TRACK AND CUT: SIMULTANEOUS TRACKING AND SEGMENTATION OF MULTIPLE OBJECTS WITH GRAPH CUTS

Aurélie Bugeau and Patrick Pérez
INRIA, Centre Rennes - Bretagne Atlantique, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France

Keywords: Tracking, Graph Cuts.

Abstract: This paper presents a new method to both track and segment multiple objects in videos using min-cut/max-flow optimizations. We introduce objective functions that combine low-level pixel-wise measures (color, motion), high-level observations obtained via an independent detection module (connected components of foreground detection masks in the experiments), motion prediction and contrast-sensitive contextual regularization. One novelty is that external observations are used without adding any association step. The minimization of these cost functions simultaneously allows "detection-before-track" tracking (track-to-observation assignment and automatic initialization of new tracks) and segmentation of tracked objects. When several tracked objects get mixed up by the detection module (e.g., single foreground detection mask for objects close to each other), a second stage of minimization allows the proper tracking and segmentation of these individual entities despite the observation confusion. Experiments on sequences from PETS 2006 corpus demonstrate the ability of the method to detect, track and precisely segment persons as they enter and traverse the field of view, even in cases of occlusions (partial or total), temporary grouping and frame dropping.

1 INTRODUCTION

Visual tracking is an important and challenging problem. Depending on applicative context under concern, it comes into various forms (automatic or manual initialization, single or multiple objects, still or moving camera, etc.), each of which being associated with an abundant literature. In a recent review on visual tracking (Yilmaz et al., 2006), 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. Such a tracking can be performed with either deterministic or probabilistic methods. Deterministic methods amount to matching by minimizing a distance based on certain descriptors of the object. Probabilistic methods provide means to take measurement uncertainties into account and are often based on a state space model of the object properties.

Dynamic segmentation aims to extract successive segmentations over time. A detailed silhouette of the target object is thus sought in each frame. This is often done by making evolve the silhouette obtained in the previous frame towards a new configuration in current frame. It can be done using a state space model defined in terms of shape and motion parameters of the contour (Isard and Blake, 1998; Terzopoulos and Szeliski, 1993) or by the minimization of a contour energy functional. The contour energy includes temporal information in the form of either temporal gradients (optical flow) (Bertalmio et al., 2000; Cremers and C. Schnörr, 2003; Mansouri, 2002) or appearance statistics originated from the object and its surroundings in previous images (Ronfard, 1994; Yilmaz, 2004). In (Xu and Ahuja, 2002) the authors use graph cuts to minimize such an energy functional. The advantages of min-cut/max-flow optimization are its low computational cost, the fact that it converges to the global minimum without getting stuck in local minima and that no *a priori* on the global shape model is needed.

In the last group of methods ("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 for the

Bugeau A. and Pérez P. (2008).
TRACK AND CUT: SIMULTANEOUS TRACKING AND SEGMENTATION OF MULTIPLE OBJECTS WITH GRAPH CUTS.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 447-454
DOI: 10.5220/0001075704470454

tracked object. The most popular method in this class is the one of Comaniciu *et al.* (Comaniciu et al., 2000; Comaniciu et al., 2003), where approximate "mean shift" iterations are used to conduct the iterative search. Graph cuts have also been used for illumination invariant kernel tracking in (Freedman and Turek, 2005).

These three types of tracking techniques have different advantages and limitations, and can serve different purposes. The "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. However this kind of tracking usually outputs bounding boxes only. By contrast, 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 minima 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 moving objects are approximately known from a preprocessing algorithm. Here, we use a simple background subtraction but more complex alternatives could be applied. An important novelty of our method is that the use of external observations does not require the addition of a preliminary association step. The association between the tracked objects and the observations is jointly conducted with the segmentation and the tracking within the proposed minimization method. The connected components of the detected foreground mask serve as high-level observations. At each time instant, tracked object masks are propagated using their associated optical flow, which provides predictions. Color and motion distributions are computed on the objects segmented in previous frame and used to evaluate individual pixel likelihood in the current frame. We introduce for each object a binary labeling objective function that combines all these ingredients (low-level pixel-wise features, high-level observations obtained via an independent detection module and motion predictions) with a contrast-sensitive contextual regularization. The minimization of each of these energy functions with min-cut/max-flow provides the segmentation of one of the tracked objects in the new frame. Our algorithm also deals with the introduction of new objects and their associated tracker. When multiple objects trigger a single detection due to their spatial vicinity, the proposed method, as most detect-before-track approaches, can get confused. To circumvent this problem, we propose to minimize a secondary multi-label energy function which allows the individual segmentation of concerned objects.

