Segmentation of Global Motion using Temporal Probabilistic Classification

P.R. Giaccone and G.A. Jones
School of Computer Science and Electronic Systems
Kingston University
Kingston upon Thames
Surrey KT1 2EE, UK
[k941319|g.jones]@kingston.ac.uk

Abstract

The segmentation of pixels belonging to different moving *elements* within a cinematographic image sequence underpins a range of post-production special effects. In this work, the separation of foreground elements, such as actors, from arbitrary backgrounds rather than from a *blue screen* is accomplished by accurately estimating the visual motion induced by a moving camera. The optical-flow field of the background is recovered using a parametric motion model (motivated by the three-dimensional pan-and-zoom motion of a camera) embedded in a *spatiotemporal* least-squares minimisation framework. A maximum *a posteriori* probability (MAP) approach is used to assign pixel membership (*background, uncovered, covered* and *foreground*) defined relative to the background element. The standard approach, based on class-conditional *a priori* distributions of *displaced-frame differences*, is augmented by information capturing the expected temporal transitions of pixel labels.

1 Introduction

Visual motion promises to provide a powerful cue to image segmentation. Yet, despite considerable research effort, the generation and segmentation of motion fields remains a largely unsolved problem. The work presented below is motivated by the need to separate moving foreground elements, such as actors, from rigid backgrounds in cinematographic imagery. Even in these relatively constrained circumstances, a number of important issues arise, including the choice of an appropriate motion estimation procedure; the definition of the motion model; and the method of classifying pixels as belonging to the background or to foreground elements.

The commonest approaches to generating dense optical-flow fields use the *constant brightness* constraint but employ different sources of additional information to constrain fully the 2D visual motion at each pixel. These include *regularisation* using *local smoothness* constraints [3] and neighbourhood-based *parametric motion models* [1]. In common with many other examples of the parametric type, the proposed algorithm uses motion models [2, 4, 5, 10] but differs in both the formulation of the motion estimation problem

BMVC 1998 doi:10.5244/C.12.62

and the choice of motion model. In section 2, a global motion model is derived that describes in uncalibrated pixel coordinates the motion of a rigid distant background. The model is embedded in a *three-frame* least-squares estimator that not only employs the enormous amount of constraint available using three consecutive images, but also automatically generates two registered *displaced-frame difference* images (our backward and forward error maps) measuring the greylevel difference between the current frame and the motion-compensated previous and next images respectively.

Having computed the global motion of the image, each pixel may then be classified into one of four classes: *background* pixels that belong to the global motion; *uncovered* background pixels that were occluded by the foreground in the previous frame; *covered* pixels that were previously projections of the background and are now occluded by the foreground; and *foreground* pixels that belong to some object occluding the background. Applications of such a classification process include frame interpolation [13] in video-coding [12] and, in our case, the creation of special effects in the post-production industry [8]. Typically, pixels are classified using *change detection* performed on the previous, current and next frames [11, 13], *ie*, if the greylevel of a 2D projected world event does not change significantly between between frames then the pixel is assumed to belong to the motion field. Pixels are thus classified as follows:

background: No significant greylevel change between the previous and current frames or between the current and next. The pixel belongs to the background in both intervals.

covered: No significant change between the previous and current frames means that the pixel initially belongs to the background. Significant change between the current and next indicates occlusion, *ie, covering* of the background.

uncovered: Significant change between the previous and current frames, indicating previous occlusion, while subsequent lack of change between the current and next frames indicates that the pixel now belongs to the background motion field.

foreground: Greylevels change significantly both between the previous and current frames and between the current and next frames.

Significant changes are recovered by pixel differencing with two important modifications. First, frames are motion-compensated prior to differencing to ensure that equivalent world events are compared. Second, because of noise, *changed* and *unchanged* greylevel differences belong to two overlapping distributions requiring probabilistic interpretation. In section 3, a *maximum a posteriori* estimation framework is introduced to classify pixels based on the forward and backward error maps. Since noise can have a significant effect on this per-frame classification approach, a temporal dimension is introduced into the process by including knowledge about the previous classification of a pixel into the calculation of *a posteriori* probabilities.

