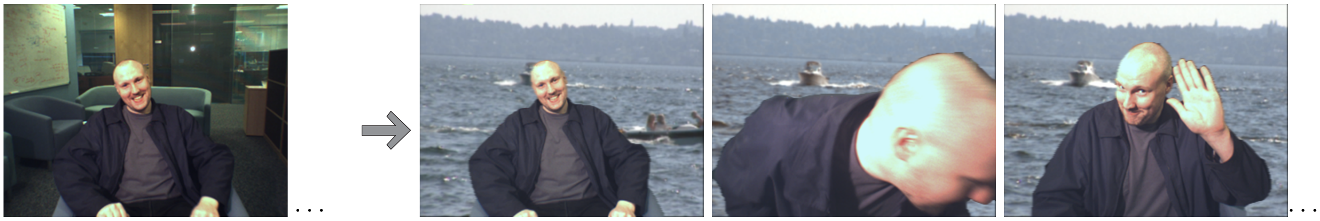
Bilayer Segmentation of Live Video

A. Criminisi, G. Cross, A. Blake, V. Kolmogorov
Microsoft Research Ltd., Cambridge, CB3 0FB, United Kingdom
http://research.microsoft.com/vision/cambridge/i2i



a *input sequence* **b** *automatic layer separation and background substitution in three different frames*

Figure 1: **An example of automatic foreground/background segmentation in monocular image sequences.** Despite the challenging foreground motion the person is accurately extracted from the sequence and then composited free of aliasing upon a different background; a useful tool in video-conferencing applications. The sequences and ground truth data used throughout this paper are available from [1].

Abstract

This paper presents an algorithm capable of real-time separation of foreground from background in monocular video sequences.

Automatic segmentation of layers from colour/contrast or from motion alone is known to be error-prone. Here motion, colour *and* contrast *cues are probabilistically fused together with* spatial *and* temporal priors *to infer layers accurately and efficiently. Central to our algorithm is the fact that pixel velocities are not needed, thus removing the need for optical flow estimation, with its tendency to error and computational expense. Instead, an efficient motion vs non-motion classifier is trained to operate directly and jointly on intensity-change and contrast. Its output is then fused with colour information. The prior on segmentation is represented by a second order, temporal, Hidden Markov Model, together with a spatial MRF favouring coherence except where contrast is high. Finally, accurate layer segmentation and explicit occlusion detection are efficiently achieved by binary graph cut.*

The segmentation accuracy of the proposed algorithm is quantitatively evaluated with respect to existing ground-truth data and found to be comparable to the accuracy of a state of the art stereo segmentation algorithm. Foreground/background segmentation is demonstrated in the application of live background substitution and shown to generate convincingly good quality composite video.

1. Introduction

This paper addresses the problem of accurately extracting a foreground layer from video in real time. A prime application is live background substitution in teleconferencing. This demands layer separation to near Computer Graphics quality, including transparency determination as in video-matting [8, 9], but with computational efficiency sufficient to attain live streaming speed.

Layer extraction from images or sequences has long been an active area of research [2, 4, 10, 13, 20, 21, 22, 23]. The challenge addressed here is to segment the foreground layer *efficiently* without restrictions on appearance, motion, camera viewpoint or shape, and sufficiently *accurately* for use in background substitution and other synthesis applications. Frequently, motion-based segmentation has been achieved by estimating optical flow (*i.e.* pixel velocities) [3] and then grouping pixels into regions according to predefined motion models. Spatial priors can also be imposed by means of graph-cut [7, 12, 13, 22, 23]. However, the grouping principle generally requires some assumption about the nature of the underlying motion (translational, affine etc.), which is restrictive. Furthermore, regularization to constrain ill posed optical flow solutions tends to introduce undesirable inaccuracies along layer boundaries. Lastly, accurate estimation of optical flow is computationally expensive, requiring an extensive search in the neighbourhood of each point. In our approach, explicit estimation of pixel velocities is altogether avoided. Instead, an efficient discriminative model, to separate motion from stasis using spatio-temporal derivatives, is learned from labelled data.

Recently, interactive segmentation techniques exploiting colour/contrast cues have been demonstrated to be very effective for static images [6, 16]. Segmentation based on colour/contrast alone is nonetheless beyond the capability of fully automatic methods. This suggests a robust approach that fuses a variety of cues, for example stereo, colour, contrast and spatial priors [11] is known to be effective and computable comfortably in real time. This paper shows that comparable segmentation accuracy can be achieved monocularly, avoiding the need for stereo cameras with their inconvenient necessity for calibration. Efficency with motion in place of stereo is actually enhanced, in that stereo match likelihoods need no longer be evaluated, and the other significant computational costs remaining approximately the same. Additionally, temporal consistency is imposed for increased segmentation accuracy, and temporal transitions probabilities are modelled with reduction of flicker artifacts and explicit detection of temporal occlusions.

