In EMMCVPR'01, Energy Minimization Methods in Computer Vision and Pattern Recognition, UK, June 2001.

Maximum Likelihood Estimation of the Template of a Rigid Moving Object

Pedro M. Q. Aguiar[1] and José M. F. Moura[2]

[1] Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal
E-mail: aguiar@isr.ist.utl.pt, WWW page: www.isr.ist.utl.pt/~aguiar
[2] Electrical and Computer Eng., Carnegie Mellon University, Pittsburgh PA, USA
E-mail: moura@ece.cmu.edu, WWW page: http://www.ece.cmu.edu/~moura

Abstract. Motion segmentation methods often fail to detect the motions of low textured regions. We develop an algorithm for segmentation of low textured moving objects. While usually current motion segmentation methods use only two or three consecutive images our method refines the shape of the moving object by processing successively the new frames as they become available. We formulate the segmentation as a parameter estimation problem. The images in the sequence are modeled taking into account the *rigidity* of the moving object and the *oclusion* of the background by the moving object. The segmentation algorithm is derived as a computationally simple approximation to the *Maximum Likelihood* estimate of the parameters involved in the image sequence model: the motions, the template of the moving object, its intensity levels, and the intensity levels of the background pixels. We describe experiments that demonstrate the good performance of our algorithm.

1 Introduction

The segmentation of an image into regions that undergo different motions has received the attention of a large number of researchers. According to their research focus, different scientific communities addressed the motion segmentation task from distinct viewpoints.

Several papers on image sequence coding address the motion segmentation task with computation time concerns. They reduce temporal redundancy by predicting each frame from the previous one through motion compensation. See reference [15] for a review on very low bit rate video coding. Regions undergoing different movements are compensated in different ways, according to their motion. The techniques used in image sequence coding attempt to segment the moving objects by processing only two consecutive frames. Since their focus is on compression and not in developing a high level representation, these efforts have not considered low textured scenes, and regions with no texture are considered unchanged. As an example, we applied the algorithm of reference [6] to segmenting a low textured moving object. Two consecutive frames of a traffic road video clip are shown in the left side of Figure 1. In the right side of Figure 1, the template of the moving car was found by excluding from the regions

that changed between the two co-registered frames the ones that correspond to uncovered background areas, see reference [6]. The small regions that due to the noise are misclassified as belonging to the car template can be discarded by an adequate morphological post-processing. However, due to the low texture of the car, the regions in the interior of the car are misclassified as belonging to the background, leading to a highly incomplete car template.

Fig. 1. Motion segmentation in low texture.

High level representation in image sequence understanding has been considered in the computer vision literature. Their approach to motion-based segmentation copes with low textured scenes by coupling motion-based segmentation with prior knowledge about the scenes as in statistical regularization techniques, or by combining motion with other atributtes. For example, reference [7] uses a *Markov Random Field* (MRF) prior and a *Bayesian Maximum a Posteriori* (MAP) criterion to segment moving regions. The authors suggest a multiscale MRF modeling to resolve large regions of uniform intensity. In reference [5], the contour of a moving object is estimated by fusing motion with color segmentation and edge detection. In general, these methods lead to complex and time consuming algorithms.

References [8, 9] describe one of the few approaches using temporal integration by averaging the images registered according to the motion of the different objects in the scene. After processing a number of frames, each of these integrated images is expected to show only one sharp region corresponding to the tracked object. This region is found by detecting the stationary regions between the corresponding integrated image and the current frame. Unless the background is textured enough to blur completely the averaged images, some regions of the background can be classified as stationary. In this situation, their method overestimates the template of the moving object. This is particularly likely to happen when the background has large regions with almost constant color or intensity level.

1.1 Proposed Approach

We formulate image sequence analysis as a parameter estimation problem by using the analogy between a communications system and image sequence analysis,

see references [14] and [17]. The segmentation algorithm is derived as a computationally simple approximation to the *Maximum Likelihood* (ML) estimate of the parameters involved in the two-dimensional (2D) image sequence model: the motions, the template of the moving object, its intensity levels (the object texture), and the intensity levels of the background pixels (the background texture). The joint ML estimation of the complete set of parameters is a very complex task. Motivated by our experience with real video sequences, we decouple the estimation of the motions (moving objects and camera) from that of the remaining parameters. The motions are estimated on a frame by frame basis and then used in the estimation of the remaining parameters. Then, we introduce the motion estimates into the ML cost function and minimize this function with respect to the remaining parameters.

The estimate of the object texture is obtained in closed form. To estimate the background texture and the moving object template, we develop a fast two-step iterative algorithm. The first step estimates the background for a fixed template – the solution is obtained in closed form. The second step estimates the template for a fixed background – the solution is given by a simple binary test evaluated at each pixel. The algorithm converges in a few iterations, typically three to five iterations.

Our approach is related to the approach of references [8, 9], however, we model explicitly the *oclusion* of the background by the moving object and we use all the frames available rather than just a single frame to estimate the moving object template. Even when the moving object has a color very similar to the color of the background, our algorithm has the ability to resolve accurately the moving object from the background, because it integrates over time those small differences.

1.2 Paper Organization

In section 2, we state the segmentation problem. We define the notation, develop the observation model, and formulate the ML estimation. In section 3, we detail the two-step iterative method that minimizes the ML cost function. In section 4, we describe two experiments that demonstrate the performance of our algorithm. Section 5 concludes the paper.

For the details not included in this paper, see reference [1]. A preliminary version of this work was presented in reference [2].

