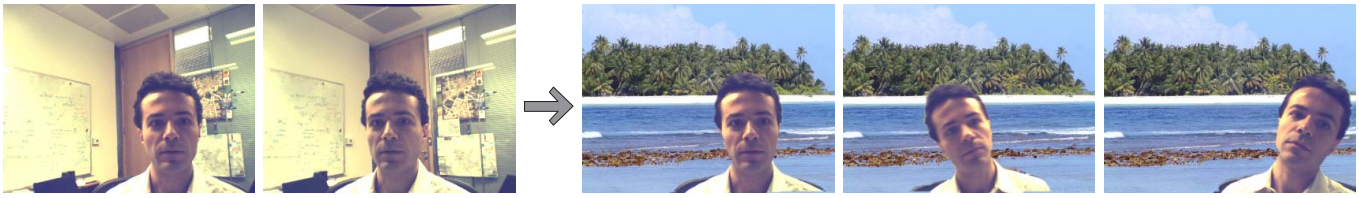
Bi-layer segmentation of binocular stereo video

V. Kolmogorov A. Criminisi A. Blake G. Cross C. Rother

Microsoft Research Ltd., 7 J J Thomson Ave, Cambridge, CB3 0FB, UK

http://research.microsoft.com/vision/cambridge



input left view *input right view* *automatic layer separation and background substitution in three different frames*

Figure 1: An example of automatic foreground/background separation in binocular stereo sequences. The extracted foreground sequence can be composited free of aliasing with different static or moving backgrounds; a useful tool in video-conferencing applications. Note, the input synchronized stereo sequences used throughout this paper can be downloaded from [1], as well as hand-labeled segmentations.

Abstract

This paper describes two algorithms capable of real-time segmentation of foreground from background layers in stereo video sequences. Automatic separation of layers from colour/contrast or from stereo alone is known to be error-prone. Here, colour, contrast and stereo matching information are fused to infer layers accurately and efficiently. The first algorithm, Layered Dynamic Programming *(LDP), solves stereo in an extended 6-state space that represents both foreground/background layers and occluded regions. The stereo-match likelihood is then fused with a contrast-sensitive colour model that is learned on the fly, and stereo disparities are obtained by dynamic programming. The second algorithm,* Layered Graph Cut *(LGC), does not directly solve stereo. Instead the stereo match likelihood is marginalised over foreground and background hypotheses, and fused with a contrast-sensitive colour model like the one used in LDP. Segmentation is solved efficiently by ternary graph cut.*

Both algorithms are evaluated with respect to ground truth data and found to have similar perfomance, substantially better than stereo or colour/contrast alone. However, their characteristics with respect to computational efficiency are rather different. The algorithms are demonstrated in the application of background substitution and shown to give good quality composite video output.

1. Introduction

This paper addresses the problem of separating a foreground layer from stereo video in real time. A prime application is for teleconferencing in which the use of a stereo webcam already makes possible various transformations of the video stream including digital pan/zoom/tilt and object insertion[1]. Here we concentrate on providing the infrastructure for live background substitution. This demands foreground layer separation to near Computer Graphics quality, including α-channel determination as in video-matting [9], but with computational efficiency sufficient to attain live streaming speed.

Layer extraction from images has long been an active area of research [6, 4, 18, 24, 25]. The challenge addressed here is to segment the foreground layer both accurately and efficiently. Conventional stereo algorithms e.g. [19, 10] have proven competent at computing depth. Stereo occlusion is a further cue that needs to be accurately computed [15, 5, 17, 11] to achieve good layer extraction. However, the strength of stereo cues degrade over low-texture regions such as blank walls, sky or saturated image areas. Recently interactive colour/contrast-based segmentation techniques have been demonstrated to be very effective [7, 20], even in the absence of texture. Segmentation based on colour/contrast alone is nonetheless beyond the capability of fully automatic methods. This suggests a robust approach that exploits fusion of a variety of cues. Here we propose a model and algorithms for fusion of stereo with colour and contrast, and a prior for intra-layer spatial coherence.

The efficiency requirements of live background substitution have restricted us to algorithms that are known to be capable of near frame-rate operation, specifically dynamic programming and ternary graph cut (i.e. α-expansion algorithm [8] with three labels). Therefore two approaches to segmentation are proposed here: Layered Dynamic Pro-

[1] research.microsoft.com/vision/cambridge/i2i

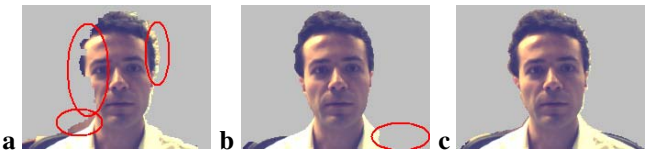


Figure 2: **Segmentation by fusing colour, contrast and stereo.** Results of three different segmentation algorithms run on the left input image of fig. 1 (see [2] or video in the CD-ROM proceedings for more examples). (**a**) Stereo-based segmentation. (**b**) Colour/contrast-based segmentation. (**c**) The algorithm proposed here, by fusing colour, contrast and stereo achieves more accurate segmentation. The foreground artefacts visible in (a) and (b) (marked in red) are corrected in (c), where the person and chair are correctly extracted. Note, we do not just combine images (a) and (b) to produce (c); see text for algorithmic details.

gramming (LDP) and Layered Graph Cut (LGC). Each works by *fusing* likelihoods for stereo-matching, colour and contrast to achieve segmentation quality unnattainable from either stereo or colour/contrast on their own (see fig. 2). This claim is verified by evaluation on stereo videos with respect to ground truth (section 5). Finally, efficient post-processing for matting [13] is applied to obtain good video quality as illustrated in stills and accompanying video in the CD-ROM proceedings.