Notation and image observables. Given an input sequence of images, a frame is represented as an array $\mathbf{z} = (z_1, z_2, \cdots, z_n, \cdots, z_N)$ of pixels in YUV colour space, indexed by the single index n. The frame at time t is denoted $\mathbf{z}^t$. Temporal derivatives are denoted

$$\dot{\mathbf{z}} = (\dot{z}_1, \dot{z}_2, \cdots, \dot{z}_n, \cdots, \dot{z}_N), \tag{1}$$

and at each time t, are computed as $\dot{z}_n^t = |G(z_n^t) - G(z_n^{t-1})|$ with $G(.)$ a Gaussian kernel at the scale of σ_t pixels. Then, also spatial gradients

$$\mathbf{g} = (g_1, g_2, \cdots, g_n, \cdots, g_N) \text{ where } g_n = |\nabla z_n|, \tag{2}$$

are computed by convolving the images with first-order derivative of Gaussian kernels with standard deviation σ_s. Here we use $\sigma_s = \sigma_t = 0.8$, approximating a Nyquist sampling filter. Spatio-temporal derivatives are computed on the Y colour-space channel only. Motion observables are denoted

$$\mathbf{m} = (\mathbf{g}, \dot{\mathbf{z}}) \tag{3}$$

and are used as the raw image features for discrimination between motion and stasis.

Segmentation is expressed as an array of opacity values $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_n, \cdots, \alpha_N)$. We focus on binary segmentation, i.e. $\alpha \in \{\mathtt{F}, \mathtt{B}\}$ with F and B denoting foreground and background, respectively. Fractional opacities are discussed briefly in section 3.

2. Probabilistic segmentation model

This section describes the probabilistic model for foreground/background segmentation, in an energy minimization framework. This extends previous energy models for segmentation [6, 11, 16] by the addition of a second order, temporal, Markov Chain prior, and an observation likelihood for image motion. The posterior model is a Conditional Random Field (CRF) [15] with a factorisation that contains some recognisably generative structure, and this is used to determine the precise algebraic forms of the factors. Various parameters are then set discriminatively [14]. The CRF is denoted

$$\begin{aligned} p(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^t \mid \mathbf{z}^1, \ldots, \mathbf{z}^t,\ \mathbf{m}^1, \ldots, \mathbf{m}^t) \\ \propto \exp -\left\{ \sum_{t'=1}^{t} E^{t'} \right\} \end{aligned} \tag{4}$$

$$\text{where } E^t = E(\boldsymbol{\alpha}^t, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}^{t-2}, \mathbf{z}^t, \mathbf{m}^t). \tag{5}$$

Note the second order temporal dependence in the Markov model, to be discussed more fully later. The whole aim is to estimate $\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^t$ given the image and motion data, and in principle this would be done by joint maximisation of the posterior, or equivalently minimisation of energy:

$$(\hat{\boldsymbol{\alpha}}^1, \ldots, \hat{\boldsymbol{\alpha}}^t) = \arg\min \sum_{t'=1}^{t} E^{t'}. \tag{6}$$

However, such batch computation is of no interest for real-time applications because of the causality constraint — each $\hat{\boldsymbol{\alpha}}^{t'}$ must be delivered on the evidence from its past, without using any evidence from the future. Therefore estimation will be done by separate minimisation of each term E^t and details are given later.

2.1. Conditional Random Field energy terms

The Energy E^t associated with time t is a sum of terms in which likelihood and prior are not entirely separated, and so does not represent a pure generative model, although some of the terms have clearly generative interpretations. The energy decomposes as a sum of four terms:

$$\begin{aligned} E(\boldsymbol{\alpha}^t, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}^{t-2}, \mathbf{z}^t, \mathbf{m}^t) = \\ V^{\mathrm{T}}(\boldsymbol{\alpha}^t, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}^{t-2}) + V^{\mathrm{S}}(\boldsymbol{\alpha}^t, \mathbf{z}^t) \\ + U^{\mathrm{C}}(\boldsymbol{\alpha}^t, \mathbf{z}) + U^{\mathrm{M}}(\boldsymbol{\alpha}^t, \boldsymbol{\alpha}^{t-1}, \mathbf{m}^t), \end{aligned} \tag{7}$$

in which the first two terms are "prior-like" and the second two are observation likelihoods. Briefly, the roles of the four terms are:

Temporal prior term $V^{\mathrm{T}}(\ldots)$ is a second-order Markov chain that imposes a tendency to temporal continuity of segmentation labels.

Spatial prior term $V^{\mathrm{S}}(\ldots)$ is an Ising term, imposing a tendency to spatial continuity of labels, and the term is inhibited by high contrast.

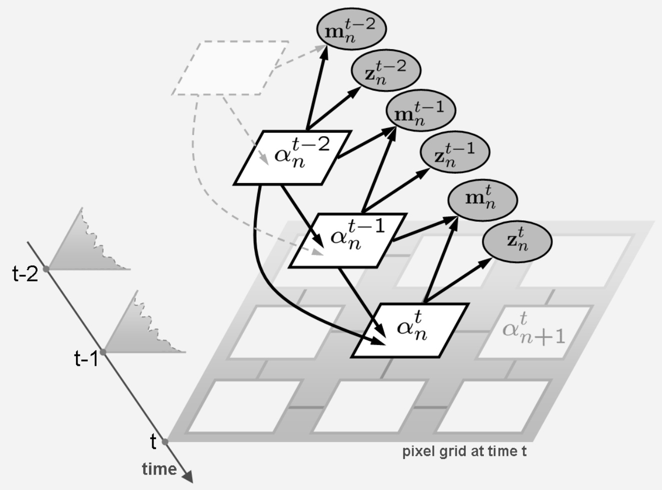


Figure 2. **Spatio-temporal Hidden Markov Model.** This graphical model illustrates both the colour likelihood and motion likelihoods together with the spatial and temporal priors. The same temporal chain is repeated at each pixel position. Spatial dependencies are illustrated for a 4-neighborhood system.