2 Global motion estimation

Many commonly used motion models, such as the 2-parameter *uniform translation*, the 4-parameter *2D rotation and translation* and the 6-parameter *affine*, are not motivated by three-dimensional concerns but are nonetheless applied to moving three-dimensional scenes because of their computational simplicity. The accuracy of the motion field is however extremely important when segmenting motion fields: misalignment leads to poor localisation of object boundaries. Consequently, we have derived a linear motion model capable of modelling the visual motion of the rigid background of an image sequence. This depth-independent parametric model, defined in equation 5, assumes that the camera motion is composed only of small rotational, translational and zoom velocities. Translational components should ideally be zero, as these are responsible for parallax distortions. Moreover, no camera calibration is required (assuming no significant spherical aberration).

2.1 Global motion model

The global motion model is formulated by examining the three-dimensional motion of a point in the view volume. The point $\mathbf{X} = (X, Y, Z)^T$ (measured in a camera coordinate system where the Z-axis is aligned along the optical axis) projects onto the 2D pixel position $\mathbf{x} = (x, y)^T$ under a perspective transformation, given the current focal length f. If $\mathbf{x}_0 = (x_0, y_0)^T$ is the intersection point of the optical axis with the image plane and α and β are the pixel dimensions, then $\mathbf{x}$ is related to the 3D position as follows:

$$\frac{X}{Z} = \frac{\alpha}{f}(x - x_0), \qquad \frac{Y}{Z} = \frac{\beta}{f}(y - y_0). \tag{1}$$

Under a small rotational and translational motion over a small time interval Δt, the point $\mathbf{X}$ in the view volume moves to the new position $\mathbf{X}' = (X', Y', Z')^T$:

$$\mathbf{X}' = \mathbf{X} + \Delta t \left\{ \Omega \times \mathbf{X} + \mathbf{T} \right\}, \tag{2}$$

where $\Omega = (\omega_x, \omega_y, \omega_z)^T$ and $\mathbf{T} = (T_x, T_y, T_z)^T$ are the rotational and translational velocities respectively. This generates the following expressions for each component in terms of the original position $\mathbf{X}$ and the motion parameters:

$$\begin{aligned} X' &= X + \Delta t \left\{ (Z\omega_y - Y\omega_z) + T_x \right\}, \\ Y' &= Y + \Delta t \left\{ (X\omega_z - Z\omega_x) + T_y \right\}, \\ Z' &= Z + \Delta t \left\{ (Y\omega_x - X\omega_y) + T_z \right\}. \end{aligned} \tag{3}$$

Under the perspective transformation, the 2D pixel location of the point $\mathbf{X}'$ is

$$x' = x_0 + \frac{f + \phi\Delta t}{\alpha}\frac{X'}{Z'}, \qquad y' = y_0 + \frac{f + \phi\Delta t}{\beta}\frac{Y'}{Z'}, \tag{4}$$

where ϕ is the zoom velocity.

We may assume, since the object distance is much greater than the translation, that

$$\frac{T_x}{Z} \approx 0, \qquad \frac{T_y}{Z} \approx 0, \qquad \frac{T_z}{Z} \approx 0.$$

Moreover, we may further assume that the depth of a point is much greater than the change in depth between frames, *ie*, that $Z \gg (Y\omega_x - X\omega_y) + T_z$ in equation 3. The Taylor-series approximation $(1 + \Delta Z/Z)^{-1} \approx 1 - \Delta Z/Z$ can thus be used to rewrite equation 4 in the following linear form:

$$\begin{aligned} x' &\approx x_0 + \frac{f}{\alpha}\left(1 + \frac{\phi\Delta t}{f}\right)\left[\frac{X}{Z} + \Delta t\left(\omega_y - \omega_z\frac{Y}{Z}\right)\right]\left[1 - \Delta t\left(\frac{Y}{Z}\omega_x - \frac{X}{Z}\omega_y\right)\right], \\ y' &\approx y_0 + \frac{f}{\beta}\left(1 + \frac{\phi\Delta t}{f}\right)\left[\frac{Y}{Z} + \Delta t\left(\frac{X}{Z}\omega_z - \omega_x\right)\right]\left[1 - \Delta t\left(\frac{Y}{Z}\omega_x - \frac{X}{Z}\omega_y\right)\right]. \end{aligned}$$