The paper is organised as follows. In section 2 we describe components of our probabilistic model that are common in both techniques. In sections 3 and 4 we present LDP and LGC algorithms, respectively. Experimental results are given in section 5. Finally, section 6 contains conclusions. Note that due to space limitations some details of the algorithms have been omitted, but can be found in [16].

2. Probabilistic models for bi-layer segmentation of stereo images

First we outline the probabilistic structure of the stereo and colour/contrast models.

2.1 Notation and basic framework

Pixels in the left and right images are m, n respectively and index either the entire images, or just a pair of matching epipolar lines, as required. Over epipolar lines, the intensity functions from left and right images are

$$\mathbf{L} = \{L_m,\ m = 0, \ldots, N\},\ \mathbf{R} = \{R_n,\ n = 0, \ldots, N\}.$$

Stereo disparity along the cyclopean[2] epipolar line is $\mathbf{d} = \{d_k,\ k = 0, \ldots, 2N\}$ and disparity is simply related to image coordinates:

$$d_k = m - n \ \text{ with } \ m = \frac{(k + d_k)}{2} \ \text{ and } \ n = \frac{(k - d_k)}{2}. \tag{1}$$

[2] *cyclopean* here means mid-way between left and right input cameras.

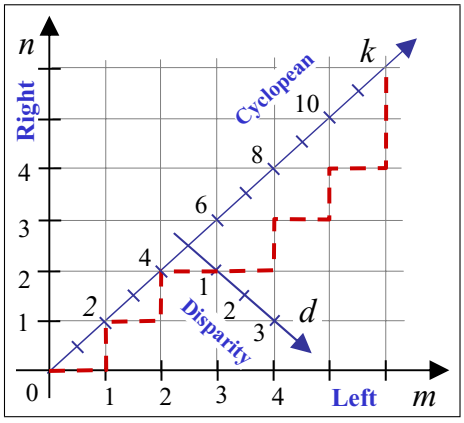


Figure 3: **Disparity and the cyclopean image.** Notation conventions for left and right epipolar lines with pixel coordinates m, n, cyclopean coordinates k and stereo disparity $d = m - n$. Possible matching path shown dashed (*cf.* [5, 10]).

Left and right pixels are ordered by any particular matching path (fig. 3) to give $2N$ cyclopean pixels

$$\mathbf{z} = \{z_k,\ k = 0, \ldots, 2N\},$$

where $k = m + n$. Only single-step horizontal and vertical moves are allowed — no diagonal or multistep moves. This means that, for a given path, $\mathbf{z}$ consists of a sequence of L_m and R_n elements, such that each left and right pixel appears exactly once on the path. This is essential to a consistent probabilistic interpretation, as explained shortly. In addition an array $\mathbf{x}$ of state variables, either in cyclopean coordinates $\mathbf{x} = \{x_k\}$ or image coordinates $\mathbf{x} = \{x_m\}$, takes values $x_k \in \{\mathrm{F}, \mathrm{B}, \mathrm{O}\}$ according to whether the pixels is a foreground match, a background match or occluded.

Sets of model parameters: Φ are defined for priors on stereo; Θ for colour/contrast and match likelihoods. Details are given later. This enables Gibbs energies to be defined, in terms of probabilistic models, which are globally minimised to obtain a segmentation. The LDP algorithm minimises, independently over each epipolar line, an energy $E(\mathbf{z}, \mathbf{d}, \mathbf{x}; \Theta, \Phi)$ in which there is explicit dependency on disparity $\mathbf{d}$. The presence of parameters Φ indicates that the LDP energy incorporates priors on stereo disparity as a further constraint on the solution for segmentation. Conversely LGC minimises, globally over an image, an energy $E(\mathbf{z}, \mathbf{x}; \Theta)$ in which disparity variables do not explicitly appear.

2.2. Likelihood for stereo

We need to model the stereo-matching likelihood function $p(\mathbf{z} \mid \mathbf{x}, \mathbf{d})$ and this is expanded as

$$\begin{aligned} p(\mathbf{z} \mid \mathbf{x}, \mathbf{d}) &= \prod_k p(z_k \mid x_k, d_k, z_1, \ldots, z_{k-1}) \\ &= K(\mathbf{z}) \prod_k \exp -\mathcal{L}_k(x_k, d_k) \end{aligned} \tag{2}$$