Colour likelihood term $U^{\mathrm{C}}(\ldots)$ evaluates the evidence for pixel labels based on colour distributions in foreground and background.

Motion likelihood term $U^{\mathrm{M}}(\ldots)$ evaluates the evidence for pixel labels based on the expectation of stasis in the background and frequently occurring motion in the foreground. Note that motion $\mathbf{m}^t$ is explained in terms of the labelling at both the current frame $\boldsymbol{\alpha}^t$ and the previous one $\boldsymbol{\alpha}^{t-1}$.

This energy resembles a spatio-temporal Hidden Markov Model (HMM), and this is illustrated graphically in figure 2. Details of the "prior" and likelihood factors are given in the remainder of this section.

2.2. Temporal prior term

Figure 3 illustrates the four different kinds of temporal transitions a pixel can undergo in a bilayer scene, based on a two-frame analysis. For instance, a foreground pixel may remain in the foreground (pixels labelled FF in fig. 3c) or move to the background (pixels labelled FB) etc. The critical point here is that a first-order Markov chain is inadequate to convey the nature of temporal coherence in this problem; a second-order Markov chain is required. For example, a pixel that was in the background at time $t-2$ and is in the foreground at time $t-1$ is far more likely to remain in the foreground at time t than to go back to the background. Note that BF and FB transitions correspond to temporal occlusion and disocclusion events, and that a pixel cannot change layer without going through an occlusion event.

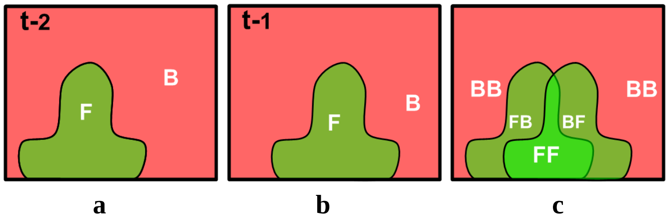


Figure 3. **Temporal transitions at a pixel.** (a,b) An object moves towards the right from frame $t-2$ to frame $t-1$. (c) Between the two frames pixels may remain in their own foreground or background layer (denoted F and B, respectively) or change layer; thus defining four different kinds of temporal transitions: $B \to B$, $F \to B$, $F \to F$, $B \to F$. Those transitions influence the label that a pixel is going to assume at frame t.

α^{t-1}	α^{t-2}	$p(\alpha^t = \mathtt{F} \mid \alpha^{t-1}, \alpha^{t-2})$
F	F	β_{FF}
F	B	β_{FB}
B	F	β_{BF}
B	B	β_{BB}

Figure 4. **Learned priors for temporal transitions.** The background probabilities are the complement of the foreground ones. See text for details.

These intuitions are captured probabilistically and incorporated in our energy minimization framework by means of a second order Markov chain, as illustrated in the graphical model of fig. 2. The temporal transition priors are learned from labelled data and then stored in a table, as in fig. 4. Note that despite there being eight (2^3) possible transitions, due to probabilistic normalization ($p(\alpha^t = \mathtt{B}|\alpha^{t-1}, \alpha^{t-2}) = 1 - p(\alpha^t = \mathtt{F}|\alpha^{t-1}, \alpha^{t-2})$) the temporal prior table has only four degrees of freedom, represented by the four parameters $\beta_{\mathrm{FF}}, \beta_{\mathrm{FB}}, \beta_{\mathrm{BF}}, \beta_{\mathrm{BB}}$. This leads to the following joint temporal prior term:

$$V^{\mathrm{T}}(\boldsymbol{\alpha}^t, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}^{t-2}) = \eta \sum_{n}^{N} \left[-\log p(\alpha_n^t | \alpha_n^{t-1}, \alpha_n^{t-2}) \right] \tag{8}$$

in which $\eta < 1$ is a discount factor to allow for multiple counting across non-independent pixels. As explained later the optimal value of η (as well as the other parameters of the CRF) is trained discriminatively from ground-truth.

2.3. Ising spatial energy

There is a natural tendency for segmentation boundaries to align with contours of high image contrast. Similarly to [6, 16], this is represented by an energy term of the form

$$V^{\mathrm{S}}(\boldsymbol{\alpha}, \mathbf{z}) = \gamma \sum_{(m,n) \in \mathbf{C}} [\alpha_m \neq \alpha_n] \left(\frac{\epsilon + e^{-\mu ||z_m - z_n||^2}}{1 + \epsilon} \right) \tag{9}$$

where (m, n) index neighbouring pixel-pairs. $\mathbf{C}$ is the set of pairs of neighbouring pixels. The contrast parameter μ is chosen to be $\mu = \left(2\left\langle \|z_m - z_n\|^2 \right\rangle\right)^{-1}$; where $< . >$ denotes expectation over all pairs of neighbours in an image sample. The energy term $V(\boldsymbol{\alpha}, \mathbf{z})$ represents a combination of an Ising prior for labelling coherence together with a contrast likelihood that acts to discount partially the coherence terms. The constant γ is a strength parameter for the coherence prior and also the contrast likelihood. The constant ϵ is a "dilution" constant for contrast, previously [6] set to $\epsilon = 0$ for pure colour segmentation. However, multiple cue experiments with colour and stereo [11] have suggested $\epsilon = 1$ as a more appropriate value.

