Tracking as Repeated Figure/Ground Segmentation

Xiaofeng Ren
Toyota Technological Institute at Chicago
1427 E. 60th Street, Chicago, IL 60637
xren@tti-c.org

Jitendra Malik
Computer Science Division
University of California, Berkeley, CA 94720
malik@cs.berkeley.edu

Abstract

Tracking over a long period of time is challenging as the appearance, shape and scale of the object in question may vary. We propose a paradigm of tracking by repeatedly segmenting figure from background. Accurate spatial support obtained in segmentation provides rich information about the track and enables reliable tracking of non-rigid objects without drifting.

Figure/ground segmentation operates sequentially in each frame by utilizing both static image cues and temporal coherence cues, which include an appearance model of brightness (or color) and a spatial model propagating figure/ground masks through low-level region correspondence. A superpixel-based conditional random field linearly combines cues and loopy belief propagation is used to estimate marginal posteriors of figure vs background. We demonstrate our approach on long sequences of sports video, including figure skating and football.

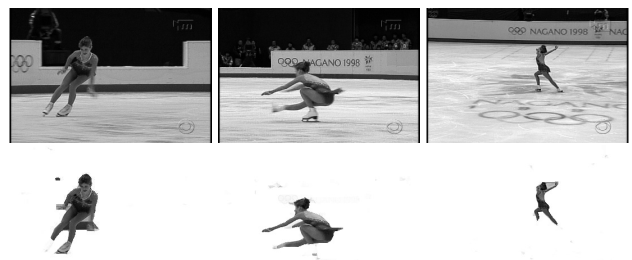


Figure 1. Frame #1, #2020 and #2764 from a figure skating sequence (Row 1). We intend to track an object over a long period of time, under substantial variations of object shape, appearance and scale, without a priori knowledge about the object. We take a figure/ground segmentation approach, tracking by sequentially segmenting out the figure in each frame (Row 2).

1. Introduction

Object tracking is a fundamental problem in computer vision and has been a focus of research for many decades. Success has been declared in many limited settings, such as the case of rigid objects or static cameras. Object tracking in its full generality, however, remains a challenging and unsolved problem. Well-known difficulties include non-rigid shape change, lack of distinctive features, complex scenes, occlusion and, last but not the least, the issue of drifting.

Regardless of the tracking paradigm, all trackers explicitly or implicitly maintain several models of temporal coherence, including:

1. An "appearance" model telling us **what** is being tracked; it could be an image patch, a histogram of color and texture, a smooth contour or a collection of local features.

2. A "spatial" model telling us **where** the object currently is; it could be estimated either at low-level (e.g. using optical flow) or high-level (e.g. through linear and non-linear models of dynamics).

Additional information that a tracker maintains may include scale as well as a background model.

It is self-evident that if a tracker knows the accurate support mask of the object, tracking becomes much easier. A spatial model that knows a support mask, rather than just the center, may predict more reliably where parts of the object will be in the future. An appearance model may also be more reliably updated if a support mask is available, with the interference of background clutter greatly reduced.

Most existing approaches to tracking, however, does not compute such a support mask. Many assume that the object in question has a rectangular or elliptical shape . Such a simplifying assumption of support may work well for objects that have an approximate shape of rectangle or ellipse, such as faces or cars. It would have trouble tracking non-rigid objects in cluttered scenes without drifting.

In this paper we propose a paradigm, *tracking* by *repeated figure/ground segmentation*, for tracking an object under large variations of shape, appearance and scale. Figure 1 shows an example of our approach. In each frame of a video, we use a superpixel-based conditional random field to combine both static image cues and models of temporal coherence. Models of temporal coherence include appearance, scale, and spatial support. A soft figure/ground mask is computed by estimating posterior marginal probabilities of figure vs background in the conditional random field.

Tracking becomes easy using segmentation. It makes full use of low-level and mid-level cues and does not require a rigid shape, distinctive local features or unique color. Accurate spatial support enables reliable updates of appearance and scale. We show that we can track complex motions for a long time without drifting, in color and grayscale, without any model of dynamics or a priori knowledge of the object. Segmentations obtained in the process also provide much richer information about the object in motion than just knowing the center.