2 Problem Formulation

We discuss motion segmentation in the context of *Generative Video* (GV), see references [11–13]. GV is a framework for the analysis and synthesis of video sequences. In GV the operational units are not the individual images in the original sequence, as in standard methods, but rather the world images and the ancillary data. The world images encode the non-redundant information about the video sequence. They are augmented views of the world – background

world image – and complete views of moving objects – figure world images. The ancillary data registers the world images, stratifies them at each time instant, and positions the camera with respect to the layering of world images. The world images and the ancillary data are the GV representation, the information that is needed to regenerate the original video sequence. We formulate the moving object segmentation task as the problem of generating the world images and ancillary data for the GV representation of a video clip.

2.1 Notation

An image is a real function defined on a subset of the real plane. The image space is a set $\{\mathbf{I} : \mathcal{D} \to \mathcal{R}\}$, where $\mathbf{I}$ is an image, $\mathcal{D}$ is the domain of the image, and $\mathcal{R}$ is the range of the image. The domain $\mathcal{D}$ is a compact subset of the real plane $\mathbb{R}^2$, and the range $\mathcal{R}$ is a subset of the real line $\mathbb{R}$. Examples of images are the frame f in the video sequence, denoted by $\mathbf{I}_f$, the background world image, denoted by $\mathbf{B}$, the moving object world image, denoted by $\mathbf{O}$, and the moving object template, denoted by $\mathbf{T}$. The images $\mathbf{I}_f$, $\mathbf{B}$, and $\mathbf{O}$ have range $\mathcal{R} = \mathbb{R}$. They code intensity gray levels[1]. The template of the moving object is a binary image, i.e., an image with range $\mathcal{R} = \{0, 1\}$, defining the region occupied by the moving object. The domain of the images $\mathbf{I}_f$ and $\mathbf{T}$ is a rectangle corresponding to the support of the frames. The domain of the background world image $\mathbf{B}$ is a subset $\mathcal{D}$ of the plane whose shape and size depends on the camera motion, i.e., $\mathcal{D}$ is the region of the background observed in the entire sequence. The domain $\mathcal{D}$ of the moving object world image is the subset of $\mathbb{R}^2$ where the template $\mathbf{T}$ takes the value 1, i.e., $\mathcal{D} = \{(x, y) : \mathbf{T}(x, y) = 1\}$.

In our implementation, the domain of each image is rectangular shaped with size fitting the needs of the corresponding image. Although we use a continuous spatial dependence for commodity, in practice the domains are discretized and the images are stored as matrices. We index the entries of each of these matrices by the pixels (x, y) of each image and refer to the value of image $\mathbf{I}$ at pixel (x, y) as $\mathbf{I}(x, y)$. Throughout the text, we refer to the image product of two images $\mathbf{A}$ and $\mathbf{B}$, i.e., the image whose value at pixel (x, y) equals $\mathbf{A}(x, y)\mathbf{B}(x, y)$, as the image $\mathbf{AB}$. Note that this product corresponds to the Hadamard product, or elementwise product, of the matrices representing images $\mathbf{A}$ and $\mathbf{B}$, not their matrix product.

We consider two-dimensional (2D) parallel motions, i.e., all motions (translations and rotations) are parallel to the camera plane. We represent this kind of

[1] The intensity values of the images in the video sequence are positive. In our experiments, these values are coded by a binary word of eight bits. Thus, the intensity values of a gray level image are in the set of integers in the interval $[0, 255]$. For simplicity, we do not take into account the discretization and the saturations, i.e., we consider the intensity values to be real numbers and the gray level images to have range $\mathcal{R} = \mathbb{R}$. The analysis in the thesis is easily extended to color images. A color is represented by specifying three intensities, either of the perceptual attributes *brightness*, *hue*, and *saturation*; or of the primary colors *red*, *green*, and *blue*, see reference [10]. The range of a color image is then $\mathcal{R} = \mathbb{R}^3$.

motions by specifying time varying position vectors. These vectors code rotation-translation pairs that take values in the group of rigid transformations of the plane, the special Euclidean group SE(2). The image obtained by applying the rigid motion coded by the vector $\mathbf{p}$ to the image $\mathbf{I}$ is denoted by $\mathcal{M}(\mathbf{p})\mathbf{I}$. The image $\mathcal{M}(\mathbf{p})\mathbf{I}$ is also usually called the registration of the image $\mathbf{I}$ according to the position vector $\mathbf{p}$. The entity represented by $\mathcal{M}(\mathbf{p})$ is seen as a motion operator. In practice, the (x, y) entry of the matrix representing the image $\mathcal{M}(\mathbf{p})\mathbf{I}$ is given by $\mathcal{M}(\mathbf{p})\mathbf{I}(x, y) = \mathbf{I}(f_x(\mathbf{p}; x, y), f_y(\mathbf{p}; x, y))$ where $f_x(\mathbf{p}; x, y)$ and $f_y(\mathbf{p}; x, y)$ represent the coordinate transformation imposed by the 2D rigid motion. We use bilinear interpolation to compute the intensity values at points that fall in between the stored samples of an image.

The motion operators can be composed. The registration of the image $\mathcal{M}(\mathbf{p})\mathbf{I}$ according to the position vector $\mathbf{q}$ is denoted by $\mathcal{M}(\mathbf{qp})\mathbf{I}$. By doing this we are using the notation $\mathbf{qp}$ for the composition of the two elements of SE(2), $\mathbf{q}$ and $\mathbf{p}$. We denote the inverse of $\mathbf{p}$ by $\mathbf{p}^{\#}$, i.e., the vector $\mathbf{p}^{\#}$ is such that when composed with $\mathbf{p}$ we obtain the identity element of SE(2). Thus, the registration of the image $\mathcal{M}(\mathbf{p})\mathbf{I}$ according to the position vector $\mathbf{p}^{\#}$ obtains the original image $\mathbf{I}$, so we have $\mathcal{M}(\mathbf{p}^{\#}\mathbf{p})\mathbf{I} = \mathcal{M}(\mathbf{pp}^{\#})\mathbf{I} = \mathbf{I}$. Note that, in general, the elements of SE(2) do not commute, i.e., we have $\mathbf{qp} \neq \mathbf{pq}$, and $\mathcal{M}(\mathbf{qp})\mathbf{I} \neq \mathcal{M}(\mathbf{pq})\mathbf{I}$. Only in special cases is the composition of the motion operators not affected by the order of application, as for example when the motions $\mathbf{p}$ and $\mathbf{q}$ are pure translations or pure rotations.