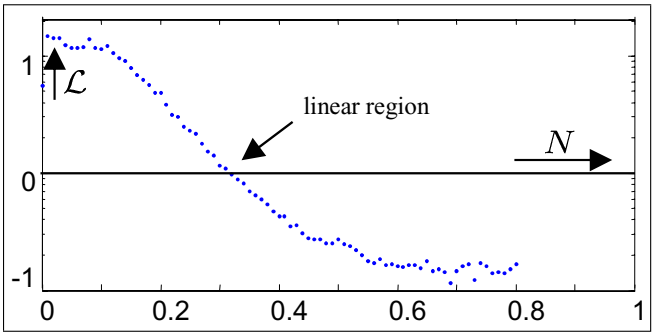


Figure 4: **Likelihood model:** the empirical log-likelihood ratio $-\mathcal{L}_k$ is shown for stereo matches, plotted here as a function of the NSSD measure $N(L^{\mathrm{P}},\ R^{\mathrm{P}})$, using the ground truth stereo data "Teddy" from the Middlebury set [3]. Note the linearity in the region of $\mathcal{L} = 0$, where most data falls. Similar behaviour has been observed for other ground-truth datasets.

where the pixelwise negative log-likelihood *ratio*, for match vs. non-match, is

$$\begin{aligned}\mathcal{L}_k(x_k, d_k) &= -\log p(z_k \mid x_k, d_k, z_1, \ldots, z_{k-1}) \\ &+ \log p(z_k \mid x_k = \mathrm{O}). \end{aligned} \tag{3}$$

According to the definition, $\mathcal{L}_k(x_k = \mathrm{O}, d_k) = 0$. Commonly [22] stereo matches are scored using SSD (sum-squared difference), that is L^2-norm of difference between image patches $L^{\mathrm{P}}_m,\ R^{\mathrm{P}}_n$ surrounding hypothetically matching pixels m, n. Like [11, 12] we model $\mathcal{L}_k$ using SSD with additive and multiplicative normalisation for robustness to non-Lambertian effects (NSSD - normalized SSD):

$$\mathcal{L}_k(x_k, d_k) = \begin{cases} M(L^{\mathrm{P}}_m,\ R^{\mathrm{P}}_n) - M_0 & \text{if } x_k \in \{\mathrm{F}, \mathrm{B}\} \\ 0 & \text{if } x_k = \mathrm{O}, \end{cases} \tag{4}$$

where $M = \lambda N$ with λ a constant, and

$$N(L^{\mathrm{P}},\ R^{\mathrm{P}}) = \frac{1}{2} \frac{\|L^{\mathrm{P}} - R^{\mathrm{P}}\|^2}{\|L^{\mathrm{P}} - \overline{L^{\mathrm{P}}}\|^2 + \|R^{\mathrm{P}} - \overline{R^{\mathrm{P}}}\|^2} \in [0, 1]. \tag{5}$$

This model has been tested against the Middlebury datasets [3] and found to be reasonable — an example of results is given in fig. 4. Such analysis gives also useful working values for λ (typical value for monochrome images is $\lambda = 10$, which holds for a variety of patch sizes; we used 3×7 patches for LGC and 5×5 patches for LGC). For M_0 this analysis yields value of approximately 0.3. However, we found that discriminatively learned M_0 is usually larger: a typical value is $M_0 = 0.4$, and that value gives better error rates in practice.

2.3 Stepwise restriction for matching paths

Previous algorithms e.g. [10, 14] have allowed multiple and/or diagonal moves on the stereo matching paths. However, the single-step restriction (fig. 3) allows for a consistent probabilistic interpretation of the sequence matching problem to exist (see [16] for details). With the restriction in place, each element L_m and R_n is "explained" once and only once: it appears once and only once as z_k in the $p(z_k|\ldots)$ term of (2), as required. The existence of a probabilistic interpretation then allows a consistent account of fusion of different modalities, by multiplication of likelihoods. The practical benefit is that the weighting coefficients of the various energy terms are mostly determined automatically, from statistics, rather than having to be set by hand.

2.4. Likelihood for colour

Following previous approaches to two-layer segmentation [7, 20] we model likelihoods for colour in foreground and background using Gaussian mixtures in RGB colour space, learned from image frames labelled (automatically) from earlier in the sequence. In addition, the background model is enhanced by mixing in a probability density learned, for each pixel, by pixelwise background maintenance [21, 23].

The foreground colour model $p(z \mid x = \mathrm{F})$ is simply a spatially global Gaussian mixture learned from foreground pixels. In the background there is a similar learned Gaussian mixture $p(z \mid x = \mathrm{B})$ and also a per-pixel single Gaussian density $p_k(z_k)$ available wherever the stability flag $s_k \in \{0, 1\}$ indicates that there has been stasis over a sufficient number of previous frames. The occluding state $x = \mathrm{O}$ refers to background pixels and therefore shares a colour model with $x = \mathrm{B}$. The combined colour model is then given by an energy U^C_k:

$$U^C_k(z_k, x_k) = -\log p(z_k \mid x_k) \ \text{ if } \ x = \mathrm{F}, \tag{6}$$

and for $x = \mathrm{B}, \mathrm{O}$:

$$U^C_k(z_k, x_k) = -\log \left[(1 - \frac{s_k}{2}) p(z_k | x_k = \mathrm{B}) + \frac{s_k}{2} p_k(z_k) \right] \tag{7}$$

a mixture between the global background model and the pixelwise one. This approach is both powerful and robust: pixelwise densities U^C_k are typically strongly peaked, and hence very informative, but sensitive to movement in the background. That sensitivity is robustified by adding in the general background distribution $p(z_k|x_k = \mathrm{B})$ as the contamination component in the mixture.