2. Related Work

Traditional approaches to tracking represent objects as either a collection of local features, a boundary contour, or a color blob. Lucas and Kanade [15] introduced an iterative image registration technique that has been widely applied to tracking local features [27]. Distinctive local features however are not always available, and active contour models were developed [11, 7, 19] to track the boundary of an object and snap to high gradient locations. Incorporating dynamics helps improve tracking, traditionally done with linear models and Kalman Filtering. The Particle Filtering approach of Isard and Blake [10] has a number of advantages over Kalman filtering, being a non-parametric representation that can maintain multi-modal hypotheses. More recently, color or appearance-based tracking [6, 8] has been popular, being robust to occlusion and clutter if the object has a distinctive appearance.

All tracking approaches are subject to the problem of *drifting* as errors gradually accumulate over time. A local feature tracker is susceptible to distractions from occlusions and clutters. Appearance-based trackers that assume a rectangular or elliptical object shape work well for cars and faces [1, 5, 33] but have trouble updating appearance models for non-rigid objects. Not updating the model is a seemingly easy workaround; but it severely limits the potential of the tracker. Heuristics have been proposed to anchor the tracker to its initialization [17].

An alternative way to improve robustness is to incorporate high-level knowledge into tracking [9, 30, 18, 29]. By matching candidate tracks to stored object models, this tracking-by-recognition paradigm avoids drifting into clutter. An example in extreme is 2D or 3D part-based tracking for articulated objects [21, 4, 28, 20]. These approaches require detailed knowledge of the object, for instance the body model of people. In this work we study visual tracking as a low-level and mid-level problem and do not use any part-based model.

Image segmentation is a huge field of research itself and discussing it is beyond the scope of this paper. Figure/ground segmentation with low-level cues only is in general impossible, as the appearance of both the object and the background may be complex; a good knowledge of the object would be required. Recently, interesting work has been done on figure/ground segmentation combining low-level image cues and high-level object knowledge [3, 12, 23]. Comparing to these approaches that train on a collection of images off-line, we utilize cues from temporal coherence available in the tracking setting.

Motion segmentation [32, 26] is another field closely related to the theme of this work. Typically motion segmentation relies on differential motion cues such as optical flow, and focuses on a short span of time when object appearance as well as motion remain consistent. The scenario we study in this work is different: given an initialization, we intend to track an object for a long period of time under large variations of shape, appearance and scale. We take an "on-line" approach and compute figure/ground segmentations sequentially in each frame as it comes in (e.g. [2]).

3. Tracking as Figure/Ground Segmentation

Object tracking is usually considered as an inference problem about where a given object is throughout a video sequence. A typical approach to tracking satisfies itself with knowing the location of object center. We aim for more: we seek an accurate spatial support of the object in each frame.

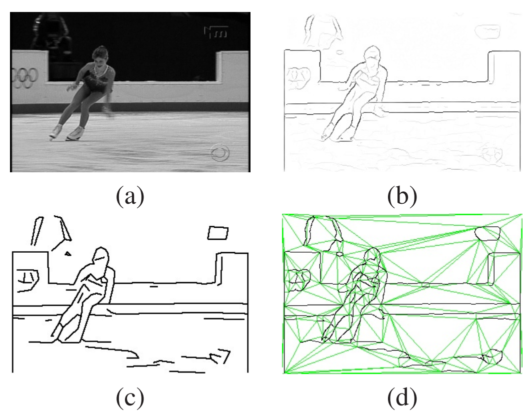


Figure 2. Pre-processing: for each frame (a), we use the *Probability-of-Boundary* operator [16] to compute a soft boundary map (b) that summarizes local brightness, color and texture contrasts. We use a fast image partitioning technique [23], which builds a piecewise straight approximation of the boundary map (c) and applies *constrained Delaunay triangulation* (CDT) to partition the image into a set of triangles (d). In the triangulation, black pixels are edges from the boundary map (b) and green pixels are completions. We use triangles in this triangulation as *superpixels* or atomic units in later stages of processing.

