Probabilistic fusion of stereo with color and contrast for bi-layer segmentation

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

Microsoft Research TR–2005–35

ABSTRACT

This paper describes two algorithms for the 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 disparities to evaluate foreground and background hypotheses, and then 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 either 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.

I. INTRODUCTION

This paper addresses the problem of separating a foreground layer, from stereo video, as in figure 1, in real time. A prime application is for teleconferencing in which the use of a stereo webcam already makes possible various

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

Fig. 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. Stereo sequence AC used here. Note: the input synchronized stereo sequences used throughout this paper can be downloaded from [1], together with hand-labeled segmentations.

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 [12], but with computational efficiency sufficient to attain live streaming speed.

Layer extraction from images has long been an active area of research [6], [4], [22], [31], [33]. The challenge addressed here is to segment the foreground layer both accurately and efficiently. Conventional stereo algorithms e.g. [25], [13] have proven competent at computing depth. Stereo occlusion is a further cue that needs to be accurately computed [20], [5], [23], [16] to achieve good layer extraction. However, the strength of stereo cues degrades 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 [10], [27], even in the absence of texture. Segmentation based on colour/contrast alone is nonetheless beyond the capability of fully automatic methods. This

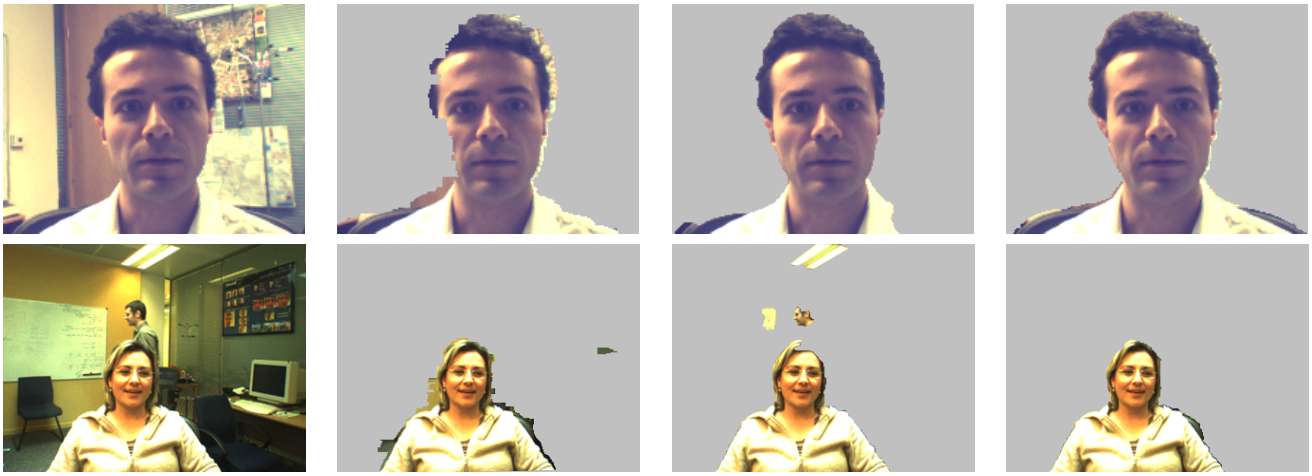


Fig. 2. **Segmentation by fusing colour, contrast and stereo.** Results of three different segmentation algorithms run on two different stereo-pairs (see [1] for more examples). **a)** data (left image); **b)** Segmentation based on stereo [16]; **c)** Segmentation based on colour/contrast [27]; **d)** The LGC algorithm proposed here fuses colour, contrast and stereo to achieve a more accurate segmentation. The foreground artefacts visible in b) and c) are corrected in d).

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 graph cut [10], [11]. Therefore two approaches to segmentation are proposed here: Layered Dynamic Programming (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 V). Finally, efficient post-processing for matting [14] 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 and then conclusions in section 6.

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

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

A. Notation and basic framework

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

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

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

$$\mathbf{z} = \{z_k,\ k = 0, \ldots, 2N + 1\},$$

where $k = m + n$. The k-axis is the so-called cyclopean[1] coordinate axis. Conventionally in stereo matching the so-called "ordering constraint" is imposed, and this means that each move in figure 3 is allowed only in the positive quadrant [3], [25]. Furthermore, in our framework, only single-step horizontal and vertical moves are allowed — no diagonal or multistep moves. The reason for this — it makes for a cleaner probabilistic model — is explained

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

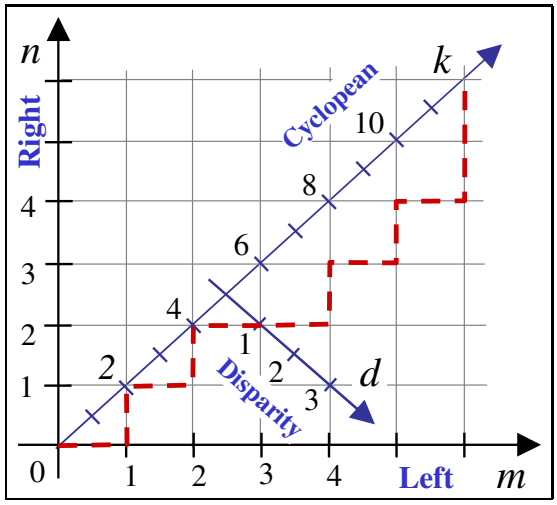


Fig. 3. **Stereo match-space.** Notation conventions for left and right epipolar lines with pixel coordinates m, n, cyclopean coordinates k and stereo disparity $d = m - n$. Possible matching path shown dashed (*cf.* [5], [13]).