2.5 Contrast model

There is a natural tendency for segmentation boundaries to align with contours of high image contrast. Similarly to [7], this is represented by an image energy of the form

$$V_{k,k'} = F_{k,k'}[x_k, x_{k'}] V^*(z_k, z_{k'}), \tag{8}$$

where k, k' are neighbouring pixel-pairs in the cyclopean image. Function $F_{k,k'}[x_k, x_{k'}]$ is the potential coefficient which implements geometric constraints (it must be

anisotropic because of epipolarity). The exact form of $F_{k,k'}$ is different for LDP and LGC, and it is given later in corresponding sections. The term V^* applies contrast sensitivity:

$$V^*(z,z') = \frac{1}{1+\epsilon}\left(\epsilon + \exp -\frac{\|z-z'\|^2}{2\sigma^2}\right) \qquad (9)$$

with $\sigma^2 = \left\langle \|z-z'\|^2 \right\rangle$, a mean over all pairs of neighbours in the left and right images.

The energy made by summing up $V_{k,k'}$ in fact represents an Ising prior for labelling coherence, modified by a contrast factor that acts to discount partially the coherence terms. The constant ϵ is a "dilution" constant for contrast, previously [7] set to $\epsilon = 0$ for pure colour segmentation. Here, $\epsilon = 1$ is more appropriate — diluting the influence of contrast in recognition of the increased diversity of segmentation cues.

3. Layered Dynamic Programming (LDP)

The LDP algorithm solves for disparity over individual scanlines on the (virtual) cyclopean image z_k. It is based on the classic dynamic programming approach [10, 19] together with augmentation of the state space to handle occlusion [11, 12]. The 4-state model of [12] is described in section 3.1. The foreground/background states are then added in the 6-state model (section 3.2).

3.1 4-state stereo with occlusions

This can be expressed concisely as a 4-state system that is summarised in fig. 5. A basic 4-state system is annotated with transitions and associated energy terms to define a global energy

$$E(\mathbf{z},\mathbf{d},\mathbf{x};\Theta,\Phi) = \sum_k E_k(d_k, d_{k-1}, x_k, x_{k-1}) \qquad (10)$$

where $x_k \in \{\mathrm{M},\mathrm{O}\}$ in which M denotes a stereo match and O an occlusion. Each $E_k(\ldots)$ term consists of the sum

$$E_k = U_k^{\mathrm{M}} + V_{k-1,k} \qquad (11)$$

of a cost $V_{k-1,k}$ of transition $k-1 \rightarrow k$ (on arcs) and a state cost U_k^{M} (inside nodes) on the diagram of fig. 5. The occluding state $x_k = \mathrm{O}$ is split into two sub-states (red circles in fig. 5), left-occluding and right-occluding (which do not intercommunicate, reflecting geometric constraints). The matching state $x_k = \mathrm{M}$ also has two substates (green circles in fig. 5):

$$\begin{array}{rll} \text{Left match} & \text{if} & d_k = d_{k-1}+1 \\ \text{Right match} & \text{if} & d_k = d_{k-1}-1 \end{array} \qquad (12)$$

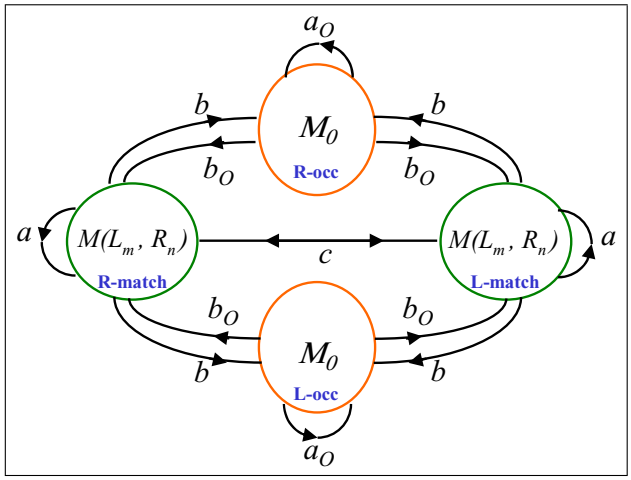


Figure 5: **State space for stereo matching with occlusion.** Matched and occluded states (each in left and right versions) form a 4-state system. Successive pixels along a cyclopean epipolar line (fig. 3) incur a cost increment (e.g. b) for the arc $k-1 \rightarrow k$ traversed, plus an increment (e.g. M_0) for the new node k.

representing the typical stepwise progress of the matching path as in figure 3. There are a total then of 4 possible states: $x_k \in \{\text{L-match, R-match, L-occ, R-occ}\}$.

