Joint Tracking and Segmentation of Objects using Graph Cuts

Aurélie Bugeau and Patrick Pérez

IRISA / INRIA,
Campus de Beaulieu,
35 042 Rennes Cedex, France
{aurelie.bugeau,perez}@irisa.fr

Abstract. This paper presents a new method to both track and segment objects in videos. It includes predictions and observations inside an energy function that is minimized with graph cuts. The min-cut/max-flow algorithm provides a segmentation as the global minimum of the energy function, at a modest computational cost. Simultaneously, our algorithm associates the tracked objects to the observations during the tracking. It thus combines “detect-before-track” tracking algorithms and segmentation methods based on color/motion distributions and/or temporal consistency. Results on real sequences are presented in which the robustness to partial occlusions and to missing observations is shown.

1 Introduction

In recent and thorough review on tracking techniques [20], tracking methods are divided into three categories : point tracking, silhouette tracking and kernel tracking. These three categories can be recast as "detect-before-track" tracking, dynamic segmentation and tracking based on distributions (color in particular).

The principle of "detect-before-track" methods is to match the tracked objects with observations provided by an independent detection module. This tracking can be done using deterministic methods or probabilistic methods. Deterministic methods correspond to matching by minimizing a distance based on certain descriptors of the object. Probabilistic methods allow taking measurement uncertainties into account. They are often based on a state space model of the object properties.

Dynamic segmentation corresponds to a succession of segmentations. These silhouette tracking methods usually make evolve an initial contour to its new position in the current frame. This can be done using a state space model defined in terms of shape and motion parameters of the contour [9], [16] or by the minimization of a contour-based energy function. In latter case, the energy function includes temporal information in the form of either the temporal gradient (optical flow)[1], [7], [13] or appearance statistics originated from the object and the background regions in previous images [15] [19]. In [18] the authors use graph cuts to minimize such an energy function. The advantages of min-cut/max-flow optimization are its low computational cost, the fact that it converges to a global minimum (as opposed to local methods that get stuck in local minima) and that no *a priori* on the global shape model is needed.

The last group of methods is based on kernel tracking. The best location for a tracked object in the current frame is the one for which some feature distribution (*e.g.*, color) is the closest to the reference one. The most used method in this class is the "mean shift" tracker [5], [6]. Graph cuts have also been used for illumination invariant kernel tracking in [8].

These three types of tracking techniques have different advantages and limitations, and can serve different purposes. "Detect-before-track" methods can deal with the entries of new objects and the exit of existing ones. They use external observations that, if they are of good quality, might allow robust tracking and possibly accurate segmentations. Silhouette tracking has the advantage of directly providing the segmentation of the tracked object. With the use of recent graph cuts techniques, convergence to the global minimum is obtained for modest computational cost. Finally kernel tracking methods, by capturing global color distribution of a tracked object, allow robust tracking at low cost in a wide range of color videos. In this paper, we address the problem of multiple objects tracking and segmentation by combining the advantages of the three classes of approaches. We suppose that, at each instant, the objects of interest are approximately known as the output of a preprocessing algorithm. Here, we use a simple background subtraction but more complex alternative techniques could be applied. These objects are the "observations" as in Bayesian filtering. At each time the extracted objects are propagated using their associated optical flow, which gives the predictions. Intensity and motion distributions are computed on the objects of previous frame. For each tracked object, an energy function is defined using the observations and these distributions, and minimized using graph cuts. The use of graph cuts directly gives the segmentation of the tracked object in the new frame. Our algorithm also deals with the introduction of new objects and their associated trackers.

In section 2, an overview of the method and the notations is given. The graph and associated energy function are then defined in section 3. Experimental results are shown in section 4, where we demonstrate in particular the robustness of our technique in case of partial occlusions and missing observations. We conclude in section 5.

2 Principle and Notations

Before explaining the scheme of the algorithm, the notations and definitions must be introduced for the objects and the observations.

2.1 Notations

In all this paper, $\mathcal{P}$ will denote the set of N pixels of a frame from an input sequence of images. To each pixel s of the image at time t is associated a feature vector $\mathbf{z}_{s,t} = (\mathbf{z}_{s,t}^{(C)}, \mathbf{z}_{s,t}^{(M)})$, where $\mathbf{z}_{s,t}^{(C)}$ is a 3-dimensional vector in RGB color space and $\mathbf{z}_{s,t}^{(M)}$ is a 2-dimensional optical flow vector. The optical flow is computed using Lucas and Kanade algorithm [12] with incremental multiscale implementation.