In section 2, notations are introduced and an overview of the method is given. The primary energy function associated to each tracked object is introduced in section 3. The introduction of new objects and the handling of complete occlusions are also explained in this section. The secondary energy function permitting the separation of objects wrongly merged in the first stage is introduced in section 4. Experimental results are reported in section 5, where we demonstrate the ability of the method to detect, track and precisely segment persons and groups, possibly with partial or complete occlusions and missing observations. The experiments also demonstrate that the second stage of minimization allows the segmentation of individual persons when spatial proximity makes them merge at the foreground detection level.

2 PRINCIPLE AND NOTATIONS

In all this paper, $\mathcal{P}$ will denote the set of N pixels of a frame from an input image sequence. 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 vector of optical flow values. Using an incremental multi-scale implementation of Lucas and Kanade algorithm (Lucas and Kanade, 1981), the optical flow is in fact only computed at pixels with sufficiently contrasted surroundings. For the other pixels, color constitutes the only low-level feature. However, for notational convenience, we shall assume in the following that optical flow is available at each pixel.

We assume that, at time t, k_t objects are tracked. The i^{th} object at time t is denoted as $O_t^{(i)}$ and is defined as a mask of pixels, $O_t^{(i)} \subset \mathcal{P}$.

The goal of this paper is to perform both segmentation and tracking to get the object $O_t^{(i)}$ corresponding to the object $O_{t-1}^{(i)}$ of previous frame. Contrary to sequential segmentation techniques (Juan and Boykov, 2006; Kohli and Torr, 2005; Paragios and Deriche, 1999), we bring in object-level "observations". They may be of various kinds (*e.g.*, obtained by a class-specific object detector, or 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 (Wang et al., 2000) 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 mask of pixels, $\mathcal{M}_t^{(j)} \subset \mathcal{P}$. Each observation is characterized by its mean feature vector:

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

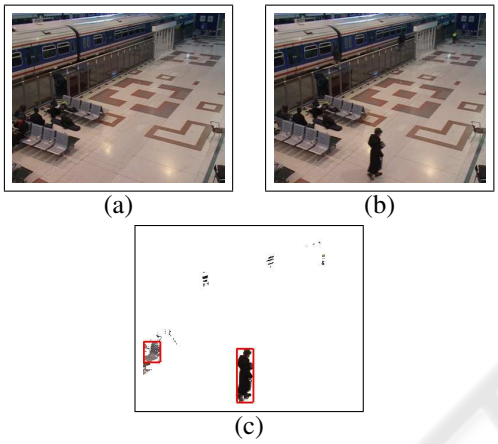


Figure 1: Observations obtained with background subtraction. (a) Reference frame. (b) Current frame. (c) Result of background subtraction (pixels in black are labeled as foreground) and derived object detections (indicated with red bounding boxes).

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

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

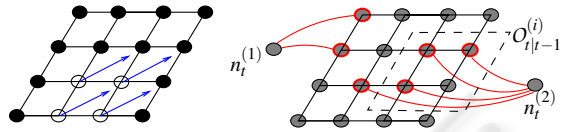
The prediction is obtained by translating each pixels belonging to $O_{t-1}^{(i)}$ by this average optical flow:

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

Using this prediction, the new observations, as well as color and motion distributions of $O_{t-1}^{(i)}$, an energy function is built. The energy is minimized using min-cut/max-flow algorithm (Boykov et al., 2001), which gives the new segmented object at time t, $O_t^{(i)}$. The minimization also provides the correspondences of the object with all the available observations.

3 ENERGY FUNCTIONS

We define one tracker for each object. To each tracker corresponds, for each frame, one graph and one energy function that is minimized using the min-cut/max-flow algorithm (Boykov et al., 2001). Nodes and edges of the graph can be seen in figure 2.