We begin by the processing of each individual frame. Static image cues mostly come in a form of contrast, be-

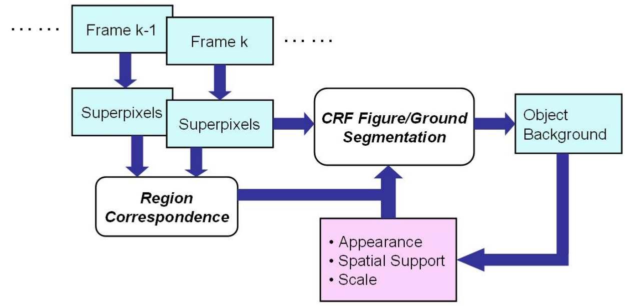


Figure 3. Summary of our approach: images are represented as sets of superpixels. A conditional random field operates independently in each frame to segment figure from background, using both static image cues and temporal coherence cues of appearance, spatial support and scale. A region correspondence carries figure/ground mask in the previous frame into the current frame. Once a soft figure/ground segmentation mask is obtained, the mask is used to update temporal coherence cues.

ing that of brightness, color or texture. We apply the *Probability-of-Boundary* (Pb) operator [16], which returns a soft boundary map that summarizes local contrast cues.

To represent an image in a more compact and perceptually meaningful way, we group pixels in the image into *superpixels* [24], or coherent atomic regions, using a fast image partitioning technique based on *constrained Delaunay triangulation* (CDT) [23]. Figure 2 shows an example of this pre-processing process. The superpixel representation not only reduces computational complexity in later stages of processing, but also makes computation more robust by enforcing consistency inside superpixels.

On top of the CDT triangulation, figure/ground segmentation is done sequentially in each frame. Figure 3 summarizes our approach: segmentation takes both static image cues and tracking cues, i.e. models of temporal coherence. Temporal coherence consists of three parts: an appearance model of brightness or color, telling us what the object looks like; a simple scale model of size and aspect ratio; and a spatial model telling us where the object is expected to be, which carries figure/ground mask in the previous frame to the current one using low-level superpixel correspondence.

We use a conditional random field model [14] to combine cues for figure/ground segmentation. Let $\{T_i\}$ be the collection of triangles, or superpixels, in the current frame. A binary random variable $\mathbf{X}_i$ is associated with each superpixel T_i, $\mathbf{X}_i = 1$ if T_i belongs to the figure, and -1 if the background. We use loopy belief propagation to estimate the marginal posterior probability $F_i = E[\mathbf{X}_i = 1]$, as a soft figure mask. Once we obtain the figure mask, the models of temporal coherence may be updated.

We show results on challenging sequences of sports video. Our figure/ground segmentation approach reliably tracks people under large variations of pose, appearance and scale. Comparing to existing approaches (e.g. [18, 20]) on similar sports data, our segmentation paradigm achieves high tracking performance without using a part-based body model (hence applicable to generic non-rigid objects) or relying on color cues.

4. Temporal Coherence

During the process of tracking, we maintain and update three models of temporal coherence: scale, appearance, and spatial support. These are the internal states of the tracker, representing the tracker's current knowledge about the object being tracked.

For scale, we use a set of three parameters: S, size of the object in pixels, σ_x, median distance to object center in the horizontal direction, and σ_y, median distance to object center in the vertical direction.

For appearance, we model brightness (or color) distributions of both the foreground object and the background, h_F and h_G. We represent both distributions as histograms in the RGB space.

The spatial model tells us where the object is expected to be, given its location in the previous frame. To handle complex motions and non-rigid deformations, we avoid using any dynamics model and seek to transfer figure/ground masks across frames using low-level cues.

Let $\{T_i^{(-1)}\}$ be the set of superpixels in the previous frame, and let $\{F_i^{(-1)}\}$ be the soft mask, or figureness values, associated with $\{T_i^{(-1)}\}$. Let $\{T_j\}$ be the set of superpixels in the current frame. We want to estimate a set of features $\hat{F}_j$, how likely a superpixel T_j in the current frame is part of the figure, based on $\{F_i^{-1}\}$ from the previous frame. This demands correspondence between the two sets of superpixels $\{T_i^{(-1)}\}$ and $\{T_j\}$.