We assume that, at time t, k_t objects are tracked. The i^{th} object at time t is denoted as $\mathcal{O}_t^{(i)}$ and is defined as a set of pixels, $\mathcal{O}_t^{(i)} \subset \mathcal{P}$. The pixels of a frame not belonging to the object $\mathcal{O}_t^{(i)}$ belong to the "background" of this object.

The goal of this paper is to perform both segmentation and tracking to get the object $\mathcal{O}_t^{(i)}$ corresponding to the object $\mathcal{O}_{t-1}^{(i)}$ of previous frame. Contrary to sequential segmentation techniques [10], [11], [14], we bring in object-level "observations". They may be of various kinds (*e.g.*, boxes or masks obtained by a class specific object detector, or static motion/color detectors). Here we consider that these observations come from a preprocessing step of background subtraction. Each observation amounts to a connected component of the foreground map after background subtraction (figure 1). The connected components are obtained using the "gap/mountain" method described in [17] and ignoring small objects. For the first frame, the tracked objects are initialized as the observations themselves. We assume that, at each time t, there are m_t observations. The j^{th} observation at time t is denoted as $\mathcal{M}_t^{(j)}$ and is defined as a set of pixels, $\mathcal{M}_t^{(j)} \subset \mathcal{P}$. Each observation is characterized by its mean feature:

$$\overline{\mathbf{z}}_t^{(j)} = \frac{\sum_{s \in \mathcal{M}_t^{(j)}} \mathbf{z}_{s,t}}{|\mathcal{M}_t^{(j)}|} \quad . \tag{1}$$

(a) (b) (c)

Fig. 1. Observations obtained with background subtraction and object isolation. (a) Reference frame. (b) Current frame (c) Result of background subtraction and derived object detection (two objects with red bounding boxes).

2.2 Principle of the algorithm

The principle of our algorithm is as follows. A prediction $\mathcal{O}_{t|t-1}^{(i)}$ is made for each object i of time $t-1$. Once again, the prediction is a set of pixels, $\mathcal{O}_{t|t-1}^{(i)} \subset \mathcal{P}$. We denote as $\mathbf{d}_{t-1}^{(i)}$ the mean, over all pixels of the object at time $t-1$, of optical flow vectors:

$$\mathbf{d}_{t-1}^{(i)} = \frac{\sum_{s \in \mathcal{O}_{t-1}^{(i)}} \mathbf{z}_{s,t-1}^{(M)}}{|\mathcal{O}_{t-1}^{(i)}|} \quad . \tag{2}$$

The prediction is obtained by translating each pixel belonging to $\mathcal{O}_{t-1}^{(i)}$ by this average optical flow:

$$\mathcal{O}_{t|t-1}^{(i)} = \{s + \mathbf{d}_{t-1}^{(i)}, s \in \mathcal{O}_{t-1}^{(i)}\} \quad . \tag{3}$$

Using this prediction, the new observations, as well as color and motion distributions of $\mathcal{O}_{t-1}^{(i)}$, a graph and an associated energy function are built. The energy is minimized using min-cut/max-flow algorithm [4], which gives the new segmented object at time t, $\mathcal{O}_t^{(i)}$. The minimization also provides the correspondences of the object $\mathcal{O}_{t-1}^{(i)}$ with all the available observations. The sketch of our algorithm is presented in figure 2.

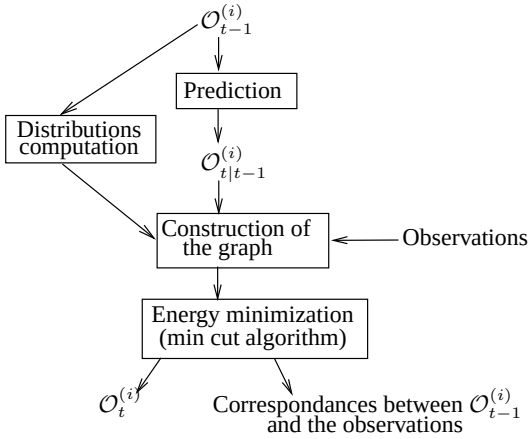
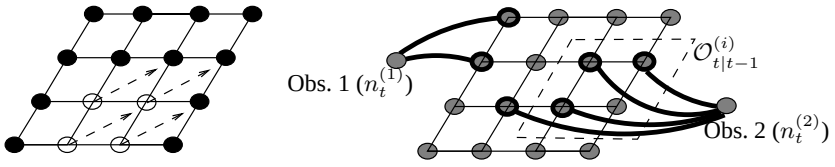