Object i at time t-1 — Graph for object i at time t

Figure 2: 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 shown in blue. The right figure shows the graph at time t. Two observations are available, each of which giving rise to a special "observation" node. The pixel nodes circled in red correspond to the masks of these two observations. Dashed box indicates predicted mask.

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 composed of two subsets. The first subset is the set of N pixels of the image grid $\mathcal{P}$. The second subset corresponds to the observations: to each observation mask $\mathcal{M}_t^{(j)}$ is associated a node $n_t^{(j)}$. We call these nodes "observation nodes". 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 $\mathcal{P}$, and $\mathcal{E}_{\mathcal{M}_t^{(j)}}$ is the set of unordered pairs $\{s, n_t^{(j)}\}$, with $s \in \mathcal{M}_t^{(j)}$.

Segmenting the object $O_t^{(i)}$ 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 $l_{j,t}^{(i)}$ ("bg" or "fg") to each observation node $n_t^{(j)}$. The set of all the node labels forms $L_t^{(i)}$.

3.2 Energy

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

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

3.2.1 Data Term

The data term only concerns the pixel nodes lying in the predicted regions and the observation nodes. For all the other pixel nodes, labeling will only be controlled by the neighbors via binary terms. More precisely, the first part of energy in (4) reads:

$$\sum_{s\in\mathcal{V}_t} R^{(i)}_{s,t}(l^{(i)}_{s,t}) = \sum_{s\in O^{(i)}_{t|t-1}} -\ln(p^{(i)}_1(s,l^{(i)}_{s,t}))+ \sum_{j=1}^{m_t} -\ln(p^{(i)}_2(j,l_{j,t})) \ . \tag{5}$$

Segmented object at time t should be similar, in terms of motion and color, to the preceding instance of this object at times $t-1$. To exploit this consistency assumption, color and motion distributions of the object and the background are extracted from previous image. The distribution $p^{(i,C)}_{t-1}$ for color, respectively $p^{(i,M)}_{t-1}$ for motion, is a Gaussian mixture model fitted to the set of values $\{\mathbf{z}^{(C)}_{s,t-1}\}_{s\in O^{(i)}_{t-1}}$, respectively $\{\mathbf{z}^{(M)}_{s,t-1}\}_{s\in O^{(i)}_{t-1}}$. Under independency assumption for color and motion, the likelihood of individual pixel feature $\mathbf{z}_{s,t}$ according to previous joint model is:

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

The two distributions for the background are $q^{(i,M)}_{t-1}$ and $q^{(i,C)}_{t-1}$. They are Gaussian mixture models built on the sets $\{\mathbf{z}^{(M)}_{s,t-1}\}_{s\in\mathcal{P}\setminus O^{(i)}_{t-1}}$ and $\{\mathbf{z}^{(C)}_{s,t-1}\}_{s\in\mathcal{P}\setminus O^{(i)}_{t-1}}$ respectively. Foreground likelihood at pixel s then reads:

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

The likelihood p_1, invoked in (5) within predicted region, can now be defined as:

$$p^{(i)}_1(s,l) = \begin{cases} p^{(i)}_{t-1}(\mathbf{z}_{s,t}) \text{ if } l = \text{“fg”}, \\ q^{(i)}_{t-1}(\mathbf{z}_{s,t}) \text{ if } l = \text{“bg”} \ . \end{cases} \tag{8}$$

An observation should be used only if it is likely to correspond to the tracked object. Therefore, we use a similar definition for p_2. However we do not evaluate the likelihood of each pixel of the observation mask but only the one of its mean feature $\bar{\mathbf{z}}^{(j)}_t$. The likelihood p_2 for the observation node $n^{(j)}_t$ is defined as:

$$p^{(i)}_2(j,l) = \begin{cases} p^{(i)}_{t-1}(\bar{\mathbf{z}}^{(j)}_t) \text{ if } l = \text{“fg”}, \\ q^{(i)}_{t-1}(\bar{\mathbf{z}}^{(j)}_t) \text{ if } l = \text{“bg”} \ . \end{cases} \tag{9}$$