2.4. Likelihood for colour

The term $U^{\mathrm{C}}(.)$ in (7) is the log of the colour likelihood. In [5, 11, 16] colour likelihoods were modeled in terms of Gaussian Mixture Models (GMM) in RGB, where foreground and background mixtures were learned via Expectation Maximization (EM). However, we have found that issues with the initialization of EM and with local minima affect the discrimination power of the final likelihood ratio. Instead, here we model the foreground and background colour likelihoods non-parametrically, as histograms in the YUV colour space. The colour term $U^{\mathrm{C}}(.)$ is defined as:

$$U^{\mathrm{C}}(\boldsymbol{\alpha}, \mathbf{z}) = -\rho \sum_{n}^{N} \log p(z_n|\alpha_n). \qquad (10)$$

Probabilistic normalization requires that $\sum_z p(z|\alpha = \mathtt{F}) = 1$, and similarly for the background likelihood. This non-parametric representation negates the need for having to set the number of GMM components as well as having to wait for EM convergence.

The foreground colour likelihood model is learned adaptively over successive frames, similarly to [11], based on data from the segmented foreground in the previous frame. The likelihoods are then stored in 3D look-up tables, constructed from the raw colour histograms, with a modest degree of smoothing, to avoid overlearning. The background colour distribution is constructed from an initial extended observation of the background, rather as in [17, 18], to build in variability of appearance. The distribution is then static over time. It is also shared by the entire background, to give additional robustness against camera shake, and studies suggest that the loss of precision in segmentation, compared with pixelwise colour models (such as those used in [19]), should not be very great [11]. Again, the distribution is represented as a smoothed histogram, rather than as a Gaussian mixture, to avoid the problems with initialisation.

2.5. Likelihood for motion

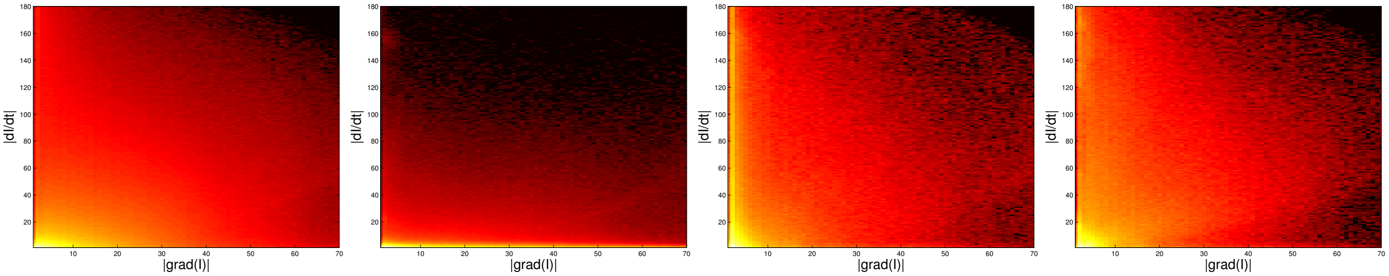
The treatment of motion could have been addressed via an intermediate computation of optical flow. However reliable computation of flow is expensive and beset with difficulties concerning the aperture problem and regularisation. Those difficulties can be finessed in the segmentation application by bypassing flow and modelling directly the characteristics of the feature normally used to compute flow, namely the spatial and temporal derivatives $\mathbf{m} = (\mathbf{g}, \dot{\mathbf{z}})$. The motion likelihood therefore captures the characteristics of those features under foreground and background conditions respectively.

However, the nature of our generative model suggests an approach to motion likelihood modelling that should capture even richer information about segmentation. Referring back to fig. 3, the immediate history of the segmentation of a pixel falls into one of four classes, FF, BB, FB, BF. We model the observed image motion features $\mathbf{m}_n^t = (g_n^t, \dot{z}_n^t)$, at time t and for pixel n, as conditioned on those combinations of the segmentation labels α_n^{t-1} and α_n^t. This is a natural model because the temporal derivative $\dot{z}_n^t$ is computed from frames $t-1$ and t, so clearly it should depend on segmentations of those frames. Illustrations of the joint distributions learned for each of the four label combinations are shown in figure 5. Empirically, the BB distribution reflects the relative constancy of the background state — temporal derivatives are small in magnitude. The FF distribution reflects larger temporal change, and as expected that is somewhat correlated with spatial gradient magnitude. Transitional FB and BF distributions show the largest temporal changes since the temporal samples at time $t-1$ and t straddle an object boundary. Note that the distributions for BF and FB are distinct in shape from those for BB and FF, and this is one indication that the second order model does indeed capture additional motion information, compared with a first order model. (The first order model would be conditioned on just F and B, for which the likelihoods are essentially identical to those for FF and BB, as illustrated in the figure.)