Fig. 2. Principle of the algorithm

3 Energy function

We define one tracker for each object. To each tracker corresponds, for each frame, one graph (figure 3) and one energy function that is minimized using the min-cut/max-flow algorithm [4]. Details of the approach are given in the following subsections.



Result for object i at time t-1 Graph for object i at time t

Fig. 3. Description of the graph. The left figure is the result of the energy minimization at time $t-1$. White nodes are labeled as object and black nodes as background. The optical flow vectors for the object are the dashed line arrows. The right figure shows the graph at time t. Two observations are available. Thick nodes correspond to the observations. See text for explanations and details on the edges.

3.1 Graph

The undirected graph $G_t = (\mathcal{V}_t, \mathcal{E}_t)$ is defined as a set of nodes $\mathcal{V}_t$ and a set of edges $\mathcal{E}_t$. The set of nodes is divided in two subsets. The first subset is the set of the N

pixels of the image grid $\mathcal{P}$. The second subset corresponds to the observations : to each observation set $\mathcal{M}_t^{(j)}$ is associated a node $n_t^{(j)}$. The set of nodes thus reads $\mathcal{V}_t = \mathcal{P} \bigcup_{j=1}^{m_t} n_t^{(j)}$. The set of edges is divided in two subsets: $\mathcal{E}_t = \mathcal{E}_\mathcal{P} \bigcup_{j=1}^{m_t} \mathcal{E}_{\mathcal{M}_t^{(j)}}$. The set $\mathcal{E}_\mathcal{P}$ represents all unordered pairs $\{s, r\}$ of neighboring elements of $\mathcal{P}$ (thin black edges on right part of figure 3), and $\mathcal{E}_{\mathcal{M}_t^{(j)}}$ is the set of unordered pairs $\{s, n_t^{(j)}\}$, with $s \in \mathcal{M}_t^{(j)}$ (thick black edges on right part of figure 3).

Segmenting the object $\mathcal{O}_t^{(i)}$ amounts to assigning a label $l_{s,t}^{(i)}$, either background, "bg", or object, "fg", to each pixel node s of the graph. Associating observations to tracked objects amounts to assigning a binary label ("bg" of "fg") to each observation node. The set of all the node labels is $L_t^{(i)}$.

3.2 Energy

An energy function is defined for each object at each time. It is composed of unary data terms $R_{s,t}^{(i)}$ and smoothness binary terms $B_{s,r,t}^{(i)}$:

$$E_t^{(i)}(L_t^{(i)}) = \sum_{s \in \mathcal{V}_t} R_{s,t}^{(i)}(l_{s,t}^{(i)}) + \lambda \sum_{\{s,r\} \in \mathcal{E}_t} B_{s,r,t}^{(i)}(1 - \delta(l_{s,t}^{(i)}, l_{r,t}^{(i)})) \ . \tag{4}$$

Following [2], the parameter λ is set to 20.

Data term The data term can be decomposed into two parts. While the first one corresponds to the prediction, the second corresponds to the observations. For all the other nodes, we do not want to give any *a priori* on whether the node is part of the object or the background (labeling of these nodes will then be controlled by the influence of neighbors via binary terms).The first part of energy in (4) reads :

$$\sum_{s \in \mathcal{V}_t} R_{s,t}^{(i)}(l_{s,t}^{(i)}) = \sum_{s \in \mathcal{O}_{t|t-1}} -\ln(p_1^{(i)}(s, l_{s,t}^{(i)})) + \sum_{j=1}^{m_t} -\ln(p_2^{(i)}(n_t^{(j)}, l_{n_t^{(j)},t})) \ . \tag{5}$$