3.2.2 Binary Term

Following (Boykov and Jolly, 2001), the binary term between neighboring pairs of pixels $\{s,r\}$ of $\mathcal{P}$ is based on color gradients and has the form

$$B^{(i)}_{s,r,t} = \lambda_1 \frac{1}{\text{dist}(s,r)} e^{-\frac{\|\mathbf{z}^{(C)}_{s,t}-\mathbf{z}^{(C)}_{r,t}\|^2}{\sigma_T^2}} \ . \tag{10}$$

As in (Blake et al., 2004), the parameter σ_T is set to $\sigma_T = 4\cdot\langle(\mathbf{z}^{(i,C)}_{s,t}-\mathbf{z}^{(i,C)}_{r,t})^2\rangle$, where $\langle.\rangle$ denotes expectation over a box surrounding the object. For edges between one pixel node and one observation node, the binary term is similar:

$$B^{(i)}_{s,n^{(j)}_t,t} = \lambda_2\, e^{-\frac{\|\mathbf{z}^{(C)}_{s,t}-\bar{\mathbf{z}}^{(j,C)}_t\|^2}{\sigma_T^2}} \ . \tag{11}$$

Parameters λ_1 and λ_2 are discussed in the experiments.

3.2.3 Energy Minimization

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

$$\hat{L}^{(i)}_t = \arg\min_{L^{(i)}_t} E^{(i)}_t(L^{(i)}_t). \tag{12}$$

This labeling gives the segmentation of the i-th object at time t as:

$$O^{(i)}_t = \{s\in\mathcal{P} : \hat{l}^{(i)}_{s,t} = \text{“fg”}\}. \tag{13}$$

3.3 Handling Complete Occlusions

When the number of pixels belonging to a tracked object becomes equal to zero, this object is likely to have disappeared due to either its exit of the field of view or its complete occlusion. If it is occluded, we want to recover it as soon as it reappears. Let t_o be the time at which the size drops to zero, and $S^{(i)}_t$ be the size of object i at time t. The simplest way to handle occlusions is to keep predicting the object using information available just before its complete disappearance:

$$O^{(i)}_{t|t-1} = \{s+(t-t_o+1)\mathbf{d}^{(i)}_{t_o-1}, s\in O^{(i)}_{t_o-1}\} \ \ ,t>t_o \ \ , \tag{14}$$

and minimizing the energy function with

$$p^{(i)}_{t-1} \equiv p^{(i)}_{t_o-1},\ q^{(i)}_{t-1} \equiv q^{(i)}_{t_o-1}. \tag{15}$$

However, before being completely occluded, an object is usually partially occluded, which influences its shape, its motion and the feature distributions. Therefore, using only information at time t_0-1 is not

sufficient and a more complex scheme must be applied. To this end, we try to find the instant t_p at which the object started to be occluded. A Gaussian distribution $\mathcal{N}(\overline{S}^{(i)}, \sigma_S^{(i)})$ on the size of the object is built and updated at each instant. If $|\mathcal{N}(S_t^{(i)}; \overline{S}^{(i)}, \sigma_S^{(i)}) - \overline{S}^{(i)}| < 3\sigma_S^{(i)}$, then we consider that the object is partially occluded and $t_p = t-1$. The prediction and the distributions are finally built on averages over the 5 frames before t_p:

$$O_{t|t-1}^{(i)} = \{s + \frac{t-t_p+1}{5} \sum_{t'=t_p-5}^{t_p} \mathbf{d}_{t'}^{(i)}, s \in O_{t_p}^{(i)}\}, \quad (16)$$

while the distributions $p_{t-1}^{(i)}$ and $q_{t-1}^{(i)}$ are now Gaussian mixture models fitted on sets $\{\mathbf{z}_{s,t'}\}_{t'=t_p-5 \ldots t_p \ , \ s \in O_{t'}^{(i)}}$ and $\{\mathbf{z}_{s,t'}\}_{t'=t_p-5 \ldots t_p \ , \ s \in \mathcal{P} \setminus O_{t'}^{(i)}}$ respectively. Specific motion models depending on the application could have been used but this falls beyond the scope of the paper.

