the messages for temporal consistency is obtained by the sum of the terms in Eq. 14 and the messages for forward and backward update related to Eq. 17. After the iterative procedure described in Sec 3.2 converges for each frame, we update the messages in Eq. 16 and synchronously pass it to neighboring frames, which is repeated until global convergence. Note that, in the above formulation, the two level-set functions are properly registered by a rigid transformation. Such registration accounts for global rigid motion across frames, gives a better prior to the new frame, and reduces level-set iterations. When all the level-set functions $\mathbf{\Phi}^{1:T}$ are converged, the human volume is finally given by the set of points such that $\mathbf{\Phi}_i^t > 0$.

4. Experiments

We evaluate our method in terms of its segmentation accuracy on annotated frames. The *YouTube* dataset from [14] is utilized first to compare our algorithm with [14]. It is a very challenging dataset that has 50 sequences containing unknown and arbitrary camera motion, cluttered background, motion blur, compression artifacts, etc. The precision and recall are computed based on this dataset for three different algorithms—our full system, our method without global optimization and the method in [14]. Table 1 summarizes the experimental results of three systems, and shows that our methods even without global optimization improve both precision and recall significantly. We attribute this improvement to the ability of our tracker to leverage salient bottom-up cues for human/background separation that are constrained by effective top-down template-driven segmentation. Our temporally coherent optimization process further improves the precision of the system by integrating information across time. The sample comparative results are provided in Fig. 7.

$$p(\mathbf{\Phi}_i^{1:T}|\mathbf{x}^{1:T},\mathbf{y}^{1:T}) = \left[\prod_{t=1}^{T}\frac{1}{p(\mathbf{x}_t)}\sum_{M^t}p(\mathbf{x}_i^t|\mathbf{\Phi}_i^t,M^t)p(M^t|\mathbf{y}_i^t)\right]p(\mathbf{\Phi}_i^{1:T})$$

$$= \underbrace{\left[\prod_{t=1}^{T}\frac{1}{p(\mathbf{x}_t)}\sum_{M^t}p(\mathbf{x}_i^t|\mathbf{\Phi}_i^t,M^t)p(M^t|\mathbf{y}_i^t)\right]}_{\text{Pixel-wise likelihood}}\underbrace{\left[\prod_{t=1}^{T-1}\Psi(\mathbf{\Phi}_i^t,\mathbf{\Phi}_i^{t+1})\right]}_{\text{Temporal consistency}}\underbrace{\left[\prod_{t=1}^{T}p(\mathbf{\Phi}_i^t)\right]}_{\text{Geometric prior}} \quad (15)$$

We created a much larger dataset, also composed of videos downloaded from *YouTube*, and several examples of human body extraction are presented in Fig. 8.

Method	Prec	Rec	F-score
Full model	0.74	0.75	0.74
Full model without global opt.	0.62	0.76	0.68
Niebles et al. [14]	0.57	0.44	0.50

Table 1. Experimental results on the *YouTube* dataset from [14]. The segmentation of humans in videos is evaluated as a retrieval problem. Ground truth consists of a set of over 180 masks that correspond to the human regions in selected frames from the dataset. For each retrieved mask, a precision is computed as the area of the intersection of the retrieved and ground-truth mask over the area of the retrieved mask; whereas recall is the area of the intersection over the area of the ground-truth mask.

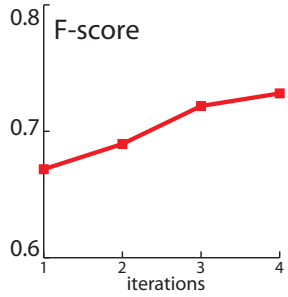


Figure 6. Segmentation accuracy vs number of global iterations. Iteratively propagating the contours across multiple frames helps reduce sporadic artifacts and produces a more temporally smooth human motion boundary.