The notation for the position vectors involved in the segmentation problem is as follows. The vector $\mathbf{p}_f$ represents the position of the background world image relative to the camera in frame f. The vector $\mathbf{q}_f$ represents the position of the moving object relative to the camera in frame f.

2.2 Observation Model

The observation model considers a scene with a moving object in front of a moving camera with two-dimensional (2D) parallel motions. The pixel (x, y) of the image $\mathbf{I}_f$ belongs either to the background world image $\mathbf{B}$ or to the object world image $\mathbf{O}$. The intensity $\mathbf{I}_f(x, y)$ of the pixel (x, y) is modeled as

$$\begin{aligned}\mathbf{I}_f(x, y) = \mathcal{M}(\mathbf{p}_f^{\#})\mathbf{B}(x, y)\ \Big[1 - \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T}(x, y)\Big] \\ +\mathcal{M}(\mathbf{q}_f^{\#})\mathbf{O}(x, y)\ \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T}(x, y) + \mathbf{W}_f(x, y). \qquad (1)\end{aligned}$$

In equation (1), $\mathbf{T}$ is the moving object template, $\mathbf{p}_f$ and $\mathbf{q}_f$ are the camera pose and the object position, and $\mathbf{W}_f$ stands for the observation noise, assumed Gaussian, zero mean, and white.

Equation (1) states that the intensity of the pixel (x, y) on frame f, $\mathbf{I}_f(x, y)$, is a noisy version of the true value of the intensity level of the pixel (x, y). If the pixel (x, y) of the current image belongs to the template of the object, $\mathbf{T}$, after the template is compensated by the object position, i.e., registered according to the vector $\mathbf{q}_f^{\#}$, then $\mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T}(x, y) = 1$. In this case, the first term of the right

hand side of (1) is zero, while the second term equals $\mathcal{M}(\mathbf{q}_f^{\#})\mathbf{O}(x,y)$, the intensity of the pixel (x,y) of the moving object. In other words, the intensity $\mathbf{I}_f(x,y)$ equals the object intensity $\mathcal{M}(\mathbf{q}_f^{\#})\mathbf{O}(x,y)$ corrupted by the noise $\mathbf{W}_f(x,y)$. On the other hand, if the pixel (x,y) does not belong to the template of the object, $\mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T}(x,y) = 0$, and this pixel belongs to the background world image $\mathbf{B}$, registered according to the inverse $\mathbf{p}_f^{\#}$ of the camera position. In this case, the intensity $\mathbf{I}_f(x,y)$ is a noisy version of the background intensity $\mathcal{M}(\mathbf{p}_f^{\#})\mathbf{B}(x,y)$. We want to emphasize that rather than modeling simply the two different motions, as usually done when processing only two consecutive frames, expression (1) models the *oclusion* of the background by the moving object explicitly.

Expression (1) is rewritten in compact form as

$$\mathbf{I}_f = \left\{ \mathcal{M}(\mathbf{p}_f^{\#})\mathbf{B} \left[\mathbf{1} - \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T} \right] + \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{O}\, \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T} + \mathbf{W}_f \right\} \mathbf{H}, \qquad (2)$$

where we assume that $\mathbf{I}_f(x,y) = 0$ for (x,y) outside the region observed by the camera. This is taken care of in equation (2) by the binary image $\mathbf{H}$ whose (x,y) entry is such that $\mathbf{H}(x,y) = 1$ if pixel (x,y) is in the observed images $\mathbf{I}_f$ or $\mathbf{H}(x,y) = 0$ if otherwise. The image $\mathbf{1}$ is constant with value 1.

2.3 Maximum Likelihood Estimation

Given F frames $\{\mathbf{I}_f, 1 \le f \le F\}$, we want to estimate the background world image $\mathbf{B}$, the object world image $\mathbf{O}$, the object template $\mathbf{T}$, the camera poses $\{\mathbf{p}_f, 1 \le f \le F\}$, and the object positions $\{\mathbf{q}_f, 1 \le f \le F\}$. The quantities $\{\mathbf{B}, \mathbf{O}, \mathbf{T}, \{\mathbf{p}_f\}, \{\mathbf{q}_f\}\}$ define the GV representation, the information that is needed to regenerate the original video sequence.

Using the observation model of expression (2) and the Gaussian white noise assumption, ML estimation leads to the minimization over all GV parameters of the functional[2]

$$C_2 = \iint \sum_{f=1}^{F} \Big\{ \mathbf{I}_f(x,y) - \mathcal{M}(\mathbf{p}_f^{\#})\mathbf{B}(x,y) \left[1 - \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T}(x,y) \right] - \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{O}(x,y)\, \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T}(x,y) \Big\}^2 \mathbf{H}(x,y)\, dx\, dy, \qquad (3)$$

where the inner sum is over the full set of F frames and the outer integral is over all pixels.