The new object should be close in terms of motion and color to the object at previous time. The color and motion distributions of the object and the background are then defined for previous time. The distribution $p_{t-1}^{(i,C)}$ for color, respectively $p_{t-1}^{(i,M)}$ for motion, is a Gaussian mixture model fitted to the set of values $\{\mathbf{z}_{s,t-1}^{(C)}\}_{s \in \mathcal{O}_{t-1}^{(i)}}$, respectively $\{\mathbf{z}_{s,t-1}^{(M)}\}_{s \in \mathcal{O}_{t-1}^{(i)}}$. Under independency assumption for color and motion, the final distribution for the object is :

$$p_{t-1}^{(i)}(\mathbf{z}_{s,t}) = p_{t-1}^{(i,C)}(\mathbf{z}_{s,t}^{(C)}) \, p_{t-1}^{(i,M)}(\mathbf{z}_{s,t}^{(M)}) \ . \tag{6}$$

The two distributions for the background are $q_{t-1}^{(i,M)}$ and $q_{t-1}^{(i,C)}$. The first one is a Gaussian mixture model built on the set of values $\{\mathbf{z}_{s,t-1}^{(M)}\}_{s \in \mathcal{P} \setminus \mathcal{O}_{t-1}^{(i)}}$. The second one is a uniform model on all color bins. The final distribution for the background is :

$$q_{t-1}^{(i)}(\mathbf{z}_{s,t}) = q_{t-1}^{(i,C)}(\mathbf{z}_{s,t}^{(C)}) \, q_{t-1}^{(i,M)}(\mathbf{z}_{s,t}^{(M)}) \ . \tag{7}$$

The likelihood p_1, which is applied to the prediction node in the energy function, can now be defined as :

$$p_1^{(i)}(s,l) = \begin{cases} p_{t-1}^{(i)}(\mathbf{z}_{s,t}) \text{ if } l = \text{“fg”} \\ q_{t-1}^{(i)}(\mathbf{z}_{s,t}) \text{ if } l = \text{“bg”} \end{cases} . \tag{8}$$

An observation should be used only if it corresponds to the tracked object. Therefore, we use the same distribution for p_2 as for p_1. However we do not evaluate the likelihood of each pixel of the observation mask but only the one of its mean feature $\overline{\mathbf{z}}_t^{(j)}$. The likelihood p_2 for the observation node $n_t^{(j)}$ is defined as

$$p_2^{(i)}(n_t^{(j)},l) = \begin{cases} p_{t-1}^{(i)}(\overline{\mathbf{z}}_t^{(j)}) \text{ if } l = \text{“fg”} \\ q_{t-1}^{(i)}(\overline{\mathbf{z}}_t^{(j)}) \text{ if } l = \text{“bg”} \end{cases} . \tag{9}$$

Binary term Following [3], the binary term between neighboring pairs of pixels $\{s,r\}$ of $\mathcal{P}$ is based on color gradients and has the form

$$B_{s,r,t}^{(i)} = \frac{1}{\text{dist}(s,r)} e^{-\frac{\|\mathbf{z}_{s,t}^{(C)} - \mathbf{z}_{r,t}^{(C)}\|^2}{\sigma_T^2}} . \tag{10}$$

As in [2], the parameter σ_T is set to

$$\sigma_T = 4 * \langle (\mathbf{z}_{s,t}^{(i,C)} - \mathbf{z}_{r,t}^{(i,C)})^2 \rangle \tag{11}$$

where $\langle . \rangle$ denotes expectation over a box surrounding the object.
For edges between the grid $\mathcal{P}$ and the observations nodes, the binary term is similar :

$$B_{s,n_t^{(j)},t}^{(i)} = e^{-\frac{\|\mathbf{z}_{s,t}^{(C)} - \overline{\mathbf{z}}_t^{(j,C)}\|^2}{\sigma_T^2}} . \tag{12}$$

Energy minimization The final labeling of pixels is obtained by minimizing the energy defined above :

$$\hat{L}_t^{(i)} = \arg\min E_t^{(i)}(L_t^{(i)}) . \tag{13}$$

Finally this labeling gives the segmentation of the object $\mathcal{O}_t^{(i)}$, defined as :

$$\mathcal{O}_t^{(i)} = \{ s \in \mathcal{P} : \hat{l}_{s,t}^{(i)} = \text{“fg”} \} . \tag{14}$$

3.3 Creation of new objects

One advantage of our method comes from the nodes corresponding to the observations. It allows the use of observations to track and segment the objects at time t as well as to establish the correspondence between an object currently tracked and all the candidate objects imperfectly detected in current frame. If, after the energy minimization for an object i, a node $n_t^{(j)}$ is labeled as “fg” it means that there is a correspondence between the object and the observation. If for all the objects, an observation node is labeled as “bg” after minimizing the energies, then the corresponding observation does not match any objects. In this case, a new object is created and is initialized as this observation.