3.4 Creation of New Objects

One advantage of our approach lies in its ability to jointly manipulate pixel labels and track-to-detection assignment labels. This allows the system to track and segment the objects at time t while establishing the correspondence between an object currently tracked and all the approximative object candidates obtained by detection in current frame. If, after the energy minimization for an object i, an observation node $n_t^{(j)}$ is labeled as "fg" it means that there is a correspondence between the i-th object and the j-th observation. If for all the objects, an observation node is labeled as "bg" ($\forall i, \ l_{t,j}^{(i)} =$ "bg"), then the corresponding observation does not match any object. In this case, a new object is created and initialized with this observation.

4 SEGMENTING MERGED OBJECTS

Assume now that the results of the segmentations for different objects overlap, that is $\cap_{i \in \mathcal{F}} O_t^{(i)} \neq \emptyset$, where $\mathcal{F}$ denotes the current set of object indices. In this case, we propose an additional step to determine whether these objects truly correspond to the same one or if they should be separated. At the end of this step, each pixel of $\cap_{i \in \mathcal{F}} O_t^{(i)}$ must belong to only one object. For this purpose, a new graph $\tilde{G}_t = (\tilde{\mathcal{V}}_t, \tilde{\mathcal{E}}_t)$ is created, where $\tilde{\mathcal{V}}_t = \cup_{i \in \mathcal{F}} O_t^{(i)}$ and $\tilde{\mathcal{E}}_t$ is composed of all unordered pairs of neighboring pixel nodes of $\tilde{\mathcal{V}}_t$. The goal is then to assign to each node s of $\tilde{\mathcal{V}}_t$ a label $\psi_s \in \mathcal{F}$. Defining $\tilde{\mathcal{L}} = \{\psi_s, s \in \tilde{\mathcal{V}}_t\}$ the labeling of $\tilde{\mathcal{V}}_t$, a new energy is defined as:

$$\tilde{E}_t(\tilde{\mathcal{L}}) = \sum_{s \in \tilde{\mathcal{V}}_t} -\ln(p_3(s, \psi_s)) + \lambda_3 \sum_{\{s,r\} \in \tilde{\mathcal{E}}_t} \frac{1}{\text{dist}(s,r)} e^{-\frac{\|\mathbf{z}_s^{(C)} - \mathbf{z}_r^{(C)}\|^2}{\sigma_3^2}} (1 - \delta(\psi_s, \psi_r)). \quad (17)$$

The parameter σ_3 is here set as $\sigma_3 = 4 \cdot \langle (\mathbf{z}_{s,t}^{(i,C)} - \mathbf{z}_{r,t}^{(i,C)})^2 \rangle$ with the averaging being over $i \in \mathcal{F}$ and $\{s,r\} \in \tilde{\mathcal{E}}$. The fact that several objects have been merged shows that their respective feature distributions at previous instant did not permit to distinguish them. A way to separate them is then to increase the role of the prediction. This is achieved by choosing function p_3 as:

$$p_3(s, \psi) = \begin{cases} p_{t-1}^{(\psi)}(\mathbf{z}_{s,t}) \text{ if } s \notin O_{t|t-1}^{(\psi)}, \\ 1 \qquad \text{otherwise} \ . \end{cases} \quad (18)$$

This multi-label energy function is minimized using the α-expansion and the swap algorithms (Boykov et al., 1998; Boykov et al., 2001). After this minimization, the objects $O_t^{(i)}, i \in \mathcal{F}$ are updated.

5 EXPERIMENTAL RESULTS

In this section we present various results on a sequence from the PETS 2006 data corpus (sequence 1 camera 4). The robustness to partial occlusions and the individual segmentation of objects that were initially merged, are first demonstrated. Then we present the handling of missing observations and of complete occlusions on other parts of the video. Following (Blake et al., 2004), the parameter λ_3 was set to 20. However parameters λ_1 and λ_2 had to be tuned by hand to get better results. Indeed λ_1 was set to 10 while λ_2 to 2. Also, the number of classes for the Gaussian mixture models was set to 10.

(a) (b)

Figure 3: Reference frames. (a) Reference frame for subsections 5.1 and 5.2. (b) Reference frame for subsection 5.3.