The estimation of the parameters of expression (2) using the F frames rather than a single pair of images is a distinguishing feature of our work. Other techniques usually process only two or three consecutive frames. We use all frames available as needed. The estimation of the parameters through the minimization

[2] We use a continuous spatial dependence for commodity. The variables x and y are continuous while f is discrete. In practice, the integral is approximated by the sum over all the pixels.

of a cost function that involves directly the image intensity values is another distinguishing feature of our approach. Other methods try to make some type of post-processing over incomplete template estimates. We process directly the image intensity values, through ML estimation.

The minimization of the functional C_2 in equation (3) with respect to the set of GV constructs $\{\mathbf{B}, \mathbf{O}, \mathbf{T}\}$ and to the motions $\{\{\mathbf{p}_f\}, \{\mathbf{q}_f\}, 1 \leq f \leq F\}$ is a highly complex task. To obtain a computationally feasible algorithm, we simplify the problem. We decouple the estimation of the motions $\{\{\mathbf{p}_f\}, \{\mathbf{q}_f\}, 1 \leq f \leq F\}$ from the determination of the GV constructs $\{\mathbf{B}, \mathbf{O}, \mathbf{T}\}$. This is reasonable from a practical point of view and is well supported by our experimental results with real videos.

The rationale behind the simplification is that the motion of the object (and the motion of the background) can be inferred without having the knowledge of the exact object template. When only two or three frames are given, even humans find it much easier to infer the motions present in the scene than to recover an accurate template of the moving object. To better appreciate the complexity of the problem, the reader can imagine an image sequence for which there is not prior knowledge available, except that there is a background and an ocluding object that moves differently from the background. Since there are no spatial cues, consider, for example, that the background texture and the object texture are spatial white noise random variables. In this situation, humans can easily infer the motion of the background and the motion of the object, even from only two consecutive frames. With respect to the template of the moving object, we are able to infer much more accurate templates if we are given a higher number of frames because in this case we easily capture the *rigidity* of the object across time. This observation motivated our approach of decoupling the estimation of the motions from the estimation of the remaining parameters.

We perform the estimation of the motions on a frame by frame basis by using a known motion estimation method [4], see reference [1] for the details. After estimating the motions, we introduce the motion estimates into the ML cost function and minimize with respect to the remaining parameters. The solution provided by our algorithm is sub-optimal, in the sense that it is an approximation to the ML estimate of the entire set of parameters, and it can be seen as an initial guess for the minimizer of the ML cost function given by expression (3). Then, we can refine the estimate by using a greedy approach. We must emphasize, however, that the key problem here is to find the initial guess in an expedite way, not the final refinement.

3 Minimization Procedure

In this section, we assume that the motions have been correctly estimated and are known. We should note that, in reality, the motions are continuously estimated. Assuming the motions are known, the problem becomes the minimization of the ML cost function with respect to the remaining parameters, i.e., with respect

to the template of the moving object, the texture of the moving object, and the texture of the background.

3.1 Two-Step Iterative Algorithm

Due to the special structure of the ML cost function C_2, we can express explicitly and with no approximations involved the estimate $\widehat{\mathbf{O}}$ of the object world image in terms of the template $\mathbf{T}$. Doing this, we are left with the minimization of C_2 with respect to the template $\mathbf{T}$ and the background world image $\mathbf{B}$, still a non-linear minimization. We approximate this minimization by a two-step iterative algorithm: (i) in step one, we solve for the background $\mathbf{B}$ while the template $\mathbf{T}$ is kept fixed; and (ii) in step two, we solve for the template $\mathbf{T}$ while the background $\mathbf{B}$ is kept fixed. We obtain closed-form solutions for the minimizers in each of the steps (i) and (ii). The two steps are repeated iteratively. The value of the ML cost function C_2 decreases along the iterative process. The algorithm proceeds till every pixel has been assigned unambiguously to either the moving object or to the background.

To initialize the segmentation algorithm, we need an initial estimate of the background. A simple, often used, estimate for the background is the average of the images in the sequence, including or not a robust statistic technique like outlier rejection, see for example reference [16]. The quality of this background estimate depends on the oclusion level of the background in the images processed. Depending on the particular characteristics of the image sequence, our algorithm can recover successfully the template of the moving object when using the average of the images as the initial estimate of the background. This is the case with the image sequence we use in the experiments reported in section 4. In reference [1], we propose a more elaborate initialization that leads to better initial estimates of the background.

3.2 Estimation of the moving object world image

We express the estimate $\widehat{\mathbf{O}}$ of the moving object world image in terms of the object template $\mathbf{T}$. By minimizing C_2 with respect to the intensity value $\mathbf{O}(x, y)$, we obtain the average of the pixels that correspond to the point (x, y) of the object. The estimate $\widehat{\mathbf{O}}$ of the moving object world image is then

$$\widehat{\mathbf{O}} = \mathbf{T}\,\frac{1}{F}\sum_{f=1}^{F}\mathcal{M}(\mathbf{q}_f)\mathbf{I}_f. \tag{4}$$

This compact expression averages the observations $\mathbf{I}$ registered according to the motion $\mathbf{q}_f$ of the object in the region corresponding to the template $\mathbf{T}$ of the moving object.

We consider now separately the two steps of the iterative algorithm described above.