4 Experimental Results

In this section results that validate the algorithm are presented. The sequences used are from the PETS 2001 data corpus (data set 1 camera 1 and dataset 3 camera 2), and the PETS 2006 data corpus (sequence 1 camera 4). The first tests are on relatively simple sequences. They are run on a subset of the PETS 2006 and on the PETS 2001, data set 3 sequence. Then the robustness to partial occlusions is shown on a subset of the PETS 2001, data set 1 sequence. Finally we present the handling of missing observations on a subset of the PETS 2006 sequence. For all the results except the first one, the frames have been cropped to show in more details the segmentation.

(a) (b) (c)

Fig. 4. Reference frames. (a) Reference frame for the PETS 2006 sequence. (b) Reference frame for the PETS 2001 sequence, dataset 1. (c) Reference frame for the PETS 2001 sequence, dataset 3.

4.1 Results with observations at each time

First results (figure 5) are on part of the PETS 2006 sequence with no particular changes. Observations are obtained by subtracting current frame with the reference frame (frame 10) shown on figure 4(a). In the first frame of test sequence, frame number 801, two objects are initialized using the observations. The chair on the left of the image is detected and always present in the tracking because a person was sited on it in the reference frame. Tracking this object is not a drawback as it could be an abandoned object. The person walking since the beginning is well tracked until it gets out of the image. A new object is then detected and a new tracker is initialized on it from frame 878. As one can see, even if the background subtraction and associated observations are not perfect, for example if part of the object is missing, our segmentation algorithm recovers the entire object.

Second results are shown in figure 6. Observations are obtained by subtracting current frame with the reference frame (frame 2200) shown on figure 4(c). Two persons are tracked in this sequence in which the light is slowly changing. In addition to this gradual change, the left person moves from light to shade. Still, our algorithm tracks correctly both persons.

(a) (b) (c)

Fig. 5. Results on the PETS 2006 sequence for frames 801, 820, 860, 900 (a) Result of simple background subtraction and extracted observations (bounding boxes) (b) Masks of tracked and segmented objects (c) Tracked objects on current frame

4.2 Results with partial occlusion

Results showing the robustness to partial occlusions are shown in figure 7. Observations are obtained by subtracting current frame with the reference frame (frame 2700) shown on figure 4(b). Three objects are tracked in this sequence. The third one, with green overlay, corresponds to the car shadow and is visible on the last frame shown. Our method allows the tracking of the car as a whole even when it is partially occluded with a lamp post.

4.3 Results with missing observations

Last result (figure 8) illustrates the capacity of the method to handle missing observations thanks to the prediction mechanism. The same part of the PETS 2006 sequence