The four motion likelihoods are learned from some labelled ground truth data and then stored as 2D histograms (smoothed) to use in likelihood evaluation. The likelihoods are evaluated as part of the total energy, in the term

$$U^{\mathrm{M}}(\boldsymbol{\alpha}^t, \boldsymbol{\alpha}^{t-1}, \mathbf{m}^t) = -\sum_n \log p(\mathbf{m}_n^t \mid \alpha_n^t, \alpha_n^{t-1}). \quad (11)$$

Illustrating the motion likelihoods. Figure 6 shows the results of a likelihood ratio test using the likelihood ratio R of the FF model versus the BB model, applied to *two* frames of the VK test sequence. Motion and non-motion events are accurately separated in textured areas. In fact, moving edges are clearly marked with bright pixels ($R > 1$) while stationary edges are marked with dark pixels ($R < 1$). However, textureless regions remain ambiguous and are automatically assigned a likelihood ratio close to unity (mid-grey in figure). This suggests that motion alone is not sufficient for an accurate segmentation. Fusing motion and



$\log p\left(\mathbf{m}^t|(\alpha^t,\alpha^{t-1})=(\texttt{F},\texttt{F})\right)$ $\log p\left(\mathbf{m}^t|(\alpha^t,\alpha^{t-1})=(\texttt{B},\texttt{B})\right)$ $\log p\left(\mathbf{m}^t|(\alpha^t,\alpha^{t-1})=(\texttt{B},\texttt{F})\right)$ $\log p\left(\mathbf{m}^t|(\alpha^t,\alpha^{t-1})=(\texttt{F},\texttt{B})\right)$

Figure 5. **Learned motion likelihoods.** Learned likelihood of motion data, conditioned on the segmentation in the two previous frames. Yellow indicates high density, red medium density and black zero density. The distributions are modelled simply as normalized histograms.

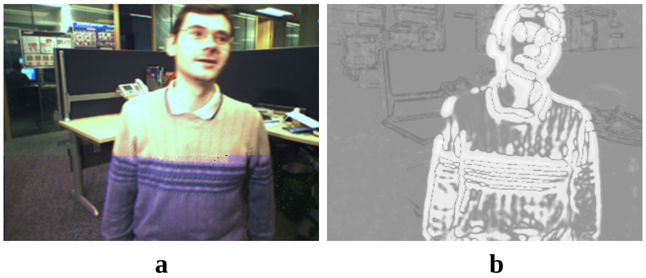


Figure 6. **Testing the motion classifier.** (a) A frame from the VK test sequence. (b) The corresponding motion likelihood map as output of a likelihood ratio test, see text for details. Bright pixels indicate motion (likelihood ratio $R > 1$) and dark ones stasis ($R < 1$). Thanks to our joint motion likelihood strong stationary edges are assigned a lower (more negative, darker) value of R than stationary textureless areas.

colour with CRF spatial and Markov chain temporal priors as in (7) is expected to help fill the remaining gaps. In stereo as opposed to motion segmentation [11], it is known that good segmentation can be achieved even without the temporal model. However, as we show later, the gaps in the motion likelihood demand also the temporal model for satisfactory filling in.

2.6. Inference by energy minimisation

At the beginning of this section the principal aim of estimation was stated to be the maximisation of the joint posterior (6). However, it was also plain that the constraints of causality in real-time systems do not allow that. Under causality, having estimated $\hat{\boldsymbol{\alpha}}^1,\ldots,\hat{\boldsymbol{\alpha}}^{t-1}$, one way to estimate $\hat{\boldsymbol{\alpha}}^t$ would simply be:

$$\hat{\boldsymbol{\alpha}}^t = \arg\min\ E(\boldsymbol{\alpha}^t,\hat{\boldsymbol{\alpha}}^{t-1},\hat{\boldsymbol{\alpha}}^{t-2},\mathbf{z}^t,\mathbf{m}^t). \tag{12}$$

Freezing all estimators before generating t is an extreme approach, and better results are obtained by acknowledging the variability in at least the immediately previous timestep. Therefore, the energy in (12) is replaced by the expected energy:

$$\mathcal{E}_{\boldsymbol{\alpha}^{t-1}|\hat{\boldsymbol{\alpha}}^{t-1}} E(\boldsymbol{\alpha}^t,\boldsymbol{\alpha}^{t-1},\hat{\boldsymbol{\alpha}}^{t-2},\mathbf{z}^t,\mathbf{m}^t). \tag{13}$$

where the conditional density for time $t-1$ is modelled as

$$p(\boldsymbol{\alpha}^{t-1}|\hat{\boldsymbol{\alpha}}^{t-1}) = \prod_n p(\alpha_n^{t-1}|\hat{\alpha}_n^{t-1}), \tag{14}$$

and

$$p(\alpha^{t-1}|\hat{\alpha}^{t-1}) = \nu + (1-\nu)\delta(\alpha^{t-1},\hat{\alpha}^{t-1}), \tag{15}$$

and ν (with $\nu \in [0,1]$) is the degree to which the binary segmentation at time $t-1$ is “softened” to give a segmentation distribution. In practice, allowing $\nu > 0$ (typically $\nu = 0.1$), prevents the segmentation becoming erroneously “stuck” in either foreground or background states.

This factorisation of the segmentation distribution across pixels makes the expectation computation (13) entirely tractable. The alternative of fully representing uncertainty in segmentation is computationally too costly. Finally, the segmentation $\hat{\boldsymbol{\alpha}}^t$ is computed by binary graph cut [7].

3. Experimental results

This section validates the proposed segmentation algorithm through comparison both with stereo-based segmentation, and with hand-labelled ground truth [1].

Bilayer segmentation, alpha matting and background substitution. In fig. 7 foreground and background of a video-sequence have been separated automatically. After an initial period where the subject is almost stationary (fig. 7a), the segmentation quickly converges to a good solution. Real-time “border matting” [16] has been used to compute fractional opacities along the boundary and this is used for anti-aliased compositing onto a new background (fig. 1b). Segmentation and background substitution for another test sequence is demonstrated in fig. 8. Notice that good segmentation is achieved even in frames containing rapid motion, as in figs. 1b, 7e and fig. 8e.