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

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

Cyclopean (k, d) coordinates form an alternative coordinate system to (m, n) in the matching diagram, and are well known to be helpful for probabilistic modelling of matching [5]. 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 pixel is a foreground match, a background match or occluded.

This sets up the notation for a path in match-space which is a sequence $(\mathbf{d}, \mathbf{x})$ of disparities and states. A Gibbs energy $E(\mathbf{z}, \mathbf{d}, \mathbf{x}; \Theta, \Phi)$ can be defined for the posterior over the inferred sequence $(\mathbf{d}, \mathbf{x})$ given the image data $\mathbf{z}$. Parameters Φ and Θ relate respectively to prior and likelihood terms in the posterior, and will be explained in more detail below. Then the Gibbs energy can be globally minimised to obtain a segmentation $\mathbf{x}$ and, as a bi-product, disparities $\mathbf{d}$. The LDP algorithm (section III) minimises the Gibbs energy separately over each epipolar line. Alternatively, the LGC algorithm (section IV) minimises, globally over the images, a modified Gibbs energy $E(\mathbf{z}, \mathbf{x}; \Theta, \Phi)$ in which disparity variables do not explicitly appear. The result is an estimate $\mathbf{x}$ of foreground/background segmentation, but without the bi-product of stereo disparities.

B. Prior distribution over matching paths

In the remainder of this section a broadly Bayesian model for the posterior distribution $p(\mathbf{x}, \mathbf{d} \mid \mathbf{z})$ is set up as a product of prior and likelihood:

$$p(\mathbf{x}, \mathbf{d} \mid \mathbf{z}) \propto p(\mathbf{x}, \mathbf{d}) p(\mathbf{z} \mid \mathbf{x}, \mathbf{d}). \tag{2}$$

The prior distribution $p(\mathbf{x}, \mathbf{d})$ is decomposed, in the interests of tractability, as a Markov model, either as Markov chains along scanlines, for LDP, or as a disparity-independent Markov Random Field (MRF) $p(\mathbf{x})$ over an entire image, for LGC. The Markov chain model decomposes the prior as

$$p(\mathbf{x}, \mathbf{d}) = p(x_0, d_0) \prod_k p(x_k, d_k \mid x_{k-1}, d_{k-1}) \tag{3}$$

in which the transition kernel $p(x_k, d_k \mid x_{k-1}, d_{k-1})$ is sparse. The sparsity has the effect of restricting the space of allowed moves in match-space (figure 3) to a small set (see below). Within that set, transition probabilities favour runs within the foreground and within the background states; within matched and unmatched states; and favour low disparity in the background with high disparity in the foreground. Details are given in section III. More generally, an MRF prior for $(\mathbf{x}, \mathbf{d})$ is specified as a product of clique potentials $F_{k,k'}$ over all pixel pairs $(k, k') \in \mathcal{N}$ deemed to be neighbouring in the cyclopean image (LDP), or the left image (LGC). For LDP we

have $F_{k,k'} = F_{k,k'}(x_k, d_k, x_{k'}, d_{k'})$ but restricted, in Markov chains, to horizontal pixel-pairs – *i.e.* pairs that are neighbours in a particular epipolar line. In LGC, where disparities do not appear explicitly, $F_{k,k'} = F_{k,k'}(x_k, x_{k'})$ which is simpler than in LDP, except that pairs occur also across neighbouring epipolar lines.

Stepwise restriction for LDP: Previous matching algorithms e.g. [13], [18] have allowed multiple and/or diagonal moves on the stereo matching paths (fig 3). However, the problem here differs significantly. In [13], [18] diagonal moves are always matched, and horizontal/vertical ones are unmatched. However the nature of the stereo matching problem demands that horizontal/vertical moves should come both in matched and unmatched forms. (Matched horizontal/vertical moves are needed to represent the deviation of a visible surface from fronto-parallel). This raises a consistency requirement between matched move types: a path consisting of a sequence of diagonal moves is exactly equivalent to a corresponding path in which horizontal and vertical moves alternate strictly. The probabilities of the two paths should therefore be identical. This is most easily achieved simply by outlawing explicit, diagonal matched moves, forcing them to be expressed instead as a horizontal/vertical pair. This restriction, illustrated in fig. 3, thus ensures a consistent probabilistic interpretation of the sequence matching problem. Furthermore, the stepwise restriction has the added virtue that each element L_m and R_n is "explained" once and only once. This is because a horizontal step in fig. 3 visits a new L_m, which is thereby "explained" but stays with the old R_n. Conversely, a vertical step visits a new R_n. Thus each L_m and each R_n appears once and only once as z_k in the $p(z_k|\ldots)$ term of the joint likelihood $p(\mathbf{z} \mid \mathbf{x}, \mathbf{d})$ (4) below, making for a consistent definition of the likelihood.