3.3 Step (i): estimation of the background for fixed template

To find the estimate $\widehat{\mathbf{B}}$ of the background world image, given the template $\mathbf{T}$, we register each term of the sum of the ML cost function C_2 in equation (3) according to the position of the camera $\mathbf{p}_f$ relative to the background. This is a valid operation because C_2 is defined as a sum over all the space $\{(x,y)\}$. We get

$$C_2 = \iint \sum_{f=1}^{F} \Big\{ \mathcal{M}(\mathbf{p}_f)\mathbf{I}_f - \mathbf{B}\ \Big[1 - \mathcal{M}(\mathbf{p}_f\mathbf{q}_f^{\#})\mathbf{T}\Big] \\ -\mathcal{M}(\mathbf{p}_f\mathbf{q}_f^{\#})\mathbf{O}\,\mathcal{M}(\mathbf{p}_f\mathbf{q}_f^{\#})\mathbf{T}(x,y)\Big\}^2 \mathcal{M}(\mathbf{p}_f)\mathbf{H}\,dx\,dy. \tag{5}$$

Minimizing the ML cost function C_2 given by expression (5) with respect to the intensity value $\mathbf{B}(x,y)$, we get the estimate $\widehat{\mathbf{B}}(x,y)$ as the average of the observed pixels that correspond to the pixel (x,y) of the background. The background world image estimate $\widehat{\mathbf{B}}$ is then written as

$$\widehat{\mathbf{B}} = \frac{\sum_{f=1}^{F}\Big[\mathbf{1} - \mathcal{M}(\mathbf{p}_f\mathbf{q}_f^{\#})\mathbf{T}\Big]\ \mathcal{M}(\mathbf{p}_f)\mathbf{I}_f}{\sum_{i=f}^{F}\Big[\mathbf{1} - \mathcal{M}(\mathbf{p}_f\mathbf{q}_f^{\#})\mathbf{T}\Big]\ \mathcal{M}(\mathbf{p}_f)\mathbf{H}}. \tag{6}$$

The estimate $\widehat{\mathbf{B}}$ of the background world image in expression (6) is the average of the observations $\mathbf{I}_f$ registered according to the background motion $\mathbf{p}_i$, in the regions $\{(x,y)\}$ not ocluded by the moving object, i.e., when $\mathcal{M}(\mathbf{p}_f\mathbf{q}_f^{\#})\mathbf{T}(x,y) = 0$. The term $\mathcal{M}(\mathbf{p}_f)\mathbf{H}$ provides the correct averaging normalization in the denominator by accounting only for the pixels seen in the corresponding image.

If we compare the moving object world image estimate $\widehat{\mathbf{O}}$ given by equation (4) with the background world image estimate $\widehat{\mathbf{B}}$ in equation (6), we see that $\widehat{\mathbf{O}}$ is linear in the template $\mathbf{T}$, while $\widehat{\mathbf{B}}$ is nonlinear in $\mathbf{T}$. This has implications when estimating the template $\mathbf{T}$ of the moving object, as we see next.

3.4 Step (ii): estimation of the template for fixed background

Let the background world image $\mathbf{B}$ be given and replace the object world image estimate $\widehat{\mathbf{O}}$ given by expression (4) in expression (3). The ML cost function C_2 becomes linearly related to the object template $\mathbf{T}$. Manipulating C_2 as described next, we obtain

$$C_2 = \iint \mathbf{T}(x,y)\ \mathbf{Q}(x,y)\ dx\,dy + \text{Constant}, \tag{7}$$

$$\mathbf{Q}(x,y) = \mathbf{Q}_1(x,y) - \mathbf{Q}_2(x,y), \tag{8}$$

$$\mathbf{Q}_1(x,y) = \frac{1}{F}\sum_{f=2}^{F}\sum_{g=1}^{f-1}\left[\mathcal{M}(\mathbf{q}_f)\mathbf{I}_f(x,y) - \mathcal{M}(\mathbf{q}_g)\mathbf{I}_g(x,y)\right]^2, \tag{9}$$

$$\mathbf{Q}_2(x,y) = \sum_{f=1}^{F} \Big[\mathcal{M}(\mathbf{q}_f)\mathbf{I}_f(x,y) - \mathcal{M}(\mathbf{q}_f\mathbf{p}_f^{\#})\mathbf{B}(x,y)\Big]^2. \tag{10}$$

We call $\mathbf{Q}$ the *segmentation matrix*.

Derivation of expressions (7) to (10)

Replace the estimate $\widehat{\mathbf{O}}$ of the moving object world image, given by expression (4), in expression (3), to obtain

$$C_2 = \iint \sum_{f=1}^{F} \Big\{ \mathbf{I} - \mathcal{M}(\mathbf{p}_f^{\#})\mathbf{B} \ \Big[1 - \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T}\Big] - \frac{1}{F}\sum_{g=1}^{F} \mathcal{M}(\mathbf{q}_f^{\#}\mathbf{q}_g)\mathbf{I}_g \, \mathcal{M}(\mathbf{q}_f^{\#})\mathbf{T} \Big\}^2 \mathbf{H}\, dx\, dy. \tag{11}$$

Register each term of the sum according to the object position $\mathbf{q}_f$. This is valid because C_2 is defined as an integral over all the space $\{(x,y)\}$. The result is

$$C_2 = \iint \sum_{f=1}^{F} \Big\{ \Big[\mathcal{M}(\mathbf{q}_f)\mathbf{I}_f - \mathcal{M}(\mathbf{q}_f\mathbf{p}_f^{\#})\mathbf{B}\Big] + \left[\mathcal{M}(\mathbf{q}_f\mathbf{p}_f^{\#})\mathbf{B} - \frac{1}{F}\sum_{g=1}^{F}\mathcal{M}(\mathbf{q}_g)\mathbf{I}_g\right]\mathbf{T} \Big\}^2 \mathcal{M}(\mathbf{q}_f)\mathbf{H}\, dx\, dy. \tag{12}$$