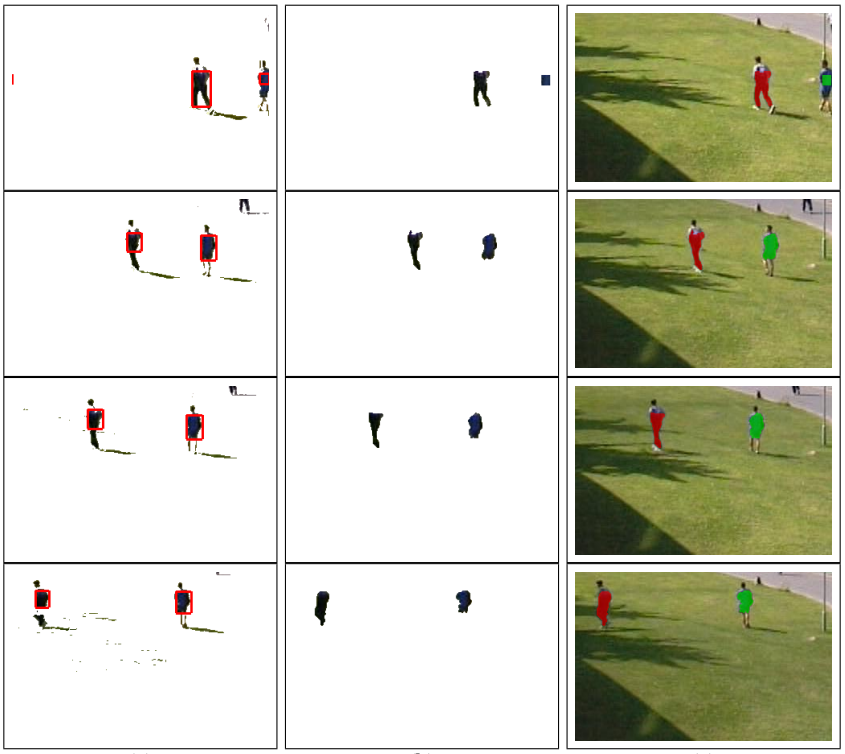
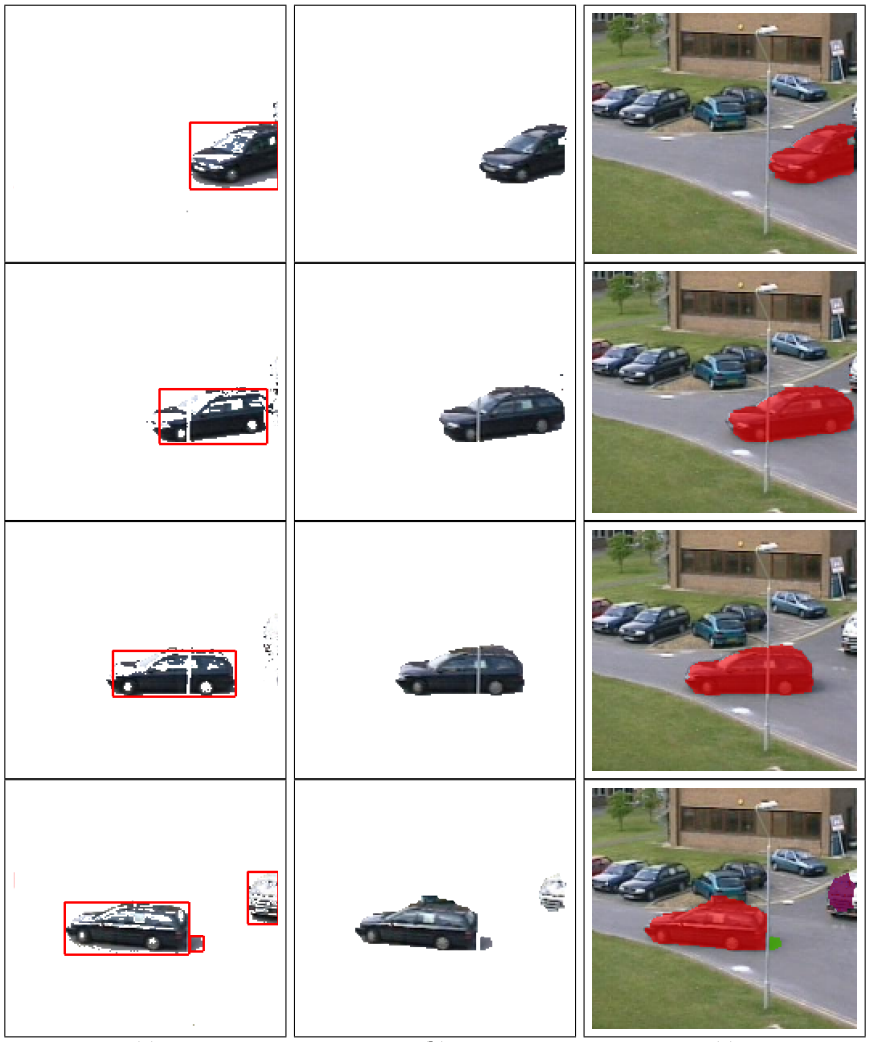


Fig. 6. Results with partial occlusions on the PETS 2001 sequence for frames 2260, 2328, 2358 and 2398 (a) Result of background subtraction and extracted observations (bounding boxes) (b) Masks of tracked and segmented objects (c) Tracked objects on current frame

as in figure 5 is used. In our test we have only performed the background subtraction on one over three frames. On figure 8, we compare the obtained segmentation with the one of figure 5 based on observations at each frame. Thanks to prediction, the result is only partially altered by this drastic temporal subsampling of observations. As one can see, even if one leg is missing in frames 805 and 806, it can be recovered as soon as a new observation is available. Conversely, this result also shows that the incorporation of observations from a detection module enables to get better segmentations than when using only predictions.