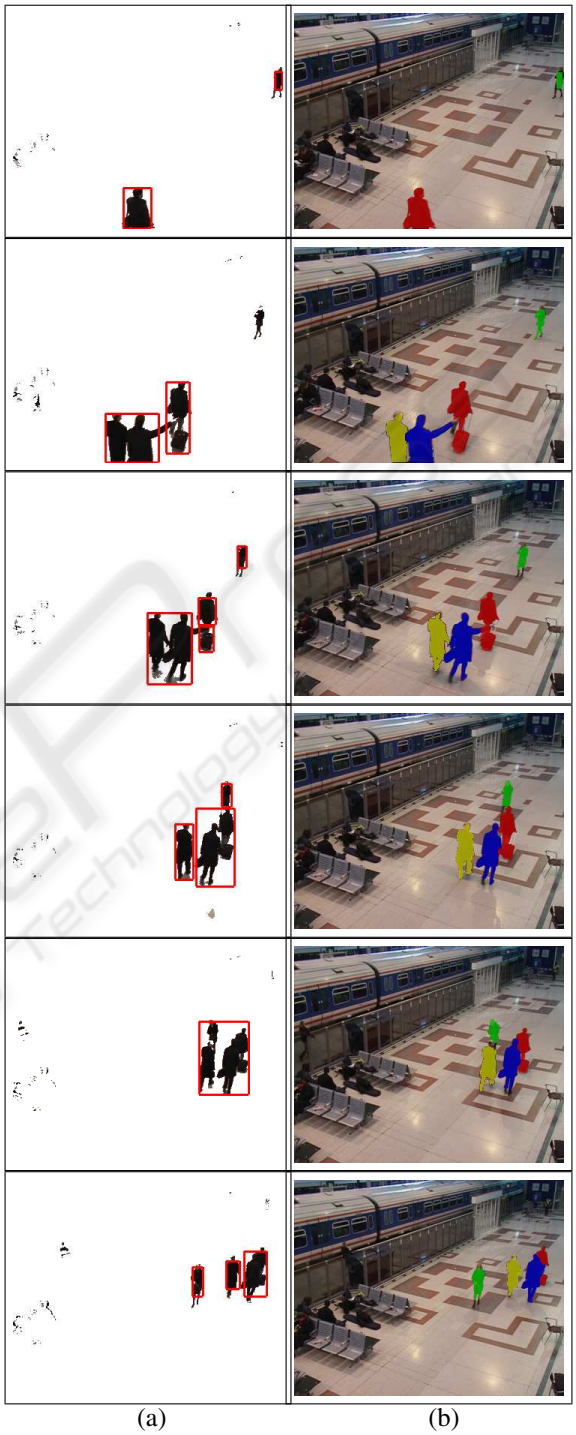
5.1 Observations at Each Time

First results (figure 4) demonstrate the good behavior of our algorithm even in the presence of partial occlusions and of object fusion. Observations, obtained by subtracting reference frame (frame 10 shown on figure 3(a)) to the current one, are visible in the first column of figure 4. The second column contains the segmentation of the objects with the use of the second energy function. Each tracked object is represented by a different color. In frame 81, two objects are initialized using the observations. Note that the connected component extracted with the "gap/mountain" method misses the legs for the person in the upper right corner. While this impacts the initial segmentation, the legs are included in the segmentation as soon as the subsequent frame. Even if from the 102^{nd} frame the two persons at the bottom of the frames correspond to only one observation, our algorithm tracks each person separately (frames 116, 146). Partial occlusions occur when the person at the top passes behind the three other ones (frames 176 and 206), which is well handled by the method, as the person is still tracked when the occlusion stops (frame 248).

In figure 5, we show in more details the influence of the second energy function by comparing the results obtained with and without it. Before frame 102, the three persons at the bottom generate three distinct observations while, passed this instant, they correspond to only one or two observations. Even if the motions and colors of the three persons are very close, the use of the secondary multi-label energy function allows their separation.