In the remainder of the derivation, the spatial dependence is not important here, and we simplify the notation by omitting (x,y). We rewrite the expression for C_2 in compact form as

$$C_2 = \iint \mathbf{C}\, dx\, dy, \quad \mathbf{C} = \sum_{f=1}^{F} \left\{ \Big[\mathcal{I}_f - \mathcal{B}_f\Big] + \left[\mathcal{B}_f - \frac{1}{F}\sum_{g=1}^{F}\mathcal{I}_g\right]\mathbf{T} \right\}^2 \mathcal{H}_f, \tag{13}$$

$$\mathcal{I}_f = \mathcal{M}(\mathbf{q}_f)\mathbf{I}_f(x,y), \quad \mathcal{B}_f = \mathcal{M}(\mathbf{q}_f\mathbf{p}_f^{\#})\mathbf{B}(x,y), \quad \mathcal{H}_f = \mathcal{M}(\mathbf{q}_f)\mathbf{H}(x,y). \tag{14}$$

We need in the sequel the following equalities

$$\left[\sum_{g=1}^{F}\mathcal{I}_g\right]^2 = \sum_{f=1}^{F}\sum_{g=1}^{F}\mathcal{I}_f\mathcal{I}_g \quad \text{and} \quad \sum_{f=2}^{F}\sum_{g=1}^{f-1}\left[\mathcal{I}_i^2 + \mathcal{I}_g^2\right] = (F-1)\sum_{g=1}^{F}\mathcal{I}_g^2. \tag{15}$$

Manipulating $\mathbf{C}$ under the assumption that the moving object is completely visible in the F images ($\mathbf{T}\mathcal{H}_f = \mathbf{T}, \forall_f$), and using the left equality in (15), we obtain

$$\mathbf{C} = \mathbf{T}\left\{\sum_{f=1}^{F}\left[2\mathcal{I}_f\mathcal{B}_f - \mathcal{B}_f^2\right] - \frac{1}{F}\left[\sum_{g=1}^{F}\mathcal{I}_g\right]^2\right\} + \sum_{f=1}^{F}\Big[\mathcal{I}_f - \mathcal{B}_f\Big]^2\mathcal{H}_f. \tag{16}$$

The second term of $\mathbf{C}$ in expression (16) is independent of the template $\mathbf{T}$. To show that the sum that multiplies $\mathbf{T}$ is the segmentation matrix $\mathbf{Q}$ as defined by expressions (8), (9), and (10), write $\mathbf{Q}$ using the notation introduced in (14):

$$\mathbf{Q} = \frac{1}{F}\sum_{f=2}^{F}\sum_{g=1}^{f-1}\left[\mathcal{I}_f^2 + \mathcal{I}_g^2 - 2\mathcal{I}_f\mathcal{I}_g\right] - \sum_{f=1}^{F}\left[\mathcal{I}_f^2 + \mathcal{B}_f^2 - 2\mathcal{I}_f\mathcal{B}_f\right]. \tag{17}$$

Manipulating this equation, using the two equalities in (15), we obtain

$$\mathbf{Q} = \sum_{f=1}^{F}\left[2\mathcal{I}_f\mathcal{B}_f - \mathcal{B}_f^2\right] - \frac{1}{F}\left[\sum_{g=1}^{F}\mathcal{I}_g^2 + 2\sum_{f=2}^{F}\sum_{g=1}^{f-1}\mathcal{I}_f\mathcal{I}_g\right]. \tag{18}$$

The following equality concludes the derivation:

$$\left[\sum_{g=1}^{F}\mathcal{I}_g\right]^2 = \sum_{g=1}^{F}\mathcal{I}_g^2 + 2\sum_{f=2}^{F}\sum_{g=1}^{f-1}\mathcal{I}_f\mathcal{I}_g. \tag{19}$$

□

We estimate the template $\mathbf{T}$ by minimizing the ML cost function given by expression (7) over the template $\mathbf{T}$, given the background world image $\mathbf{B}$. It is clear from expression (7), that the minimization of C_2 with respect to each spatial location of $\mathbf{T}$ is independent from the minimization over the other locations. The template $\widehat{\mathbf{T}}$ that minimizes the ML cost function C_2 is given by the following test evaluated at each pixel:

$$\mathbf{Q}_1(x,y) \underset{\widehat{\mathbf{T}}(x,y)=1}{\overset{\widehat{\mathbf{T}}(x,y)=0}{\gtrless}} \mathbf{Q}_2(x,y). \tag{20}$$

The estimate $\widehat{\mathbf{T}}$ of the template of the moving object in equation (20) is obtained by checking which of two accumulated square differences is greater. In the spatial locations where the accumulated differences between each frame $\mathcal{M}(\mathbf{q}_f)\mathbf{I}_f$ and the background $\mathcal{M}(\mathbf{q}_g\mathbf{p}_g^{\#})\mathbf{B}$ are greater than the accumulated differences between each pair of co-registered frames $\mathcal{M}(\mathbf{q}_f)\mathbf{I}_f$ and $\mathcal{M}(\mathbf{q}_g)\mathbf{I}_g$, we estimate $\widehat{\mathbf{T}}(x,y) = 1$, meaning that these pixels belong to the moving object. If not, the pixel is assigned to the background.

The reason why we did not replace the background world image estimate $\widehat{\mathbf{B}}$ given by (6) in (3) as we did with the object world image estimate $\widehat{\mathbf{O}}$ is that it leads to an expression for C_2 in which the minimization with respect to each different spatial location $\mathbf{T}(x,y)$ is not independent from the other locations. Solving this binary minimization problem by a conventional method is extremely time consuming. In contrast, the minimization of C_2 over $\mathbf{T}$ for fixed $\mathbf{B}$ results in a local binary test. This makes our solution computationally very simple.