Detecting temporal occlusions. Figure 9 shows examples of temporal occlusion detection for the JM sequence,

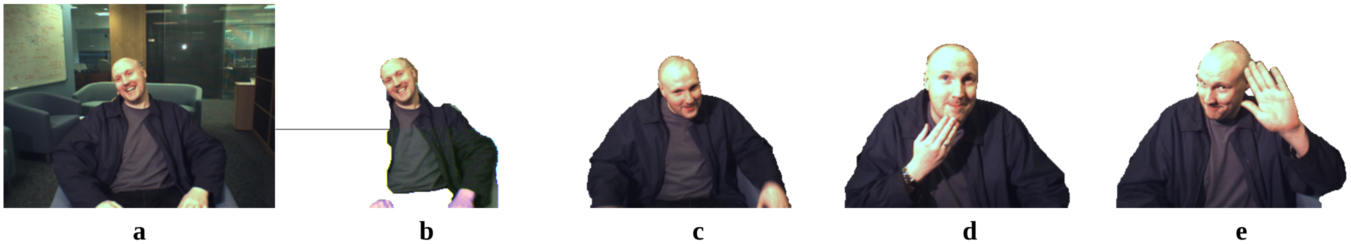


Figure 7. **Bilayer segmentation.** (a) A frame from the JM test sequence; (b,...,e) automatic foreground extraction results for several frames. (b) In the first frames only parts of the person are moving and therefore the segmentation is not accurate. (c,...,e) However, after only a few frames the segmentation converges to the correct solution. After this initial stage, the model is burnt in, and can tolerate periods of stasis. The extracted foreground can now be composited with anti-aliasing onto new backgrounds, as in fig.1b.

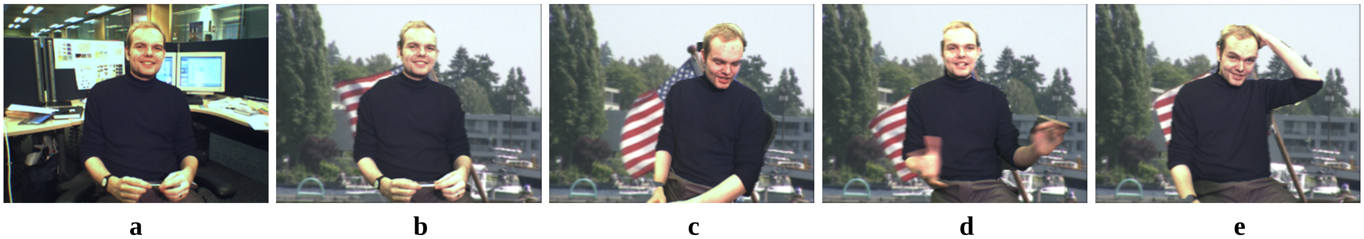


Figure 8. **Bilayer segmentation and background substitution.** (a) A frame from the MS test sequence; (b,...,e) foreground extraction and anti-aliased background substitution, over several frames. (e) The algorithm is capable of handling complex motions.

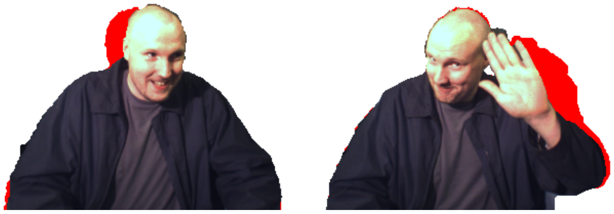


Figure 9. **Foreground extraction and occlusion detection:** two frames from the JM test sequence are shown. Pixels undergoing an F $\longrightarrow$ B transition are marked in red.

made possible by the spatio-temporal priors. Pixels transitioning from foreground to background are marked in red.

Quantitative evaluation and comparisons. Following [11] error rates are measured as a percentage of misclassified pixels, with respect to ground-truth segmentation[1]. Figure 10 presents the results for four of the six Microsoft test sequences [1]. The error rates obtained monocularly (blue) are compared to those obtained by “Layered Graph-Cut” (LGC) stereo segmentation [11]. It can be observed that while monocular segmentation cannot be expected to perform better than stereo, its accuracy is comparable with that of LGC segmentation. Figure 10 provides an objective measure of visual accuracy while videos (on our web site) offer a subjective impression that is hard to capture numerically. Despite some flicker artefacts the quality of monocular segmentation is generally convincing.

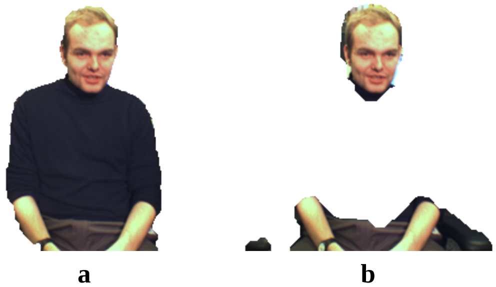


Figure 11. **The advantages of fusing motion and colour information.** (a) Segmentation of a frame of the MS sequence with the fused motion, colour/contrast and spatio-temporal information. (b) Corresponding segmentation when removing the colour likelihood. The motion cue alone produces incorrect segmentation in untextured regions. See also fig. 10c.

Figure 10 also shows that fusing colour with motion does indeed reduce error: removing the colour component U^{C} from model (7) considerably increases error rates (dotted blue lines). This effect can also be observed in fig. 11 where motion information alone produces a large (and temporally persistent) gap in the untextured region of the shirt, which is filled once colour information is added in.