We compute the correspondence by solving a linear transportation problem, analogous to the *Earth Mover's Distance* [25], based on location and brightness (or color). Let $R_i^{(-1)}$ be the mass or size of the superpixels $T_i^{(-1)}$, and R_j the size of superpixels T_j. For any pair of superpixels $(T_i^{(-1)}, T_j)$, let d_{ij} be the distance between centers of $T_i^{(-1)}$ and T_j, and let h_{ij} be the difference in average brightness. we define the cost of the match $(i \rightarrow j)$ as a linear combination $c_{ij} = w_d d_{ij} + w_h h_{ij}$. Let x_{ij} represent the amount of mass being transported from $T_i^{(-1)}$ to T_j, we solve the following linear program:

$$\min L(x) = \sum_{i,j} c_{ij} x_{ij} \tag{1}$$
$$\text{s. t.} \sum_j x_{ij} = R_i^{(-1)}, \sum_i x_{ij} = R_j, \quad x_{ij} \geq 0$$

To avoid high cost matches, we add an additional outlier node for both frames. Once we have the assignments x_{ij},

(a) (b) (c) (d) (e)

Figure 4. An example of temporal coherence cues. (a) is one frame in the skating sequence and (b) is the figure/ground mask we have obtained. (c) shows the next frame. The two frames (a) and (c) are both represented as triangulations, and we compute a region correspondence/assignment between the two sets of triangles. The correspondence is used to transfer the mask (b) into (d), the spatial "prior" that a triangle in frame (d) supports the object, dark meaning high probability. At the same time, we maintain an appearance model, as brightness histograms of both the object and the background. The appearance prior, or the likelihood ratio of the two histograms, is visualized in (e). In this example, when no color information is available, the appearance prior (e) fires on many parts of the background, while the spatial prior (d) is more focused on the figure.

we can estimate $\hat{F}_j$, the spatial "prior" that a superpixel T_j in the current frame belongs to the figure. $\hat{F}_j$ is computed as a weighted average $\hat{F}_j = \sum_i x_{ij} F_i^{(-1)}/R_j$. An example is shown in Figure 4.

After we solve the figure/ground segmentation in the current frame and obtain a soft mask $\{F_j\}$, we update the models of temporal coherence in a straightforward way. For example, we re-estimate the size of the foreground object, $S' = \sum F_j R_j$. The size S is updated as $(1-r)S + rS'$ with a fixed rate r. Other models are similarly updated.

5. Figure/Ground Segmentation

We employ a *conditional random field* (CRF) for figure/ground segmentation. Introduced in [14] as a model for labeling 1D structures in natural language, conditional random fields have become a popular technique in computer vision, being applied to a range of vision problems including labeling man-made structures [13] and object-specific segmentation [23]. A conditional random field provides a general probabilistic framework for discriminative labeling and is especially suitable for combining multiple sources of cues in our figure/ground segmentation problem.

Let $\{T_i\}$ be the set of superpixels comprising the image. Let $\{\mathbf{X}_i\}$ be the binary labels or random variables associated with $\{T_i\}$, $\mathbf{X}_i = 1$ if T_i belongs to the figure, or -1 if the background. A conditional random field for figure/ground segmentation defines a joint distribution of $\mathbf{X} = \{\mathbf{X}_i\}$:

$$P(\mathbf{X}|I;\Theta) = \frac{1}{Z(I,\Theta)} \exp\left\{-\sum \alpha_k f_k(\mathbf{X}, I; \Theta)\right\} \quad (2)$$

where the features f_k are linearly combined in an exponential function, and Z is the normalization factor or the partition function.

5.1. Cues for Figure/Ground

The *Pb* boundary map summarizes local contrasts of brightness, color and texture, and is the only static image cue we use in the model. Let T_i and T_j be a pair of adjacent superpixels in the current frame. If $\mathbf{X}_i = \mathbf{X}_j$, i.e. if they belong to the same segment, there should be no boundary between them and the contrast should be low; vice versa, if $\mathbf{X}_i \neq \mathbf{X}_j$, the contrast should be high. Let Pb_{ij} be the average Pb contrast value along the boundary between the pair, we may define a boundary feature, weighted by the length of this boundary L_{ij}:

$$f_b = \sum_{i,j} L_{ij} \left[\log\left(\frac{Pb_{ij}}{1 - Pb_{ij}}\right) - \tau_b\right] (\mathbf{X}_i \mathbf{X}_j)$$

where τ_b is an offset, roughly corresponding to the average case Pb value.