It may happen that, after processing the F available frames, the test (20) remains inconclusive at a given pixel (x,y) ($\mathbf{Q}_1(x,y) \simeq \mathbf{Q}_2(x,y)$): in other words,

it is not possible to decide if this pixel belongs to the moving object or to the background. We modify our algorithm to address this ambiguity by defining the modified cost function

$$C_{2\mathrm{MOD}} = C_2 + \alpha \,\mathrm{Area}(\mathbf{T}) = C_2 + \alpha \iint \mathbf{T}(x,y)\, dx\, dy, \tag{21}$$

where C_2 is as in equation (3), α is non-negative, and Area($\mathbf{T}$) is the area of the template. Minimizing $C_{2\mathrm{MOD}}$ balances the agreement between the observations and the model (term C_2), with minimizing the area of the template. Carrying out the minimization, first note that the second term in expression (21) does not depend on $\mathbf{O}$, neither on $\mathbf{B}$, so we get $\widehat{\mathbf{O}}_{\mathrm{MOD}} = \widehat{\mathbf{O}}$ and $\widehat{\mathbf{B}}_{\mathrm{MOD}} = \widehat{\mathbf{B}}$. By replacing $\widehat{\mathbf{O}}$ in $C_{2\mathrm{MOD}}$, we get a modified version of equation (7),

$$C_{2\mathrm{MOD}} = \iint \mathbf{T}(x,y)\,[\mathbf{Q}(x,y) + \alpha]\; dx\, dy + \mathrm{Constant}, \tag{22}$$

where $\mathbf{Q}$ is defined in equations (8), (9), and (10). The template estimate is now given by the following test, that extends test (20),

$$\mathbf{Q}(x,y) \underset{\widehat{\mathbf{T}}(x,y)=1}{\overset{\widehat{\mathbf{T}}(x,y)=0}{\gtrless}} -\alpha \ . \tag{23}$$

The parameter α may be chosen by experimentation, by using the *Minimum Description Length* (MDL) principle, see reference [3], or made adaptive by a annealing schedule like in stochastic relaxation.

4 Experiments

We describe two experiments. The first one uses a challenging computer generated image sequence to illustrate the convergence of the two-step iterative algorithm and its capability to segment complex shaped moving objects. The second experiment segments a real life traffic video clip.

4.1 Synthetic Image Sequence

We synthesized an image sequence according to the model described in section 2. Figure 2 shows the world images used. The left frame, from a real video, is the background world image. The moving object template is the logo of the *Instituto Superior Técnico* (IST) which is transparent between the letters. Its world image, shown in the right frame, is obtained by clipping with the IST logo a portion of one of the frames in the sequence. The task of reconstructing the object template is particularly challenging with this video sequence due to the low contrast between the object and the background and the complexity of the template. We synthesized a sequence of 20 images where the background is static and the IST logo moves arround.

Fig. 2. GV constructs: background and moving object.

Figure 3 shows three frames of the sequence obtained according to the image formation model introduced in section 2, expression (2), with noise variance $\sigma^2 = 4$ (the intensity values are in the interval $[0, 255]$). The object moves from the center (left frame) down by translational and rotational motion. It is difficult to recognize the logo in the right frame because its texture is confused with the texture of the background.

Fig. 3. Three frames of the synthesized image sequence.

Figure 4 illustrates the four iterations it took for the two-step estimation method of our algorithm to converge. The template estimate is initialized to zero (top left frame). Each background estimate in the right hand side was obtained using the template estimate on the left of it. Each template estimate was obtained using the previous background estimate. The arrows in Figure 4 indicate the flow of the algorithm. The good template estimate obtained, see bottom left image, illustrates that our algorithm can estimate complex templates in low contrast background.

Note that this type of complex templates (objects with transparent regions) is much easier to describe by using a binary matrix than by using contour based descriptions, like splines, Fourier descriptors, or snakes. Our algorithm overcomes the difficulty arising from the higher number of degrees of freedom of the binary template by integrating over time the small intensity differences between the background and the object. The two-step iterative algorithm performs this integrations in an expedite way.

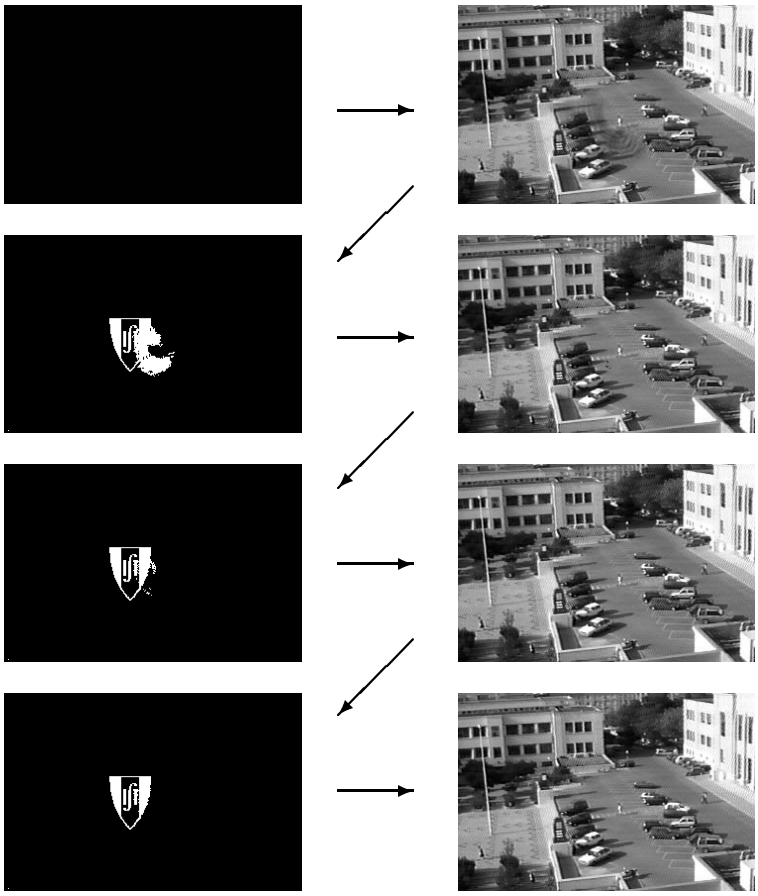

Fig. 4. Two-step iterative method: template estimates and background estimates.