Figure 12 illustrates the comparison between the motion likelihoods defined over joint spatio-temporal derivatives (section 2.5) and the more conventional likelihoods over temporal derivatives alone. The figure shows mean error rates (circles) and the associated 1-std error bars. The errors associated to spatio-temporal motion vs. stasis classification are never worse than those of temporal derivatives alone.

Finally, the contribution of the temporal model is eval-

[1] The published ground truth based on motion rather than that based on depth.

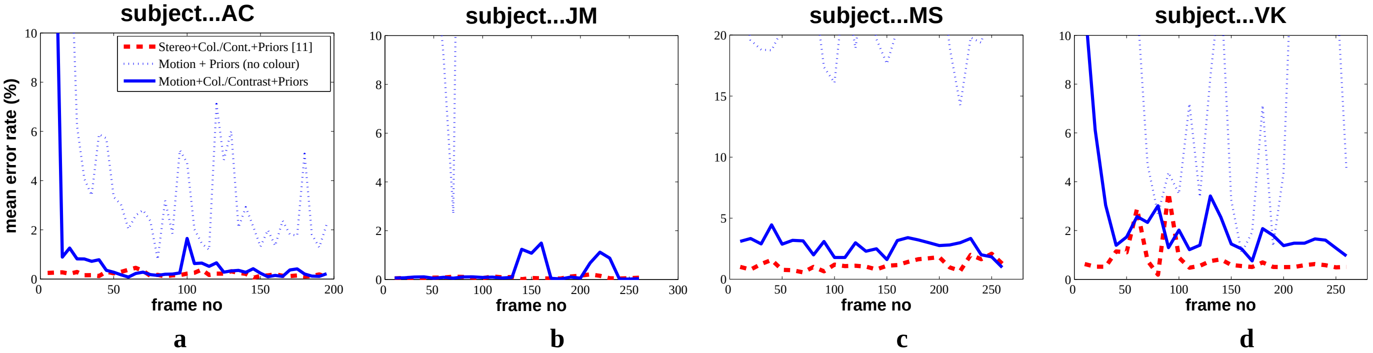


Figure 10. **Accuracy of segmentation.** (a,...,d) Error rates for the AC, JM, VK and MS sequences, respectively. The red curve (dashed) indicates the error rates obtained by LGC stereo segmentation [11]. The solid blue curve indicates the error rates obtained by the proposed monocular algorithm. For AC and VK, an initial period of stasis prevents accurate segmentation but, after a few frames the error rates drop to a value close to that of LGC stereo. After this initial stage, the model is burned in, and can tolerate periods of stasis. Omitting the colour component of the model increases error (blue dotted line).

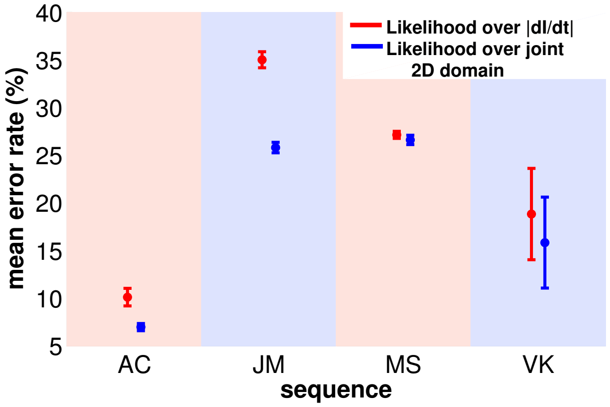


Figure 12. **Spatio-temporal derivatives perform well as features for motion vs. stasis discrimination.** Error rates are shown for a joint density over spatio-temporal derivatives, as compared with one based on temporal derivative (frame difference) alone. Spatio-temporal derivatives are never worse than simple frame difference, and clearly advantageous in two out of the four test sequences.

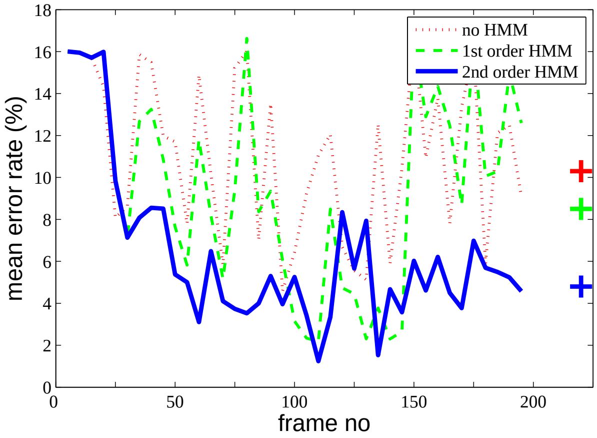


Figure 13. **The advantage of the second order temporal model.** Error plots for different orders of temporal HMM, for the AC test sequence. Crosses indicate error averaged over all frames. Averages were computed from frame 10 onwards to exclude the burn-in period. The second order model clearly achieves the lowest rate of error.

uated. Fig. 13 compares error rates for the following three cases: i) no temporal modeling, ii) first order HMM, iii) second order HMM (including both the second order temporal prior and the 2-frame motion likelihood). Error is computed for the AC test sequence with model parameters fully optimized for best performance. Colour information is omitted to avoid confounding factors in the comparison. From fig. 13 it is clear that the second order HMM model achieves the lowest error, followed by the first order model, with highest error occurring when the temporal model is entirely absent.