The model has a number of parameters $\Phi = \{a_0, b_0, a, b, c\}$ which specify the stereo prior over matching paths. It might seem problematic that so many parameters need to be set, but in fact they can be learned from previous labelled frames as follows:

$$b_0 = \log(2W_{\mathrm{O}}) \;\; b = \log(2W_{\mathrm{M}}) \qquad (13)$$

where W_{M} and W_{O} are the mean widths of matched and occlusion regions respectively. This follows simply from the fact that $2\exp -b$ is the probability of escape from a matched state, and similarly for $2\exp -b_0$ in an occluded state. Then consideration of viewing geometry (details omitted) indicates:

$$a = \log(1 + D/b) - \log(1 - 1/W_{\mathrm{M}}), \qquad (14)$$

where D is a nominal distance to objects in the scene and b is the interocular distance (camera baseline). Lastly, probabilistic normalisation demands that

$$c = -\log(1 - 2e^{-b} - e^{-a}) \;\text{ and }\; a_0 = -\log(1 - 2e^{-b_0}),$$

so there are really just 3 independent parameters in Φ. Match costs inside nodes are defined in terms of match likelihood energy, as in (4). The total energy is then minimised by Dynamic Programming in a manner similar to [11].

3.2 6-state stereo with occlusion and layers

Next, we distinguish foreground and background layers and use an extended 6-state algorithm in which matched states

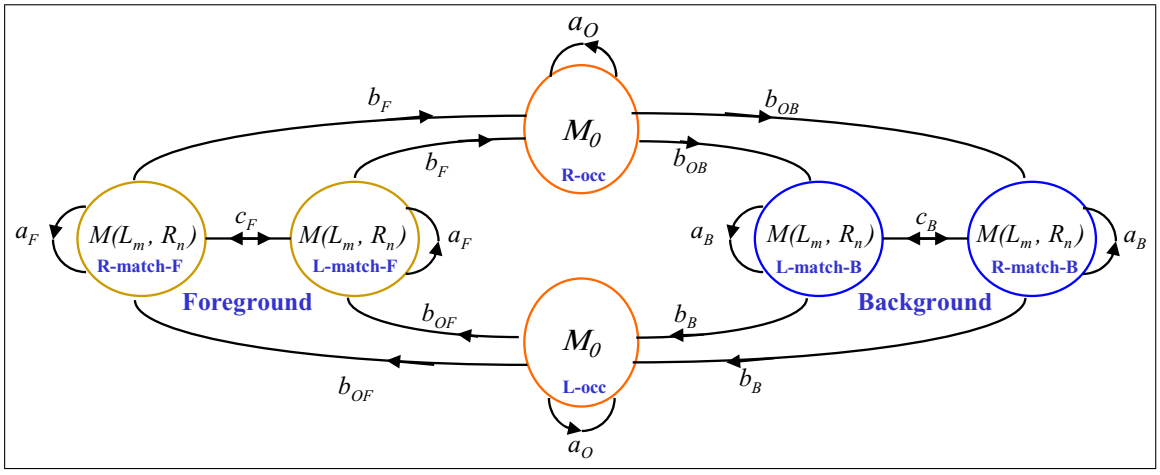


Figure 6: **Extended state space for LDP** in which the matched state of fig. 5 is split into a foregound and a background substate. Note that from the foreground state (yellow circles), only the *right* occluding state is accessible, and from background (blue circles) only the *left* occluding state, constraints of the geometry of occlusion.

from the 4-state system are split into foreground and background substates. The diagram of fig. 5 is cut by the splitting of the matched states and unfolded into the diagram of fig. 6. There are now a total of 6 possible states: $x_k \in \{$L-match-F, R-match-F, L-match-B, R-match-B, L-occ, R-occ$\}$. The model has a number of parameters $\Phi = \{a_F, a_B, a_O, b_F, b_B, b_{OF}, b_{OB}, c_F, c_B\}$ all of which can be set from statistics and geometry as before, but now statistics are collected separately for the $x_k = \mathrm{F}$ and $x_k = \mathrm{B}$ conditions.

3.3 The 6-state model with disparity-pull and colour/contrast fusion

Now the stereo infrastructure for LDP is capable of representing the two layers, it remains to add in energies for the colour and contrast likelihoods. The full energy for stereo matching, per cyclopean pixel, is now

$$E_k = U_k^{\mathrm{M}} + V_{k-1,k} + U_k^{\mathrm{C}} + U_k^{\mathrm{D}} \tag{15}$$

where U_k^{M} and $V_{k-1,k}$ are respectively the node and transition energies from section 3.2. The nodal energy is now extended, from U_k^{M} to $U_k^{\mathrm{M}}+U_k^{\mathrm{C}}+U_k^{\mathrm{D}}$, to take account of additional colour and "disparity-pull" information, respectively. The colour energy term U_k^{C} is as described earlier (6). The disparity-pull energy

$$U_k^{\mathrm{D}}(z_k, x_k) = -\log p(d_k|x_k) \tag{16}$$

represents the pull of each layer towards certain disparities, as determined by the densities $p(d_k|x_k = \mathrm{F}, \mathrm{B}, \mathrm{O})$ which are learned as Gaussians from labelled data in previous frames. Typically this term pulls the foreground/background layers towards larger/smaller values of disparity respectively.

Finally, the transition component $V_{k-1,k}$ from the 6-state model is further modified to take account of contrast (8). This is done by modifying each transition energy between occluding and foreground states (fig. 6) as follows:

$$b_F \rightarrow b_F\, V^*(z_{k-1}, z_k) \;\text{ and }\; b_{OF} \rightarrow b_{OF}\, V^*(z_{k-1}, z_k), \tag{17}$$

where V^* is the contrast term defined earlier (9). Note that colour/contrast in the 6-state model have to be computed jointly over left and right images (see [16] for details).