The above can be simplified further using the fact that since ω_x, ω_y, ω_z and ϕ are small, their products are negligible. By combining this assumption with the perspective expressions of equation 1, the above equation may be rearranged to create an 8-parameter linear visual displacement model in terms of the uncalibrated pixel coordinates x and y:

$$\begin{aligned} \Delta x(x,y) &= x' - x &&= \Delta t\left(a_0x^2 + a_1xy + a_2x + a_3y + a_4\right) \\ \Delta y(x,y) &= y' - y &&= \Delta t\left(a_0xy + a_1y^2 + a_5x + a_6y + a_7\right), \end{aligned} \tag{5}$$

which may be expressed in the alternative vector matrix formulation

$$\Delta\mathbf{x}(\mathbf{x},\mathbf{a},\Delta t) = \begin{bmatrix} \Delta x(x,y) \\ \Delta y(x,y) \end{bmatrix} = \Delta t X(\mathbf{x})\mathbf{a}, \tag{6}$$

where

$$X(\mathbf{x}) = \begin{bmatrix} x^2 & xy & x & y & 1 & 0 & 0 & 0 \\ xy & y^2 & 0 & 0 & 0 & x & y & 1 \end{bmatrix}, \qquad \mathbf{a} = (a_0\ a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ a_7)^T.$$

2.2 Generating an optical-flow estimator

The classical optical-flow approach generates *motion fields* between successive images. A more generalised *spatiotemporal* approach was introduced in [9] that computes the *displacement field* $\Delta\mathbf{x}$ defining the change in position of a pixel, *ie*, the position $\mathbf{x}'$ of a pixel at time $t + \Delta t$ is given by $\mathbf{x} + \Delta\mathbf{x}(\mathbf{x}, \mathbf{a}_t, \Delta t)$, which from equation 6 gives

$$\mathbf{x}' = \mathbf{x} + \Delta t X(\mathbf{x})\mathbf{a}_t, \tag{7}$$

where $\mathbf{a}_t$ are the motion model parameters at time t. Assuming constant intensity over short sequences of images, the estimated displacement field should warp a pixel of a particular greylevel to one of the same greylevel in another image. The accuracy of this motion can be measured by an *error term* $e(\mathbf{x})$ that compares the greylevel $I_t(\mathbf{x})$ at a point $\mathbf{x}$ in the image at time t with the greylevel of the image at time $t + \Delta t$ at the displaced pixel location, defined as

$$e(\mathbf{x}) \stackrel{\text{def}}{=} I_t(\mathbf{x}) - I_{t+\Delta t}(\mathbf{x} + \Delta t X(\mathbf{x})\mathbf{a}_t). \tag{8}$$

An iterative estimator uses a previous estimate of the motion to refine final motion parameters. To generate this estimator, the error function of equation 8 is linearised around the current i^{th} motion estimate $\mathbf{a}_{t,i}$:

$$e(\mathbf{x}) \approx \Delta I_t(\mathbf{x}, \mathbf{a}_{t,i}, \Delta t) - \Delta t \nabla I_{t+\Delta t}(\mathbf{x} + \Delta t X(\mathbf{x})\mathbf{a}_{t,i})^T X(\mathbf{x})\Delta\mathbf{a}_t, \tag{9}$$

where $\Delta\mathbf{a}_t = \mathbf{a}_{t,i+1} - \mathbf{a}_{t,i}$ is the update to the current motion parameters, and $I_{t+\Delta t}(\mathbf{x} + \Delta t X(\mathbf{x})\mathbf{a}_{t,i})$ is the *motion-compensated* frame. The *displaced-frame difference* $\Delta I_t(\mathbf{x}, \mathbf{a}_{t,i}, \Delta t)$ is given by

$$\Delta I_t(\mathbf{x}, \mathbf{a}_{t,i}, \Delta t) = I_t(\mathbf{x}) - I_{t+\Delta t}\left(\mathbf{x} + \Delta\mathbf{x}(\mathbf{x}, \mathbf{a}_{t,i}, \Delta t)\right), \tag{10}$$

where $\Delta\mathbf{x}(\mathbf{x}, \mathbf{a}_{t,i}, \Delta t)$, given by equation 6, is the displacement generated by $\mathbf{a}_{t,i}$.