Robustness to photometric variations. Accurate segmentation of *all* the six test sequences in [1] has proved difficult in view of particularly large photometric variability in some sequences. The variations have been found to be due mostly to camera AGC (automatic gain control) — see the supplementary material. We found that the IJ and IU sequences exhibit illumination variation about an order of magnitude higher than in the remaining four sequences. While stereo-based segmentation is relatively immune to such problems [11], monocular algorithms are more prone to be disturbed. However, such large levels of photometric variation are easily avoided in practice by switching off the AGC facility.

Background substitution on a new test sequence. As a final demonstration, fig.14 shows the results of our background substitution technique on a TV broadcast sequence. The original weather map has been replaced with a new one. The background model is calibrated here from a single hand-segmented frame.

4. Conclusions

This paper has presented a novel algorithm for the accurate segmentation of videos by probabilistic fusion of

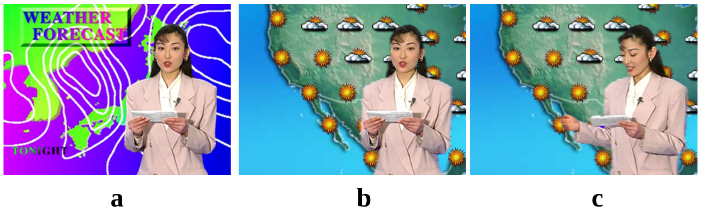


Figure 14. **Changing the weather: a final example of background substitution.** (a) A frame from the original TV sequence; (b,c) Two frames of the corresponding synthetic sequence where the original weather chart has been replaced with a different one.

motion, colour and contrast cues together with spatial and temporal priors. The model forms a Conditional Random Field, and its parameters are trained discriminatively. The motion component of the model avoids the computation of optical flow, and instead uses a novel and effective likelihood model based on spatio-temporal derivatives, and conditioned on frame-pairs. Spatio-temporal coherence is exploited via a contrast sensitive Ising energy, combined with a second order temporal Markov chain.

In terms of efficiency our algorithm compares favourably with respect to existing real-time stereo techniques [11], and achieves comparable levels of accuracy. Computationally intensive evaluation of stereo match scores is replaced by efficient motion likelihood and colour model evaluation, using efficient table look-up.

Quantitative evaluation has confirmed the validity of the proposed approach and highlighted advantages and limitations with respect to stereo-based segmentation. Finally, combining the proposed motion likelihoods and second order temporal model with stereo matching information may well, in the future, lead to greater levels of accuracy and robustness than either motion or stereo alone.

Acknowledgements. The authors acknowledge helpful discussions with C. Rother, M. Cohen and C. Zhang.

References

[1] http://research.microsoft.com/vision/cambridge/i2i/DSWeb.htm. 1, 5, 6, 7

[2] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 434–441, Santa Barbara, Jun 1998. 1

[3] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *Int. J. Computer Vision*, 12(1):43–77, 1994. 1

[4] J. Bergen, P. Burt, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(9):886–896, 1992. 1

[5] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *Proc. Europ. Conf. Computer Vision*, 2004. 4

[6] Y. Boykov and M.-P. Jollie. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proc. Int. Conf. on Computer Vision*, pages CD–ROM, 2001. 2, 3, 4

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001. 1, 5

[8] Y.-Y. Chuang, A. Agarwala, B. Curless, D. Salesin, and R. Szeliski. Video matting of complex scenes. In *Proc. Conf. Computer graphics and interactive techniques*, pages 243–248. ACM Press, 2002. 1

[9] Y.-Y. Chuang, B. Curless, D. Salesin, and R. Szeliski. A Bayesian approach to digital matting. In *Proc. Conf. Computer Vision and Pattern Recognition*, CD–ROM, 2001. 1

[10] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *Proc. Conf. Computer Vision and Pattern Recog.*, 2001. 1

[11] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Bi-layer segmentation of binocular stereo video. In *Proc. Conf. Comp. Vision Pattern Rec.*, San Diego, CA, Jun 2005. 2, 4, 5, 6, 7, 8

[12] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 26, February 2004. 1

[13] P. Kumar, P. Torr, and A. Zisserman. Learning layered motion segmentations of video. In *Proc. Int. Conf. Computer Vision*, Beijing, China, oct 2005. 1

[14] S. Kumar and M. . Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *IEEE International Conference on Computer Vision (ICCV)*, Nice, France, 2003. 2

[15] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001. 2

[16] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004. 2, 3, 4, 5

[17] S. Rowe and A. Blake. Statistical mosaics for tracking. *J. Image and Vision Computing*, 14:549–564, 1996. 4

[18] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. Conf. Comp. Vision Pattern Rec.*, Fort Collins, CO, jun 1999. 4

[19] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum. Background cut. In *Proc. Europ. Conf. Computer Vision*, Graz, Austria, 2006. 4

[20] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated bayesian approach to layer extraction from image sequences. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(3):297–303, March 2001. 1

[21] J. Y. A. Wang and E. H. Adelson. Layered representation for motion analysis. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 361–366, New York, Jun 1993. 1

[22] J. Wills, S. Agarwal, and S. Belongie. What went where. In *Proc. Conf. Comp. Vision Pattern Rec.*, pp. I:37–40, 2003. 1

[23] J. Xiao and M. Shah. Motion layer extraction in the presence of occlusion using graph cut. In *Proc. Conf. Comp. Vision Pattern Rec.*, 2004. 1