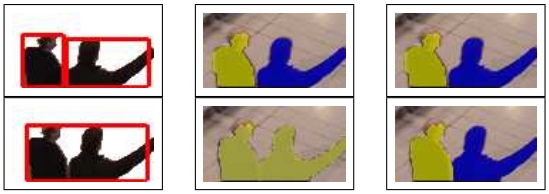
5.2 Missing Observations

On figure 6 we illustrate the capacity of the method to handle missing observations thanks to the prediction mechanism. In our test we have only performed the background subtraction on one over three frames. On figure 6, we compare the obtained segmentations with the ones based on observations at each frame. First column shows the intermittent observations, the second one the masks of the objects obtained in case of missing observations and the last one the masks with observations at each time. Thanks to the prediction, the results are 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 is recovered as soon as a new observation is available. Conversely, this result also shows that the incorporation of observations from the detection module enables to get better segmentations than when using only predictions.



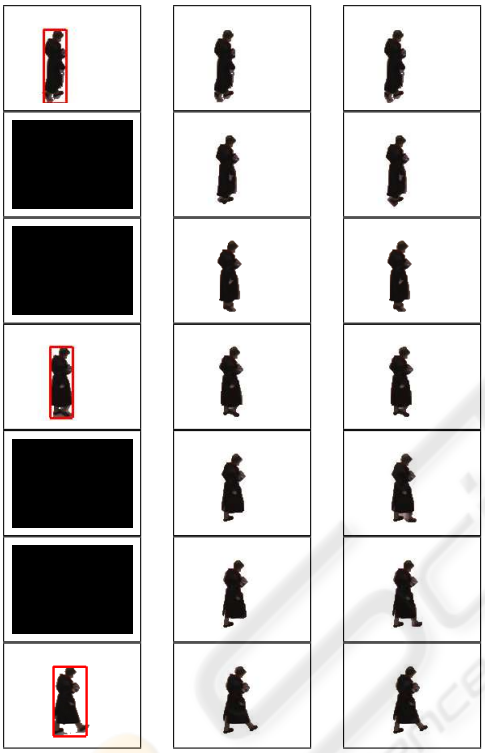
(a) (b)

Figure 4: Results on sequence from PETS 2006 (frames 81, 116, 146, 176, 206 and 248). (a) Result of simple background subtraction and extracted observations. (b) Tracked objects on current frame using the secondary energy function.



(a) (b) (c)

Figure 5: Separating merged objects with the secondary minimization (frames 101 and 102). (a) Result of simple background subtraction and extracted observations. (b) Segmentations with primary energy functions only. (c) Segmentation after post-processing with the secondary energy function.

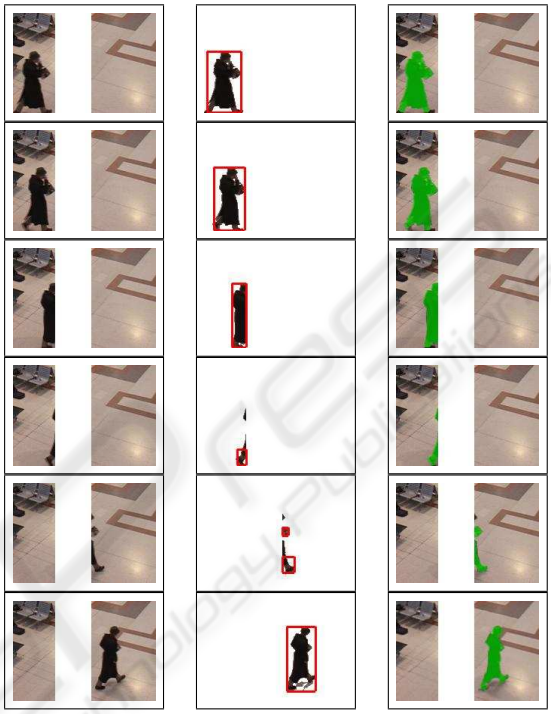


(a) (b) (c)

Figure 6: Results with observations only every 3 frames (frames 801 to 807) (a) Results of background subtraction and extracted observations. (b) Masks of tracked object. (c) Comparison with the masks obtained when there is no missing observations.

5.3 Complete Occlusions

Results in figure 7 demonstrate the ability of our method to deal with complete occlusions. In this portion of the video, we added synthetically a vertical white band in the images in order to generate complete occlusions. The reference frame can be seen on figure 3(b). On figure 7, the first column contains the original images (with the white band), the second one the observations and the last one the obtained segmentations. Our algorithm keeps tracking and segmenting the object as it progressively disappears and resumes tracking and segmenting it as soon as it reappears.