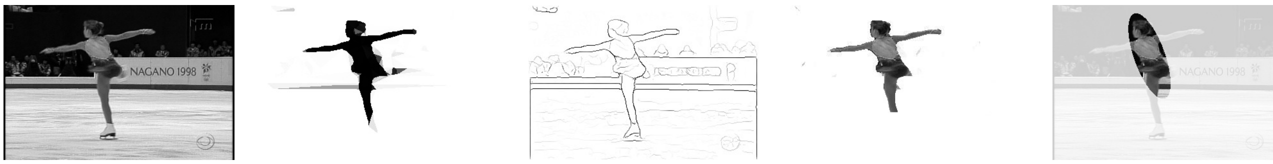
Given the appearance model h_F of the object, a brightness or color histogram, we can calculate $h_F(T_i)$, the average likelihood of the superpixel T_i under the model. Similarly we calculate the average likelihood $h_G(T_i)$ under the background model. Let R_i be the size or area of the superpixel T_i, the likelihood ratio provides the appearance cue:

$$f_c = \sum_i R_i \left[\log\left(\frac{h_F(T_i)}{h_G(T_i)}\right)\right] \mathbf{X}_i$$

As discussed in the last section, the spatial model of temporal coherence provides $\hat{F}_i$, the spatial prior carried over from figure/ground mask in the previous frame. We define the spatial support cue:

$$f_l = \sum_i R_i \left[\log\left(\frac{\hat{F}_i}{1 - \hat{F}_i}\right)\right] \mathbf{X}_i$$

The region correspondence is computed at low-level; hence the spatial model has no notion of the object being connected and likely convex. To keep the foreground mask from falling apart, we compute a tentative object center $(\hat{y}, \hat{x})$ from $\hat{F}_i$, find the average distance $(\hat{d}_y, \hat{d}_x)$ of each superpixel T_i to this center, and normalize it with the scale



(a) (b) (c) (d) (e)

Figure 5. Figure/ground segmentation enables reliable tracking and avoids drifting by combining both temporal coherence cues and static image cues. For the sample frame in (a), we show in (b) a combination of the spatial prior and the appearance prior. The combined prior is more accurate than either of the individual cues (shown in Figure 4). Nevertheless, errors inevitably occur when we transfer information to the new frame. Our segmentation tracker utilizes both the combined temporal prior and static image cues of brightness and texture contrasts, summarized in the contour contrast map (c). The resulting segmentation (d) closely follows the high-contrast contours and corrects errors in the temporal prior, largely reducing the likelihood of drafting. The accurate support mask in (d) is then used to update the object appearance and scale. As a comparison, if one approximates the object support as an ellipse (e), the ellipse cannot match the object perfectly. Updating would be much less reliable using the elliptical support.

parameters σ_x and σ_y to be a distance cue:

$$f_o = \sum_i R_i \left[\sqrt{(\frac{\hat{d}_y}{\sigma_y})^2 + (\frac{\hat{d}_x}{\sigma_x})^2} - \tau_o \right] \mathbf{X}_i$$

And finally, we want to control the total size of the superpixels assigned to the figure. Given an assignment $\{\mathbf{X}_i\} = x_i, x_i \in \{-1, 1\}$, the figure size is $\sum_{i:x_i=1} R_i$, and we want it to be close to the current scale parameter S. Therefore we add a squared penalty term:

$$f_s = \left[S - \sum_{i:\mathbf{X}_i=1} R_i \right]^2 / S^2$$

5.2. Computing Figure/Ground Mask

We use *loopy belief propagation* [31] to solve for F_i, the marginal probabilities of $\mathbf{X}_i$. Messages in the belief propagation are updated sequentially, in a fixed order. Belief propagation is facilitated by the use of superpixels. A superpixel representation greatly reduces the number of variables, and at the same time allows propagation over a long range. Loopy belief propagation converges quickly on the triangulation graphs, typically < 10 iterations.

The potential functions in our model are unary or binary on the variables $\{\mathbf{X}_i\}$, except for one, the scale potential f_s which involves all the variables. Scale, after all, is a global parameter and cannot be decomposed into local features.