Any errors in the *motion-compensated* frame are compounded when computing its spatial derivatives and hence we use the spatial derivatives $\nabla I_t(\mathbf{x})$ as a good approximation to $\nabla I_{t+\Delta t}(\mathbf{x} + \Delta t X(\mathbf{x})\mathbf{a}_{t,i})$. The error term may then be rewritten to relate the next estimate of the motion parameters $\mathbf{a}_{t,i+1}$ to the current estimate of the displacement field $\Delta\mathbf{x}(\mathbf{x}, \mathbf{a}_{t,i}, \Delta t)$:

$$e(\mathbf{x}) \approx \Delta I_t(\mathbf{x}, \mathbf{a}_{t,i}, \Delta t) - \nabla I_t(\mathbf{x})^T \left\{\Delta t X(\mathbf{x})\mathbf{a}_{t,i+1} - \Delta\mathbf{x}(\mathbf{x}, \mathbf{a}_{t,i}, \Delta t)\right\}. \tag{11}$$

Such a formulation, which relates motion parameters to the displacement field, enables accurate motion estimation even in the presence of large velocities, given an accurate initial estimate of the displacement field $\Delta\mathbf{x}_{t,0}(\mathbf{x})$, generated using an alternative technique, *eg, block matching*.

2.3 Deriving the three-frame optical-flow estimator

Considerable constraint on the global motion of pixels in a frame is available by assuming that motion is constant over the two-frame time interval centred on the current frame I_t. We can construct two error maps, the *forward* error map $e_t^f(\mathbf{x})$ between frames $I_t(\mathbf{x})$ and $I_{t+1}(\mathbf{x})$; and the *backward* error map $e_t^b(\mathbf{x})$ between frames $I_t(\mathbf{x})$ and $I_{t-1}(\mathbf{x})$, derived from equation 11 using $\Delta t = 1$ and $\Delta t = -1$ respectively as follows:

$$\begin{aligned} e^f(\mathbf{x}, \mathbf{a}_{t,i+1}) &= \left\{I_t(\mathbf{x}) - I_{t+1}\left(\mathbf{x} + \Delta\mathbf{x}_t^{t+1}(\mathbf{x}, \mathbf{a}_{t,i})\right)\right\}, \\ &\qquad -\nabla I_t(\mathbf{x})^T \left\{X(\mathbf{x})\mathbf{a}_{t,i+1} - \Delta\mathbf{x}_t^{t+1}(\mathbf{x}, \mathbf{a}_{t,i})\right\} \\ e^b(\mathbf{x}, \mathbf{a}_{t,i+1}) &= \left\{I_t(\mathbf{x}) - I_{t-1}\left(\mathbf{x} + \Delta\mathbf{x}_t^{t-1}(\mathbf{x}, \mathbf{a}_{t,i})\right)\right\} \\ &\qquad +\nabla I_t(\mathbf{x})^T \left\{X(\mathbf{x})\mathbf{a}_{t,i+1} + \Delta\mathbf{x}_t^{t-1}(\mathbf{x}, \mathbf{a}_{t,i})\right\}, \end{aligned} \tag{12}$$

where $\Delta\mathbf{x}_t^{t+1}(\mathbf{x}, \mathbf{a}_{t,i}) = X(\mathbf{x})\mathbf{a}_{t,i}$ and $\Delta\mathbf{x}_t^{t-1}(\mathbf{x}, \mathbf{a}_{t,i}) = -X(\mathbf{x})\mathbf{a}_{t,i}$.

A least-squares problem may be formulated to locate the appropriate motion parameters that minimise the set of error terms generated by the background pixels $\mathcal{B}_t$ in the frame $I_t(\mathbf{x})$. Using the error terms of equation 12, this error functional may be defined in terms of the next motion estimate $\mathbf{a}_{t,i+1}$ as follows:

$$\epsilon(\mathbf{a}_{t,i+1}) \stackrel{\text{def}}{=} \sum_{\mathbf{x}\in\mathcal{B}_t} e^b(\mathbf{x}, \mathbf{a}_{t,i+1})^2 + \sum_{\mathbf{x}\in\mathcal{B}_t} e^f(\mathbf{x}, \mathbf{a}_{t,i+1})^2. \tag{13}$$