5 Conclusion

In this paper we have presented a new method to simultaneously segment and track objects. Predictions and observations composed of detected objects are introduced in an energy function which is minimized using graph cuts. The use of graph cuts permits the segmentation of the objects at a modest computational cost. A novelty is the use



(a) (b) (c)

Fig. 7. Results with partial occlusions on the PETS 2001 sequence for frames 2481, 2496, 2511 and 2526 (a) Result of background subtraction and extracted observations (bounding boxes) (b) Masks of tracked and segmented objects (c) Tracked objects on current frame

of observation nodes in the graph which gives better segmentations but also enables the association of the tracked objects to the observations. The algorithm is robust to partial occlusion, progressive illumination changes and to missing observations. The observations used in this paper are obtained by a very simple background subtraction based on a single reference frame. More complex background subtraction or object detection could be used as well with no change to the approach. As we use distributions of objects at previous time to minimize the energy, our method would fail in case of very abrupt illumination changes. However by adding an external detector of abrupt

illumination changes, we could circumvent this problem by keeping only the prediction and update the reference frame when an abrupt change occurs. We are currently investigating a way to handle complete occlusions. Another research direction lies in handling the fusion and split of several detection masks in more cluttered scenes.

References

1. M. Bertalmio, G. Sapiro, and G. Randall. Morphing active contours. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(7):733–737, 2000.
2. A. Blake, C. Rother, M. Brown, P. Perez, and P.H.S. Torr. Interactive image segmentation using an adaptive gmmrf model. In *Proc. Europ. Conf. Computer Vision*, 2004.
3. Y. Boykov and M.P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. *Proc. Int. Conf. Computer Vision*, 2001.
4. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(11):1222–1239, 2001.
5. D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean-shift. In *Proc. Conf. Comp. Vision Pattern Rec.*, 2000.
6. D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based optical tracking. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(5):564–577, 2003.
7. D. Cremers and C. Schnörr. Statistical shape knowledge in variational motion segmentation. *Image and Vision Computing*, 21(1):77–86, 2003.
8. D. Freedman and M. W. Turek. Illumination-invariant tracking via graph cuts. *Proc. Conf. Comp. Vision Pattern Rec.*, 2005.
9. M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.
10. O. Juan and Y. Boykov. Active graph cuts. In *Proc. Conf. Comp. Vision Pattern Rec.*, 2006.
11. P. Kohli and P.H.S. Torr. Effciently solving dynamic markov random fields using graph cuts. In *Proc. Int. Conf. Computer Vision*, 2005.
12. B.D. Lucas and T. Kanade. An iterative technique of image registration and its application to stereo. *Proc. Int. Joint Conf. on Artificial Intelligence*, 1981.
13. A. Mansouri. Region tracking via level set pdes without motion computation. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(7):947–961, 2002.
14. N. Paragios and R. Deriche. Geodesic active regions for motion estimation and tracking. In *Proc. Int. Conf. Computer Vision*, 1999.
15. R. Ronfard. Region-based strategies for active contour models. *Int. J. Computer Vision*, 13(2):229–251, 1994.
16. D. Terzopoulos and R. Szeliski. Tracking with Kalman snakes. In *Active Vision*, pages 3–20. MIT Press, Cambridge, MA, 1992.
17. Y. Wang, J.F. Doherty, and R.E. Van Dyck. Moving object tracking in video. *Applied Imagery Pattern Recognition Annual Workshop*, 2000.
18. N. Xu and N. Ahuja. Object contour tracking using graph cuts based active contours. *Proc. Int. Conf. Image Processing*, 2002.
19. A. Yilmaz. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Trans. Pattern Anal. Machine Intell.*, 26(11):1531–1536, 2004.
20. A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.

(a) (b) (c)

Fig. 8. Results with observations only every 3 frames on the PETS 2006 sequence for frames 801 to 807 (a) Result of background subtraction and observations (b) Masks of tracked and segmented objects (c) Comparison with the masks obtained when there is no missing observations