Updating messages for the scale potential requires the estimation of an expectation, in the following form:

$$E_{\{\mathbf{Y}_j\} \backslash \mathbf{Y}_k} \left[\exp(-\frac{1}{S^2}(S - R_k y_k - \sum_{j:j\neq k} R_j \mathbf{Y}_j)^2) \right]$$

where $y_k \in \{0, 1\}$ is a constant and $\{\mathbf{Y}_j\} \in \{0, 1\}$ are Bernoulli random variables. This expectation is obviously too costly to compute exactly. The random variables $\mathbf{Y}_j$, however, only appears in a sum. We use the central limit theorem to approximate the distribution of the sum $\sum_{j:j\neq k} R_j \mathbf{Y}_j$ as a Gaussian. In such an approximation, the expectations may be solved in close form. We omit the details here.

The scale potential effectively acts as an adaptive gain control mechanism inside the loop of belief propagation. If the current belief states of the superpixels assign too much mass to the figure, the scale potential sends messages to all the superpixels to reduce the mass; if there is not enough mass, the scale potential sends messages to increase it.

6. Experiments

We test our approach on a number of sports sequences: a figure skating sequence of Tara Lipinski, 3117 frames, both in grayscale and color; a skating sequence of Michelle Kwan, 750 frames, in grayscale; and a football sequence, 940 frames, in color. All the images are of resolution 240-by-360. We hand-initialize each sequence with a bounding rectangle. Lacking proper training data with groundtruth, we set the parameters of the model by hand.

Our figure/ground segmentation tracker successfully tracks people through large variations of pose, appearance and scale as well as severe occlusion; sample results are shown in Figure 8. This robustness is due to the tracker's knowledge of multiple sources of information, combining temporal coherence cues and static image cues. Temporal coherence cues, including the appearance and spatial priors, roughly locate the object in a frame; static image cues, including brightness and texture contrasts, correct errors in the priors and refine the support mask to "snap" to object boundaries (see an example in Figure 5). Knowing accurate support of the object also makes it easier to update the temporal coherence models on-the-fly, with the interference of background clutters reduced to a minimum.

In Figure 6 we compare our results with a mean-shift tracker of Zivkovic and Krose [33] , where they used an ellipse to approximate the object shape. Mainly due to variation in pose, the mean-shift tracker gradually drifts and

Frame #1 Frame #153 Frame #153 Frame #918 Frame #918

Figure 6. We compare our algorithm to a mean-shift based tracker [33] on the Lipinski sequence with color. The mean-shift tracker uses an ellipse to approximate object shape. Frame #153: the mean-shift tracker confuses the object with the background, mainly because of non-rigid deformation. Frame #918: the mean-shift tracker completely loses the object. In comparison, our segmentation tracker does not drift and finds accurate figure segmentations.

Frame #1454 Frame #1455 Frame #1455 Frame #1461 Frame #1480

Figure 7. Our segmentation tracker restarts itself after a camera switch between Frame #1454 and #1455. Scale, spatial support and background appearance cues are invalid after a camera switch; the figure appearance and static image cues are still valid. At first the tracker does not know for sure where the object is, hence the figure mask spreading over the image. As time goes on, the tracker accumulates information and gradually focuses back on the object.

loses the figure. With figure/ground segmentation, we can track Lipinski under large variations of pose and scale.

There is one interesting caveat in the Lipinski sequence: on a few occasions, the camera is switched, and the skater appears at a different location with a different background. A camera switch is easy to detect as the raw image difference between adjacent frames would be large.

Figure 7 shows an example of camera switch. Many cues are not valid at a camera switch, such as object location, scale, or background appearance. The tracker relies on the foreground appearance model and static image cues to restart itself. At first, the tracker does not know exactly where the object is, hence probability mass spreading out over the image. After a few frames, however, the "belief" of the tracker converges to the object.

This ability to restart indicates that, with a single figure/ground mask, the tracker can maintain multiple hypotheses, keeping alternatives around when it is not certain. This is common when the tracker runs into an ambiguous region, as we can see in a few places in the grayscale Lipinski track and the football track in Figure 8.

In the football track in Figure 8, we see an example of how our segmentation tracker handles occlusion. On an occasion in the football sequence, the football player is severely occluded, and for about 100 frames only the upper body is visible. Although the tracker does not keep any history, it is able to "re-discover" the lower body after it reappears, when the lower body becomes distinctive enough from the background. Again, this happens because the tracker knows and utilizes both temporal coherence and static image cues.