Our model is not only more accurate in moving human extraction, but also computationally much more efficient. In our implementation with Matlab, one boundary is obtained in less than 50 ms per bottom-up propagation.Similarly, a top-down step with template-driven segmentation takes about the same time. In practice, we apply the top-down process to 20% of the pedestrian detections per person in the video. We use a set of 100 templates in the pedestrian silhouette database. The method in [14] runs in more than 20 seconds per frame per person on similar hardware.

5. Conclusion and Discussion

We have demonstrated a technique to efficiently extract moving humans from challenging sequences, where the top-down modeling provides the shape prior to the bottom-up processing, and the global optimization refines the contour of human bodies. As shown empirically, our method outperforms state-of-the-art techniques at a fraction of the computational cost. This speed allows us to collect a larger set of annotated natural videos containing human motions from *YouTube*. The dataset contains 500 sequences with over 70k frames. Our dataset and video results will be available at: `http://vision.stanford.edu/projects/niebles/cvpr2010`.

Acknowledgments

We thank internal and external reviewers for their comments on our work. This research is partially supported by an NSF CAREER grant (IIS-0845230), a Google research award, and a Microsoft Research fellowship to L.F-F.

References

[1] S. Baker and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *IJCV*, 56(3):221–255, 2004.

[2] C. Bibby and I. Reid. Robust real-time visual tracking using pixel-wise posteriors. In *ECCV*, 2008.

[3] D. Cremers. Dynamical statistical shape priors for level set-based tracking. *IEEE TPAMI*, 28(8):1262–1273, 2006.

[4] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *IJCV*, 72(2):195–215, 2007.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *CVPR*, 2000.

[6] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[7] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[8] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion: Part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1:77–254, 2005.

[9] T. Han, H. Ning, and T. Huang. Efficient nonparametric belief propagation with application to articulated body tracking. In *CVPR*, pages 214–221, 2006.

[10] X. Lan and D. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *ICCV*, volume I, pages 722–729, 2004.

[11] I. Laptev. Improvements of object detection using boosted histograms. In *BMVC*, volume III, pages 949–958, 2006.

[12] C.-S. Lee and A. Elgammal. Modeling view and posture manifolds for tracking. In *ICCV*, 2007.