Setting to zero the partial derivatives of the above functional with respect to $\mathbf{a}_{t,i+1}$ generates the following iterative estimator for $\mathbf{a}$:

$$\begin{aligned} \mathbf{a}_{t,i+1} = \frac{1}{2}\Bigg\{\sum_{\mathbf{x}\in\mathcal{B}_t} X(\mathbf{x})^T \nabla I_t(\mathbf{x}) \nabla I_t(\mathbf{x})^T X(\mathbf{x})\Bigg\}^{-1} \Bigg\{\sum_{\mathbf{x}\in\mathcal{B}_t} X(\mathbf{x})^T \nabla I_t(\mathbf{x}) \\ \left[\nabla I_t(\mathbf{x})^T \left(\Delta\mathbf{x}_t^{t+1}(\mathbf{x}, \mathbf{a}_{t,i}) - \Delta\mathbf{x}_t^{t-1}(\mathbf{x}, \mathbf{a}_{t,i})\right) - \left(I_{t+1}(\mathbf{x}, \mathbf{a}_{t,i}) - I_{t-1}(\mathbf{x}, \mathbf{a}_{t,i})\right)\right]\Bigg\}, \end{aligned} \tag{14}$$

where the first parameter estimate $\mathbf{a}_{t,1}$ can be computed from the initial displacements $\Delta\mathbf{x}_{t,0}(\mathbf{x})$ or from the projected parameters $\mathbf{a}_{t-1}$ of the previous motion.

One problem with the above estimator is that the set of pixels $\mathcal{B}_t$ to which the motion model is fitted must belong to the same moving region, *ie*, the background. The inclusion in the estimation process of any foreground pixels will significantly corrupt the motion model. While redefining a *robust-statistical* version of the estimator given in equation 14 is straightforward [4, 10], empirical evidence suggests that robust-statistical estimators do not automatically provide robust solutions. For M-estimators, this is due partly to the high breakdown value for an 8-parameter motion model. The main reason is that not all outliers necessarily generate high greylevel differences.

In the motion segmentation stage described in the next section, pixels are classified as *background*, *uncovered*, *covered* or *foreground*. The set $\mathcal{B}_t$ is defined as those pixels in the current frame I_t whose motion-compensated pixel position in the previous frame I_{t-1} has been classified as *background* or *uncovered*. Currently, the initial set $\mathcal{B}_0$ for the first frame in a sequence is the whole image; in *post-production* applications, which motivate this work, an initial crudely drawn mask or *matte* could be supplied by the special-effects operator to indicate roughly the background region.

Figure 1 shows three quarter-PAL-size frames from the *Kathy* sequence in which a foreground figure moves rapidly against a rigid distant slow-moving background. The corresponding forward and backward error maps, $e^f(\mathbf{x}, \widehat{\mathbf{a}}_t)$ and $e^b(\mathbf{x}, \widehat{\mathbf{a}}_t)$, recovered from the minimisation process, are shown in figure 1(d) and (e) respectively. As is typical of most image sequences, the foreground figure consists of a substantial minority of pixels with low greylevel error.

(a) Previous frame (b) Current frame (c) Next frame

Figure 1: (d) Forward error map (e) Backward error map

3 Segmenting motion fields

Once the global motion of the image has been computed, each pixel may be classified from the class set $\Lambda = \{\lambda_B, \lambda_U, \lambda_C, \lambda_F\}$ representing the following classes respectively: *background*, where neither forward and backward error values e^f and e^f change significantly; *uncovered* pixels, where only the forward error changes at the onset of occlusion; *covered* pixels, where a previous high error due to occlusion subsequently becomes low; and finally *foreground*, where both forward and backward errors remain high. Let λ_0 and λ_1 represent the *unchanged* and *changed* classes respectively, and $p(e|\lambda_0)$ and $p(e|\lambda_1)$ their *a priori* probability density functions. Typical probability density functions are shown in figure 2.

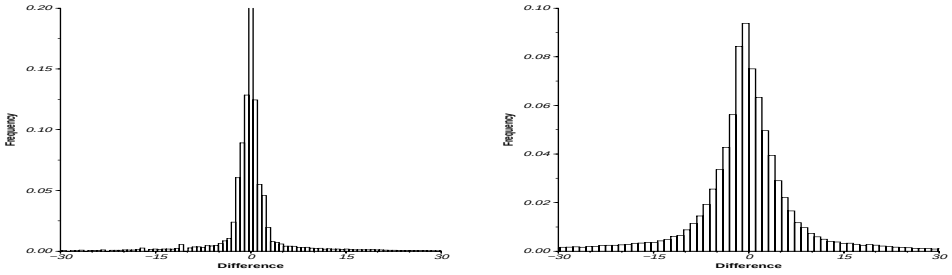


Figure 2: (a) No-change PDF (b) Change PDF

By combining the forward and backward error values e^f and e^b to generate a measurement vector $\mathbf{e} = (e^f, e^b)$, the most suitable class for each pixel may be computed using the standard maximum *a posteriori* probability decision rule

$$\widehat{\lambda} = \underset{\lambda \in \Lambda}{\operatorname{argmax}} \quad p(\mathbf{e}|\lambda)p(\lambda). \tag{15}$$