Successful tracking on these sequences suggests that our approach is capable of handling non-rigid shape, appearance variation, scale change as well as occlusion and background clutter. Moreover, the figure segmentations we obtain are fairly accurate, with arms included in most cases even when they are a few pixels wide and far from the body center, without any knowledge of arms being parts of the human body. These figure segmentations may then be used to "learn" about the object being tracked and apply the knowledge to static image detection [22].

7. Discussion

In this paper we have proposed a figure/ground segmentation approach to object tracking. Instead of assuming a rectangular or elliptical shape, we repeatedly apply a conditional random field model of figure/ground segmentation, and obtain a figure mask in each frame. Such a spatial support mask makes tracking more robust and less susceptible to drifting. We show successful tracks on long sports video with large variations in shape, appearance and scale.

In this work we have restricted ourselves to a simple set of cues as well as a straightforward superpixel correspondence algorithm. Our figure/ground segmentation framework is general and conceptually there is no difficulty in combining additional cues into the conditional random field, such as shape matching, local/point feature correspondence, or mid-level cues like the smoothness of boundaries or T-junctions. It is also conceivable that more high-level models may be added, for instance dynamics models of object center and parts, or explicit reasoning about occlusion and multiple object tracking.

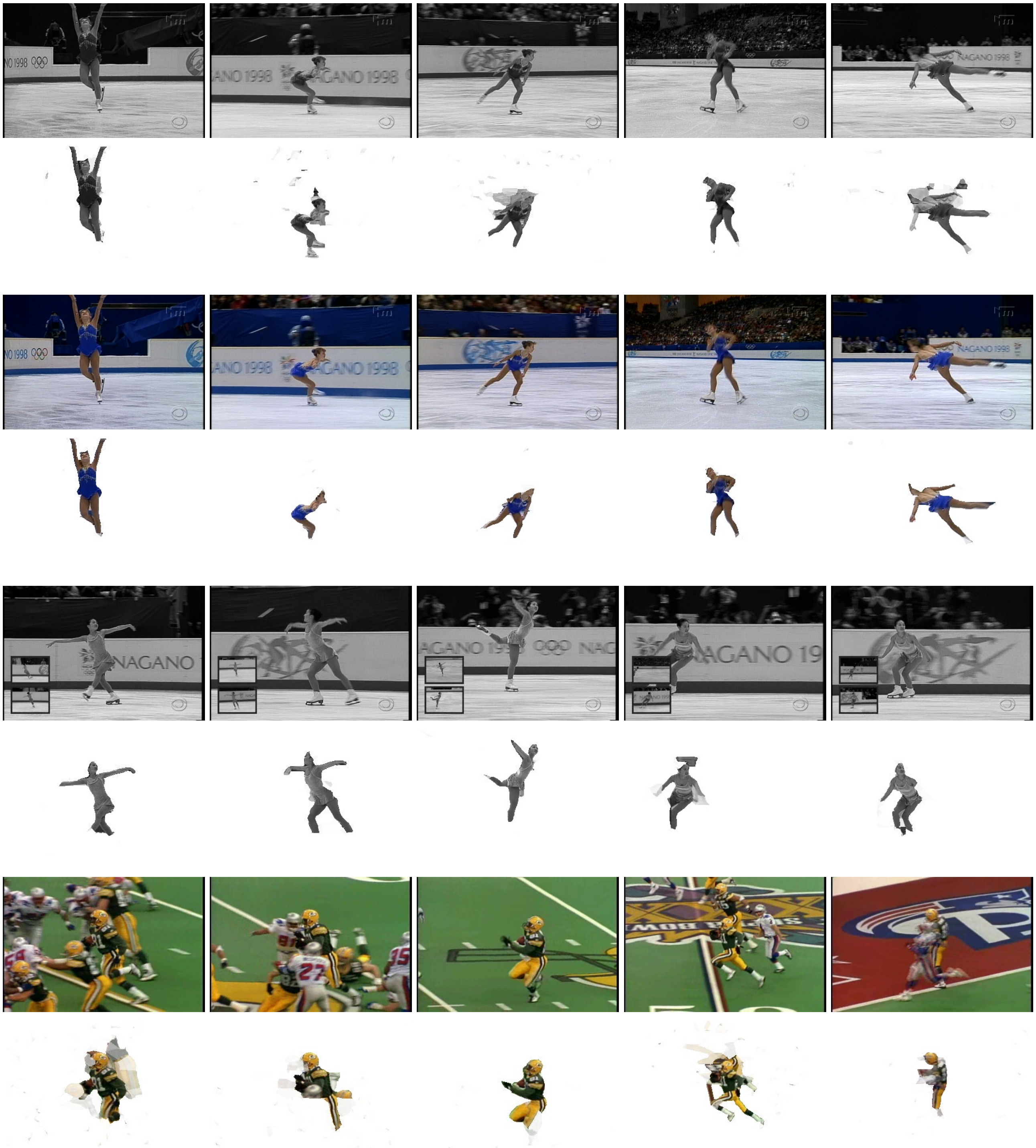


Figure 8. Sample results on four sequences of sports video: frame #88, #315, #840,#1624 and #1943 for the Tara Lipinski skating sequence, both in grayscale and color; frame #99, #194, #406, #504 and #552 for the Michelle Kwan sequence in grayscale; and frame #167, #222, #294, #451 and #882 for the football sequence. Results are shown as the original image masked by the posterior probability of figureness. The mask is soft; we can see blending in a few places when there is ambiguity. Our conceptually simple figure/ground approach reliably tracks and segments people in these video, even in the grayscale cases when no distinctive color or local features are available. It also nicely handles occlusion and background clutter.

References

[1] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR*, pages 232–237, 1998. 2

[2] M. Black. Combining intensity and motion for incremental segmentation and tracking over long image sequences. In *ECCV*, pages 485–493, 1992. 2

[3] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, volume 2, pages 109–124, 2002. 2

[4] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, pages 8–15, 1998. 2