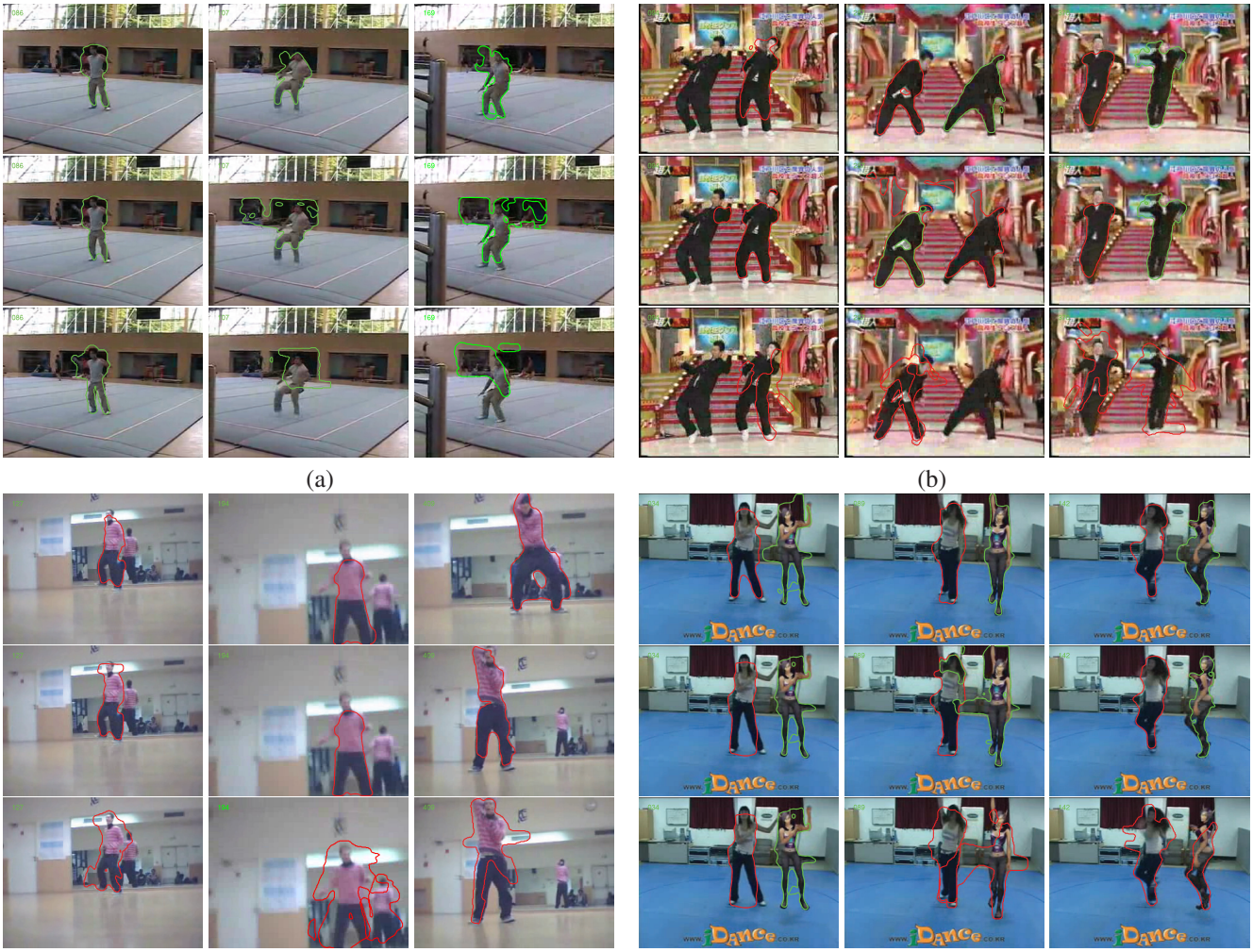


Figure 7. We show 4 example results on the *YouTube* dataset from [14]. For each video, we randomly sample three frames, and compare the extraction results of our method (row 1), with a simplified version of our method (without global optimization and top-down component, in rows 2) and our replication of the method in [14] (rows 3). The outlines of the humans are drawn in color curves. We observe that, in general, our full algorithm performs better than the other two methods.

[13] C. Li, C. Xu, C. Gui, and M. Fox. Level set evolution without re-initialization: A new variational formulation. In *CVPR*, pages 430–436, 2005.

[14] J. Niebles, B. Han, A. Ferencz, and L. Fei-Fei. Extracting moving people from internet videos. In *ECCV*, 2008.

[15] N. Paragios and R. Deriche. Geodesic active regions and level set methods for motion estimation and tracking. *CVIU*, 97(3):259–282, 2005.

[16] D. Ramanan. Learning to parse images of articulated objects. In *NIPS*, 2006.

[17] D. Ramanan, D. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *IEEE TPAMI*, 29(1):65–81, 2007.

[18] X. Ren and J. Malik. Tracking as repeated figure/ground segmentation. In *CVPR*, 2007.

[19] H. Sidenbladh, M. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *ECCV*, 2002.

[20] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, 2004.

[21] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. In *CVPR*, 2005.

[22] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. In *CVPR*, 2001.

[23] D. Tran and A. Sorokin. Human activity recognition with metric learning. In *ECCV*, 2008.

[24] L. Wang, J. Shi, G. Song, and I. Shen. Object detection combining recognition and segmentation. In *ACCV*, volume I, pages 189–199, 2007.

[25] A. Yilmaz, X. Li, and M. Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE TPAMI*, 26(11):1531–1536, 2004.

Figure 8. We show results on 8 gymnastics sequences from an assortment of about 500 *YouTube* videos. The players in the video exhibit a rich variety of challenging motions. Nevertheless, our algorithm is able to retrieve the contour of the person. The colored number at the corner of each image indicates the frame number in the original sequence. The outlines of the humans are drawn in color curves.