Now the full 6-state system, augmented both for bi-layer inference and for fusion of colour/contrast with stereo can be optimised by dynamic programming as before. Results of this approach are shown below in section 5, but in the meantime the alternative LGC algorithm is described.

4. Layered Graph Cut (LGC)

Layered Graph Cut (LGC) determines segmentation $\mathbf{x}$ as the minimum of an energy function $E(\mathbf{z}, \mathbf{x}; \Theta)$, in which, unlike LDP, stereo disparity $\mathbf{d}$ does not appear explicitly. Instead, disparity is marginalised to give a likelihood $p(\mathbf{L} \mid \mathbf{x}, \mathbf{R})$, in which stereo-match likelihoods have been aggregated to compute support for each of the three labels in $\mathbf{x}$: foreground, background and occlusion $(\mathrm{F}, \mathrm{B}, \mathrm{O})$. The segmentation is ternary so the α-expansion form of graph-cut [8] is needed. Space forbids a detailed description of the LGC algorithm, however, it represents another, very effective way of implementing the colour-stereo fusion idea. Therefore, it was felt important to include a sketch of the method. A particular difference between LDP and LGC is that LGC is specified with respect to one (*e.g.* left) image, rather than the cyclopean frame as in LDP.

The energy function for LGC is composed of three

terms:

$$E(\mathbf{z},\mathbf{x};\Theta) = U^{\mathrm{C}}(\mathbf{z},\mathbf{x};\Theta) + V(\mathbf{z},\mathbf{x};\Theta) + U^{\mathrm{S}}(\mathbf{z},\mathbf{x}), \quad (18)$$

representing energies for colour-likelihood, spatial coherence/contrast and stereo likelihood respectively. The colour energy is simply a sum over pixels in the left image

$$U^{\mathrm{C}}(\mathbf{z},\mathbf{x};\Theta) = \sum_m U^{\mathrm{C}}_m(L_m, x_m) \quad (19)$$

of the pixelwise colour energy defined earlier (6). The coherence/contrast energy is a sum of pairwise energies of the form (8) where coefficient $F_{m,m'}$ is defined as follows. For vertical and diagonal cliques it acts as a switch active across a transition in or out of the foreground state: $F_{m,m'}[x,x'] = \gamma$ if exactly one variable x, x' equals F, and $F_{m,m'}[x,x'] = 0$ otherwise. For horizontal lines it implements geometric constraints: $F_{m,m'}[x,x']$ is infinity for transitions O→B and F→O, and zero for all other transitions.

4.1 Marginalisation of stereo likelihood

The remaining term in (18) is $U^{\mathrm{S}}(\mathbf{z},\mathbf{x})$ which captures the influence of stereo matching likelihood on the probability of a particular segmentation. It is defined to be

$$U^{\mathrm{S}}(\mathbf{z},\mathbf{x}) = \sum_m U^{\mathrm{S}}_m(x_m) \quad (20)$$

$$\text{where } U^{\mathrm{S}}_m(x_m) = -\log p(L_m|x_m,\mathbf{R}) + \text{const}, \quad (21)$$

$$p(L_m|x_m,\mathbf{R}) = \sum_d p(L_m|x_m,d_m=d,\mathbf{R}) p(d_m=d|x_m) \quad (22)$$

— marginalizing over disparity, and the distributions $p(d_m = d|x_m)$ for $x_m \in \{\mathrm{F},\mathrm{B}\}$ are learned from labelled data in previous frames. The const term in (21) allows us to make use of the likelihood-ratio model of section 2.2 for stereo matches, giving

$$U^{\mathrm{S}}_m(x_m) = \quad (23)$$
$$-\log\left[\textstyle\sum_d p(d_m=d|x_m)\exp -\lambda M(L^{\mathrm{P}}_m,\ R^{\mathrm{P}}_n)\right] - M_0.$$

Results of LDP and LGC are given next.

5. Results

Performance of the LGC and LDP algorithms was evaluated with respect to ground-truth segmentations of every fifth frame (left view) in each of two test stereo sequences[3]. The data was labelled manually, labelling each pixel as background, foreground or unknown. The unknown label was used to mark mixed pixels occurring along layer boundaries. Error is then measured as percentage of misclassified pixels, ignoring "unknown" pixels.

[3]Ground truth segmentation data is available at [1].

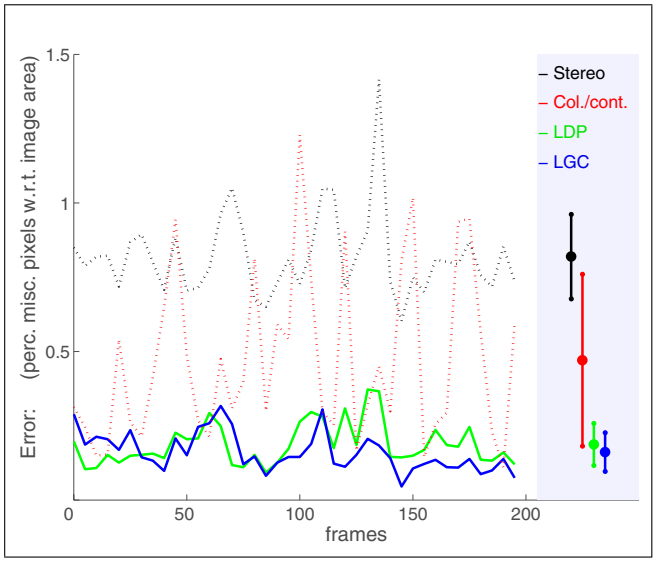


Figure 7: **Measuring segmentation performance.** Segmentation error (percentage of misclassified pixels) is computed on the `S1` sequence, with respect to ground truth. Average error values and 1-std bars are also plotted. Note that fused stereo and colour/contrast (LGC and LDP) perform substantially better than either stereo or colour/contrast alone.