(a) (b) (c)

Figure 7: Results with complete occlusions (frames 782, 785, 792, 798, 810 and 824) (a) Original frames. (b) Results of background subtraction and extracted observations. (c) Comparison with the masks obtained when there is not any missing observations.

6 CONCLUSIONS

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 of observation nodes in the graph which gives better segmentations but also enables the direct association of the tracked objects to the observations (without adding any association procedure). The algorithm is robust to partial and complete occlusions, progressive illumination changes and to missing observations. Thanks to the use of a secondary multi-label energy function, our method allows individual tracking and segmentation

of objects which where not distinguished from each other in the first stage. The observations used in this paper are obtained by a simple background subtraction based on a single reference frame. Note however that more complex background subtraction or object detection could be used as well with no change to the approach.

As we use feature distributions of objects at previous time to define current energy functions, our method breaks down in extreme cases of abrupt illumination changes. However, by adding an external detector of such changes, we could circumvent this problem by keeping only the prediction and updating the reference frame when the abrupt change occurs. Also, other cues, such as shapes, should probably be added to improve the results.

Apart from this rather specific problem, several research directions are open. One of them concerns the design of an unifying energy framework that would allow segmentation and tracking of multiple objects while precluding the incorrect merging of similar objects getting close to each other in the image plane. Another direction of research concerns the automatic tuning of the parameters, which remains an open problem in the recent literature on image labeling (e.g., figure/ground segmentation) with graph-cuts.

REFERENCES

Bertalmio, M., Sapiro, G., and Randall, G. (2000). Morphing active contours. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(7):733–737.

Blake, A., Rother, C., Brown, M., Pérez, P., and Torr, P. (2004). Interactive image segmentation using an adaptive gmmrf model. In *Proc. Europ. Conf. Computer Vision*.

Boykov, Y. and Jolly, M. (2001.). Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. *Proc. Int. Conf. Computer Vision*.

Boykov, Y., Veksler, O., and Zabih, R. (1998). Markov random fields with efficient approximations. In *Proc. Conf. Comp. Vision Pattern Rec.*

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(11):1222–1239.

Comaniciu, D., Ramesh, V., and Meer, P. (2000). Real-time tracking of non-rigid objects using mean-shift. In *Proc. Conf. Comp. Vision Pattern Rec.*

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based optical tracking. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(5):564–577.

Cremers, D. and C. Schnörr, C. (2003). Statistical shape knowledge in variational motion segmentation. *Image and Vision Computing*, 21(1):77–86.

Freedman, D. and Turek, M. (2005). Illumination-invariant tracking via graph cuts. *Proc. Conf. Comp. Vision Pattern Rec.*

Isard, M. and Blake, A. (1998). Condensation – conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28.

Juan, O. and Boykov, Y. (2006). Active graph cuts. In *Proc. Conf. Comp. Vision Pattern Rec.*

Kohli, P. and Torr, P. (2005). Effciently solving dynamic markov random fields using graph cuts. In *Proc. Int. Conf. Computer Vision*.

Lucas, B. and Kanade, T. (1981). An iterative technique of image registration and its application to stereo. *Proc. Int. Joint Conf. on Artificial Intelligence*.

Mansouri, A. (2002). Region tracking via level set pdes without motion computation. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(7):947–961.

Paragios, N. and Deriche, R. (1999). Geodesic active regions for motion estimation and tracking. In *Proc. Int. Conf. Computer Vision*.

Ronfard, R. (1994). Region-based strategies for active contour models. *Int. J. Computer Vision*, 13(2):229–251.

Terzopoulos, D. and Szeliski, R. (1993). Tracking with kalman snakes. *Active vision*, pages 3–20.

Wang, Y., Doherty, J., and Van Dyck, R. (2000). Moving object tracking in video. *Applied Imagery Pattern Recognition (AIPR) Annual Workshop*.

Xu, N. and Ahuja, N. (2002). Object contour tracking using graph cuts based active contours. *Proc. Int. Conf. Image Processing*.

Yilmaz, A. (2004). Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Trans. Pattern Anal. Machine Intell.*, 26(11):1531–1536.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13.