4.2 Road Traffic

In this experiment we use a road traffic video clip. The road traffic video sequence has 250 frames. Figure 5 shows frames 15, 166, and 225. The example given in section 1 to motivate the study of the segmentation of low textured scenes, see Figure 1, also uses frames 76 and 77 from the road traffic video clip.

In this video sequence, the camera exhibits a pronounced panning motion, while four different cars enter and leave the scene. The cars and the background have regions of low texture. The intensity of some of the cars is very similar to the intensity of parts of the background.

Figures 6 and 7 show the good results obtained after segmenting the sequence with our algorithm. Figure 7 displays the background world image, while Figure 6 shows the world images of each of the moving cars. The estimates of the templates

Fig. 5. Traffic road video sequence. Frames 15, 166, and 225.

for the cars in Figure 6 becomes unambiguous after 10, 10, and 14 frames, respectively.

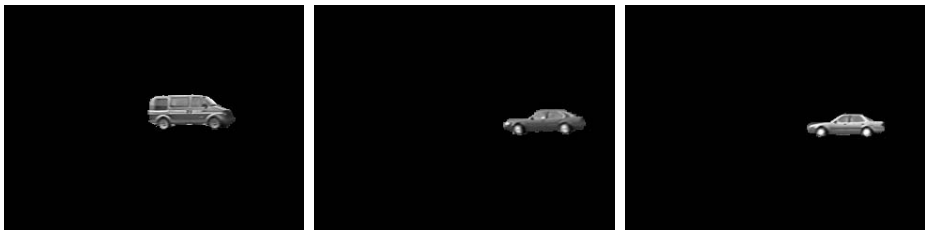

Fig. 6. Moving objects recovered from the traffic road video sequence.

Fig. 7. Background world image recovered from the traffic road video sequence.

5 Conclusion

We develop an algorithm for segmenting 2D rigid moving objects from an image sequence. Our method recovers the template of the 2D rigid moving object by processing directly the image intensity values. We model both the *rigidity* of the moving object over a set of frames and and the *oclusion* of the background by the moving object.

We motivate our algorithm by looking for a feasible approximation to the ML estimation of the unknowns involved in the segmentation problem. Our methodology introduces the 2D motion estimates into the ML cost function and uses a two-step iterative algorithm to approximate the minimization of the resultant cost function. The solutions for both steps result computationally very simple. The two-step algorithm is computationally efficient because the convergence is achieved in a small number of iterations (typically three to five iterations).

Our experiments show that the algorithm proposed can estimate complex templates in low contrast scenes.

References

1. P. M. Q. Aguiar. *Rigid Structure from Video.* PhD thesis, Instituto Superior Técnico, Lisboa, Portugal, 2000. Available at `www.isr.ist.utl.pt/~aguiar`.
2. Pedro M. Q. Aguiar and José M. F. Moura. Detecting and solving template ambiguities in motion segmentation. In *ICIP'97*, Santa Barbara, CA, USA, 1997.
3. A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. on Information Theory*, 44(6), 1998.
4. J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *ECCV'92*, Santa Margherita Ligure, Italy, 1992.
5. Patrick Bouthemy and Edouard François. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *IJCV*, 10(2), 1993.
6. Norbert Diehl. Object-oriented motion estimation and segmentation in image sequences. *Signal Processing: Image Communication*, 3(1):23–56, February 1991.
7. Marie-Pierre Dubuisson and Anil K. Jain. Contour extraction of moving objects in complex outdoor scenes. *IJCV*, 14(1), 1995.
8. Michal Irani and Shmuel Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communications and Image Representation*, 4(4):324–335, December 1993.
9. Michal Irani, Benny Rousso, and Shmuel Peleg. Computing occluding and transparent motions. *IJCV*, 12(1), February 1994.
10. Anil K. Jain. *Fundamentals of Digital Image Processing.* Prentice Hall Information and Sciences Series. Prentice-Hall International Inc., 1989.
11. R. S. Jasinschi and J. M. F. Moura. *Generative Video: Very Low Bit Rate Video Compression.* U.S. Patent and Trademark Office, S.N. 5,854,856, issued, 1998.
12. Radu S. Jasinschi. *Generative Video: A Meta Video Representation.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, September 1995.
13. Radu S. Jasinschi and José M. F. Moura. Content-based video sequence representation. In *ICIP'95*, Washigton D.C., USA, September 1995.
14. D. C. Knill, D.l Kersten, and A. Yuille. A Bayesian formulation of visual perception. In *Perception as Bayesian Inference.* Cambridge University Press, 1996.
15. Haibo Li, Astrid Lundmark, and Robert Forchheimer. Image sequence coding at very low bitrates: A review. *IEEE Trans. on Image Processing*, 3(5), 1994.
16. Wolfgang Luth. Segmentation of image sequences. *Theoretical Foundations of Computer Vision, Mathematical Research – Akademie Verlag*, pages 215–226, 1992.
17. Joseph A. O'Sullivan, Richard E. Blahut, and Donald L. Snyder. Information-theoretic image formation. *IEEE Trans. on Information Theory*, 44(6), 1998.