Measuring accuracy of segmentation. Segmentation performance for the stereo sequence pair `S1` (example input images in fig.1) is compared for colour/contrast, for stereo alone, and for colour/contrast and stereo fused together (fig. 7). The colour/contrast algorithm here is simply LGC in which the stereo component is switched off. The stereo-only algorithm is 4-state DP as in section 3.1. Fusion of colour/contrast and stereo by the LGC and LDP algorithms both show similarly enhanced performance compared with colour/contrast or stereo alone. As a test of robustness, the algorithms have also been tested on a sequence `S2` with motion in the background (example input images in fig. 12). Two people enter the scene and move around behind a person occupying the foreground. Once again the power of fusing colour/contrast and stereo is immediately apparent (fig. 8). An example of a segmented image is shown in fig. 9 and the spatial distribution of segmentation errors is illustrated in fig. 10: errors tend to cluster closely around object boundaries.

Background substitution in sequences. Finally, figs. 11-13 demonstrate the application of segmentation to background replacement in video sequences (additional results are available at [2]). Background substitution in sequences is challenging as the human eye is very sensitive to flicker artefacts. Following foreground/background segmentation, α-matting has been computed using SPS [13] as a post-process. Both the LGC and LDP algorithms give good results, with blended boundaries and little visible flicker.

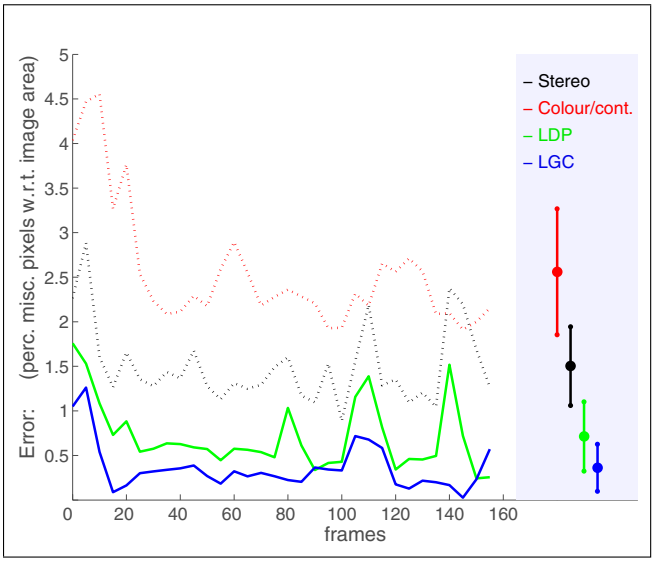
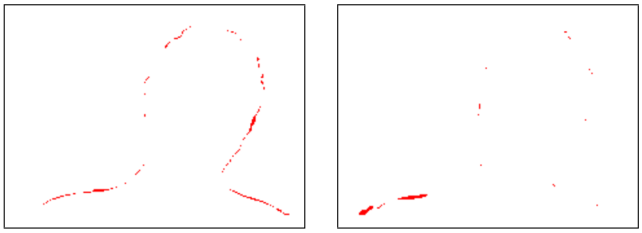


Figure 8: **Segmentation performance is robust to background motion.** As for fig. 7 but for the `S2` sequence: fusion by LDP or LGC is robust to movement in the background.

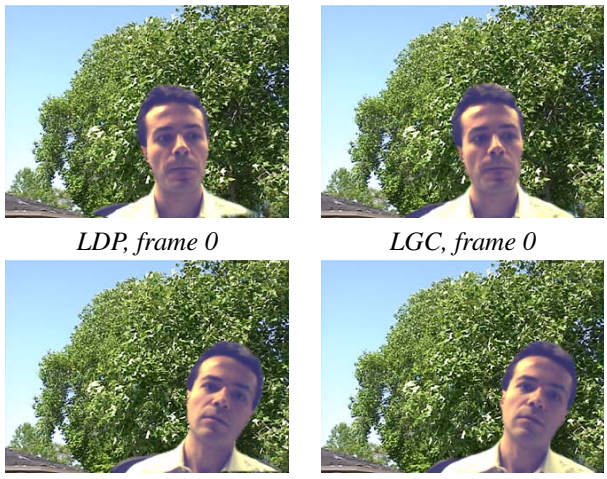
LDP *LGC*

Figure 9: **Extracted foreground layer** for the left view of `S1`, frame 100.



LDP *LGC*

Figure 10: **Spatial distribution of segmentation error.** Red pixels are misclassified (with respect to ground-truth). Results for `S1` at frame 100.