Since e^f and e^b are assumed to be independent, the *a priori* probabilities are:

$$\begin{aligned} p(\mathbf{e}|\lambda_B) &= p(e^f|\lambda_0)p(e^b|\lambda_0), \qquad & p(\mathbf{e}|\lambda_U) &= p(e^f|\lambda_0)p(e^b|\lambda_1), \\ p(\mathbf{e}|\lambda_C) &= p(e^f|\lambda_1)p(e^b|\lambda_0), \qquad & p(\mathbf{e}|\lambda_F) &= p(e^f|\lambda_1)p(e^b|\lambda_1). \end{aligned} \tag{16}$$

The *a posteriori* probability maps for the *background, uncovered, covered* and *foreground* classes given for the error maps in figure 1 are shown in figure 3. For clarity, high probability values are shown as darker shades. Note that the foreground element is moving relative to the moving background from bottom left to top right at an average velocity of 2 pixels per frame.

3.1 Temporal probabilistic updating

The classification generated by the above procedure is unsatisfactory: the classification image is speckled with incorrect labels due to noise, and, more significantly, foreground pixels in areas of low greylevel variation tend to exhibit low greylevel differences and hence become misclassified. We therefore introduce a temporal labelling to exploit the temporal continuity of pixel labels. We wish to compute the multiple-label *a posteriori* probability $p(\lambda_t, \lambda_{t-1}, \cdots | \mathbf{e}_t, \mathbf{e}_{t-1}, \cdots)$.

Assuming that the label λ_t in current frame depends only on the current error $\mathbf{e}_t$ and the label λ_{t-1} in the previous frame, and that current and previous error estimates are independent, $p(\lambda_t, \lambda_{t-1}, \cdots | \mathbf{e}_t, \mathbf{e}_{t-1}, \cdots)$ may be rewritten as

$$\begin{aligned} & p(\lambda_t, \lambda_{t-1}, \cdots | \mathbf{e}_t, \mathbf{e}_{t-1}, \cdots) \\ = \quad & p(\mathbf{e}_t|\lambda_t) p(\lambda_t|\lambda_{t-1}) p(\lambda_{t-1}, \lambda_{t-2}, \cdots | \mathbf{e}_{t-1}, \mathbf{e}_{t-2}, \cdots) \frac{p(\mathbf{e}_{t-1}, \mathbf{e}_{t-2}, \cdots)}{p(\mathbf{e}_t, \mathbf{e}_{t-1}, \cdots)}. \end{aligned} \tag{17}$$

This temporally recursive expression suggests the following decision rule for determining the current label $\widehat{\lambda}_t$ of a pixel from the class conditional probabilities computed in the previous frame:

$$\widehat{\lambda}_t = \underset{\lambda_t \in \Lambda}{\operatorname{argmax}} \left\{ p(\mathbf{e}_t|\lambda_t) \quad \max_{\lambda_{t-1} \in \Lambda} \left\{ p(\lambda_t|\lambda_{t-1}) p(\lambda_{t-1}, \widehat{\lambda}_{t-2}, \cdots | \mathbf{e}_{t-1}, \mathbf{e}_{t-2}, \cdots) \right\} \right\}. \tag{18}$$

Figure 4 compares this rule and the non-temporal rule in equation 15, applying each to the *Kathy* sequence. In the temporal classification, note the greater density of foreground-labelled pixels, reduced speckle and, most importantly, the correct labelling of covered/uncovered pixels at the edges of the actress element showing the direction of motion. Holes remain however in the foreground at low greylevel gradients.