[5] R. Collins and Y. Liu. On-line selection of discriminative tracking features. In *ICCV*, volume 1, pages 346–352, 2003. 2

[6] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, volume 2, pages 142–149, 2000. 2

[7] T. Cootes, C. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 2

[8] A. Elgammal and L. Davis. Probabilistic framework for segmenting people under occlusion. In *ICCV*, volume 2, pages 145–152, 2001. 2

[9] D. P. Huttenlocher, J. J. Noh, and W. J. Rucklidge. Tracking non-rigid objects in complex scenes. In *ICCV*, pages 93–101, 1993. 2

[10] M. Isard and A. Blake. Condensation: Conditional density propagation for visual tracking. *Int'l. J. Comp. Vision*, 29(1):5–28, 1998. 2

[11] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int'l. J. Comp. Vision*, 1(4):321–331, 1988. 2

[12] M. P. Kumar, P. Torr, and A. Zisserman. Obj cut. In *CVPR*, volume 1, pages 18–25, 2005. 2

[13] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *ICCV*, pages 1150–1159, 2003. 4

[14] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001. 3, 4

[15] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of IJCAI*, pages 674–679, 1981. 2

[16] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using brightness and texture. In *NIPS*, 2002. 2, 3

[17] I. Matthews, T. Ishikawa, and S. Baker. The template update problem. In *BMVC*, 2003. 2

[18] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV*, volume 3, pages 666–680, 2002. 2, 3

[19] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Trans. PAMI*, 22(3):266–280, 200. 2

[20] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *CVPR*, volume 2, pages 467–474, 2003. 2, 3

[21] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, pages 612–617, 1995. 2

[22] X. Ren. Learning and matching line aspects for articulated objects. In *CVPR*, 2007. 6

[23] X. Ren, C. Fowlkes, and J. Malik. Cue integration in figure/ground labeling. In *NIPS*, 2005. 2, 3, 4

[24] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, volume 1, pages 10–17, 2003. 3

[25] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *ICCV*, pages 59–66, 1998. 3

[26] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *ICCV*, pages 1154–1160, 1998. 2

[27] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600, 1994. 2

[28] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, pages 702–718, 2000. 2

[29] C. Tomasi, S. Petrov, and A. Sastry. 3d tracking = classification + interpolation. In *ICCV*, pages 1441–1448, 2003. 2

[30] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, volume 2, pages 50–57, 2001. 2

[31] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, pages 1–41, 2000. 5

[32] Y. Weiss and E. Adelson. A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models. In *CVPR*, pages 321–6, 1996. 2

[33] Z. Zivkovic and B. Krose. An em-like algorithm for color-histogram-based object tracking. In *CVPR*, volume 1, pages 798–803, 2004. 2, 5, 6