LDP, frame 0 *LGC, frame 0*

LDP, frame 100 *LGC, frame 100*

Figure 11: **Segmentation and background substitution.** Here we show background substitution for two frames of the `S1` sequence. Visual quality of LDP and LGC results are similar.

6. Conclusion

This paper has addressed the important problem of segmenting stereo sequences. Disparity-based segmentation and colour/contrast-based segmentation alone are prone to failure. LDP and LGC are algorithms capable of fusing the two kinds of information with a substantial consequent improvement in segmentation accuracy. Moreover, both algorithms are suited for real-time implementation. Fast implementations of DP techniques are well known [10, 11]. Ternary graph cut has been applied, in our laboratory, at around 10 frames per second for 320×240 image on a 3GHz Pentium desktop machine. Given that the segmentation accuracies of LDP and LGC are comparable, what is to choose between them? In fact the choice may depend on architecture: the stereo component of LGC can be done, in principle on a graphics co-processor, including the marginalisation over disparities. In LDP however, although stereo-match scores could be computed with the graphics coprocessor, communicating the entire cost array $\mathcal{L}_k(x_k, d_k)$ to the general processor is beyond the bandwidth limitations of current GPU designs. On the other hand LDP is economical in memory usage, in that it can proceed scanline by scanline.

In conclusion, we have demonstrated properties of the LDP and LGC algorithms and underlying model as follows.

- Fusion of stereo with colour and contrast can be captured in a probabilistic model, in which parameters can mostly be learned, or are otherwise stable.
- Fusion of stereo with colour and contrast makes for more powerful segmentation than for stereo or colour/contrast alone.
- Good quality segmentation of temporal sequences (stereo) can be achieved, without imposing any explicit temporal consistency between neighbouring frames.

Acknowledgements

We thank M. Isard and R. Szeliski for helpful discussions.

Figure 12: **Segmentation with non-stationary background.** (**Left**) Three frames of the input left sequence `S2` (right frame not shown here). (**Right**) Corresponding LGC segmentation and background substitution. Note the robustness of the segmentation to motion in the original background.

Figure 13: **Non-stationary background with more complex foreground.** A final example of segmentation and background substitution (test sequence `S3`). (**Left**) Input left images. A third person is moving in the original background. (**Right**) LGC background-substitution.

References

[1] http://research.microsoft.com/vision/cambridge/i2i/DSWeb.htm

[2] http://research.microsoft.com/vision/cambridge/i2i/bgsubstitution.htm

[3] http://cat.middlebury.edu/stereo/

[4] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *Proc. CVPR*, pages 434–441, Santa Barbara, 1998.

[5] P. N. Belhumeur. A Bayesian-approach to binocular stereopsis. *Int. J. Computer Vision*, 19(3):237–260, August 1996.

[6] J.R. Bergen, P.J. 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.

[7] Y. Boykov and M-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proc. Int. Conf. on Computer Vision*, 2001.

[8] Y. Boykov, O. Veksler and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on PAMI*, 23(11), 2001.

[9] Y-Y Chuang, A. Agarwala, B. Curless, D.H. Salesin, and R. Szeliski. Video matting of complex scenes. In *Proc. ACM Siggraph*, 2004.

[10] I.J. Cox, S.L. Hingorani, and S.B. Rao. A maximum likelihood stereo algorithm. *CVIU*, 63(3):542–567, 1996.

[11] A. Criminisi, J. Shotton, A. Blake, and P.H.S. Torr. Gaze manipulation for one to one teleconferencing. In *Proc. ICCV*, 2003.

[12] A. Criminisi, J. Shotton, A. Blake, C. Rother, and P.H.S. Torr. Efficient Dense-Stereo and Novel-view Synthesis for Gaze Manipulation in One-to-one Teleconferencing. Technical Report MSR-TR-2003-59, Microsoft, 2003.

[13] A. Criminisi and A. Blake. The SPS algorithm: Patching figural continuity and transparency by Split-Patch Search. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 721–728, 2004.

[14] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge University Press, 1998.

[15] D. Geiger, B. Ladendorf, and A.Yuille. Occlusions and binocular stereo. *Int. J. Computer Vision*, 14:211–226, 1995.

[16] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Probabilistic fusion of stereo with color and contrast for video segmentation. Technical Report MSR-TR-2005-36, Microsoft, 2005.

[17] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proc. ECCV*, Copenhagen, Denmark, May 2002.

[18] N. Jojic, and B. Frey. Learning flexible sprites in video layers. In *Proc. CVPR*, Hawaii, 2001.

[19] Y. Ohta and T. Kanade. Stereo by intra- and inter-scan line search using dynamic programming. *IEEE Trans. on PAMI*, 7(2), 1985.

[20] C. Rother, V. Kolmogorov, and A. Blake. GrabCut – Interactive foreground extraction using iterated graph cuts. In *Proc. Siggraph*, 2004.

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

[22] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3), 2002.

[23] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. CVPR*, pages 246–252, 1999.

[24] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated Baysesian approach to layer extraction from image sequences. *PAMI*, 2001.

[25] J. Y. A. Wang and E. H. Adelson. Layered representation for motion analysis. In *Proc. CVPR*, pages 361–366, New York, Jun 1993.