The *a priori* class probabilities $p(\lambda)$ and temporal class association probabilities $p(\lambda_x|\lambda_y)$ are user-initialised by estimating the fraction of pixels in each class in a sequence. For this 30-frame *Kathy* sequence, these values were selected empirically:

$p(\lambda_B\|\lambda_B) = 0.96$	$p(\lambda_B\|\lambda_U) = 1.00$	$p(\lambda_B\|\lambda_C) = 0.00$	$p(\lambda_B\|\lambda_F) = 0.00$
$p(\lambda_U\|\lambda_B) = 0.00$	$p(\lambda_U\|\lambda_U) = 0.00$	$p(\lambda_U\|\lambda_C) = 0.00$	$p(\lambda_U\|\lambda_F) = 0.13$
$p(\lambda_C\|\lambda_B) = 0.04$	$p(\lambda_C\|\lambda_U) = 0.00$	$p(\lambda_C\|\lambda_C) = 0.00$	$p(\lambda_C\|\lambda_F) = 0.00$
$p(\lambda_F\|\lambda_B) = 0.00$	$p(\lambda_F\|\lambda_U) = 0.00$	$p(\lambda_F\|\lambda_C) = 1.00$	$p(\lambda_F\|\lambda_F) = 0.87$
$p(\lambda_B) = 0.74$	$p(\lambda_U) = 0.03$	$p(\lambda_C) = 0.03$	$p(\lambda_F) = 0.20$

4 Conclusions

Motion segmentation is a complex problem in which motion estimation and pixel classification are intimately related. We have developed a highly accurate global motion estimator that makes use of a spatiotemporal framework to estimate an optical-flow field incorporating a motion model motivated by three-dimensional camera motion. Once motion has been computed, pixels are classified as *background, uncovered, covered* or *foreground* on the basis of interframe greylevel differences. The fact that these differences may be

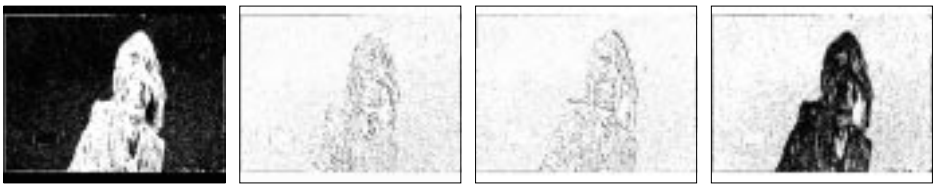


Figure 3: (*left to right*) background, uncovered, covered and foreground

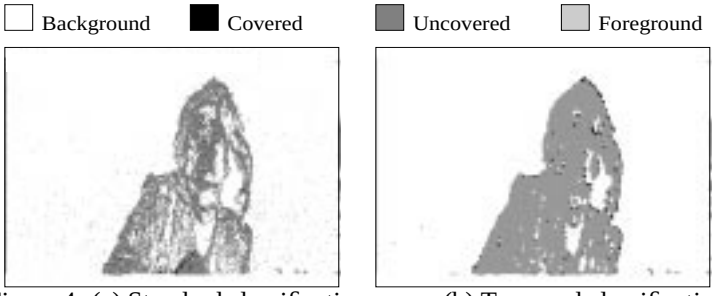
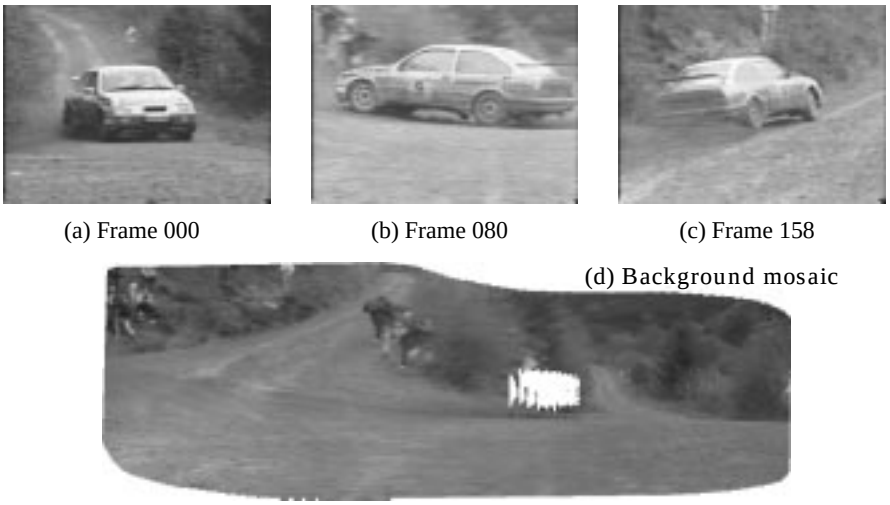


Figure 4: (a) Standard classification (b) Temporal classification

legitimately low in foreground elements, coupled with noise, can result in poor segmentation of foreground from background. Consequently, a temporal dimension is introduced that encodes information about the allowed (or expected) sequential classifications of a motion-compensated pixel over time. This temporal integration results in a significant improvement in classification and a denser, more noise-resistant segmentation.

A final demonstration of the use of our segmentation for a cinematographic special effect is shown in figure 5. Background motion is calculated for each image in a 160-frame sequence of a rally. The inevitably incomplete background mattes are filled by a operator to remove the car feature entirely. The global motion estimates are then used to merge the background mattes into a panoramic mosaic [6, 7, 8].



(a) Frame 000 (b) Frame 080 (c) Frame 158

(d) Background mosaic

Figure 5: Mosaicking segmented background mattes

5 Acknowledgement

The authors gratefully acknowledge the support and cinematographic data kindly provided by the Computer Film Company Ltd.

References

[1] P. Anandan, J.R. Bergen, K.J. Hanna, and R. Hingorani. *"Motion Analysis and Image Sequence Processing"*, chapter "Hierarchical Model-Based Motion Estimation", pages 1–22. Ed. M. Ibrahim Sezan and Reginald L. Lagendijk, Kluwer Academic Publishers, Boston, 1993.

[2] S. Ayer, P. Schroeter, and J. Bigün. "Segmentation of moving objects by robust motion parameter estimation over multiple frame". In *Proceedings of European Conference on Computer Vision*, pages 316–327, Stockholm, 1994.

[3] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. "Performance of Optical Flow Techniques". *International Journal of Computer Vision*, 12(1):43–77, 1994.

[4] M. Bober and J. Kittler. "Robust Motion Analysis". In *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, pages 947–952, 1994.

[5] B. Duc, P. Schroeter, and J. Bigün. "Spatio-Temporal Robust Motion Estimation and Segmentation". In *6th Int. Conf. Computer Analysis of Images and Patterns*, pages 238–245, Prague, September 1995. Springer-Verlag.

[6] E. François. "Rigid Layers Reconstruction Based on Motion Segmentation". In *Workshop on Image Analysis for Multimedia Interactive Services*, pages 81–86, Louvain-la-Neuve, Belgium, 24–25 June 1997.

[7] M. Gelgon and P. Bouthemy. "A Hierarchical Motion-based Segmentation and Tracking Technique for Video Storyboard-like Representation and Content-based Indexing". In *Workshop on Image Analysis for Multimedia Interactive Services*, pages 93–98, Louvain-la-Neuve, Belgium, 24–25 June 1997.

[8] P. Giaccone, D. Greenhill, and G.A. Jones. "Segmenting, Describing and Compositing Video Sequences Containing Multiple Moving Elements". In *TV and Broadcasting on Internet, WWW and Networks*, Bradford, UK, 22–23 April 1998.

[9] P.R. Giaccone and G.A. Jones. "Spatio-Temporal Approaches to the Computation of Optical Flow". In *Proceedings of the British Machine Vision Conference*, pages 420–429, Colchester, UK, September 1997.

[10] E.-P. Ong and M. Spann. "Robust Computation of Optical Flow". In *Proceed. of the British Machine Vision Conference*, volume 2, pages 573–582, 1995.

[11] N. Paragios and G. Tziritas. "Detection and Location of Moving Objects Using Deterministic Relaxation Algorithms". In *Proceedings of IEEE International Conf. Pattern Recognition*, 1996.

[12] C. Ponticos. "A Robust Real Time Face Location Algorithm for Videophones". In *Proceedings of the British Machine Vision Conference*, pages 499–458, Guildford, UK, September 1993.

[13] S. Tubaro and F. Rocca. *"Motion Analysis and Image Sequence Processing"*, chapter "Motion Field Estimators and their Application to Image Interpretation", pages 153–187. Ed. M. Ibrahim Sezan and Reginald L. Lagendijk, Kluwer Academic Publishers, Boston, 1993.