C. 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

$$\begin{aligned} p(\mathbf{z} \mid \mathbf{x}, \mathbf{d}) &= \prod_k p(z_k \mid x_k, d_k, z_1, \ldots, z_{k-1}) \\ &= K(\mathbf{z}) \prod_k \exp -U_k^{\mathrm{M}}(x_k, d_k) \end{aligned} \tag{4}$$

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

$$\begin{aligned} U_k^{\mathrm{M}}(x, d_k) &= -\log p(z_k \mid x_k = x, d_k, z_1, \ldots, z_{k-1}) \\ &+ \log p(z_k \mid x_k = \mathrm{O}). \end{aligned} \tag{5}$$

According to the definition, $U_k^{\mathrm{M}}(\mathrm{O}, d_k) = 0$. Commonly [29] stereo matches are scored using SSD (sum-squared difference), that is L^2-norm of difference between image patches $L_m^{\mathrm{P}},\ R_n^{\mathrm{P}}$ surrounding hypothetically matching pixels m, n. Following [16] we model U_k^{M} in terms of SSD but with additive and multiplicative normalisation for robustness to non-Lambertian effects and photometric calibration error. This is termed NSSD — normalized SSD:

$$U_k^{\mathrm{M}}(x_k, d_k) = \begin{cases} M(L_m^{\mathrm{P}},\ R_n^{\mathrm{P}}) & \text{if } x_k \in \{\mathrm{F}, \mathrm{B}\} \\ 0 & \text{if } x_k = \mathrm{O}, \end{cases} \tag{6}$$

where $M = \lambda(N - N_0)$ with λ a constant, and the NSSD N is:

$$N(L^{\mathrm{P}},\ R^{\mathrm{P}}) = \frac{1}{2} \frac{\|L^{\mathrm{P}} - R^{\mathrm{P}}\|^2}{\|L^{\mathrm{P}} - \overline{L^{\mathrm{P}}}\|^2 + \|R^{\mathrm{P}} - \overline{R^{\mathrm{P}}}\|^2} \in [0, 1], \tag{7}$$

in which $\overline{R^{\mathrm{P}}}$ denotes the mean value over the patch R^{P}. As a refinement, we further allow for subpixel offset by parabolic interpolation, along epipolar lines, of the values of

$$N(L_m^{\mathrm{P}},\ R_{n-1}^{\mathrm{P}}),\ N(L_m^{\mathrm{P}},\ R_n^{\mathrm{P}}),\ N(L_m^{\mathrm{P}},\ R_{n+1}^{\mathrm{P}}),$$

and take the minimum value of the parabola to replace the value of $N(L_m^{\mathrm{P}},\ R_n^{\mathrm{P}})$, where it is needed in the matching algorithm. This subpixel refinement was found to improve error rates mildly, and was similar in effect to alterative interpolation schemes [24], [7]. This model has been tested against the Middlebury data-sets [2] and found to be reasonable — examples of results are given in fig. 4a). Importanly, such analysis gives useful working values for

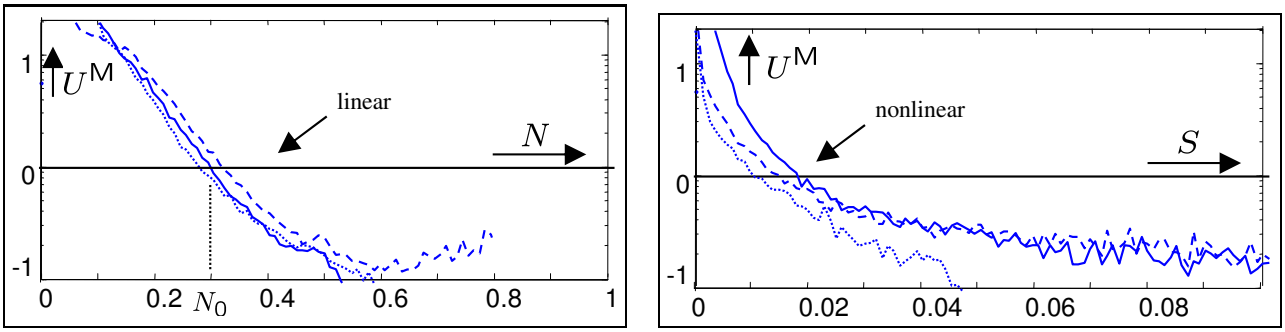


Fig. 4. **Likelihood model:** the empirical negative-log-likelihood ratio U^{M} is shown for stereo matches, plotted here (a) as a function of the NSSD measure $N(L^{\mathrm{P}},\ R^{\mathrm{P}})$, using the ground truth stereo data from three of the Middlebury data sets [2] ("cones", "teddy", and "sawtooth"). Note the linearity in the region of $U^{\mathrm{M}} = 0$, where discrimination is most critical. The more commonly used SSD measure is also analysed (b) but gives a non-linear U^{M}, which is also less consistent across datasets.

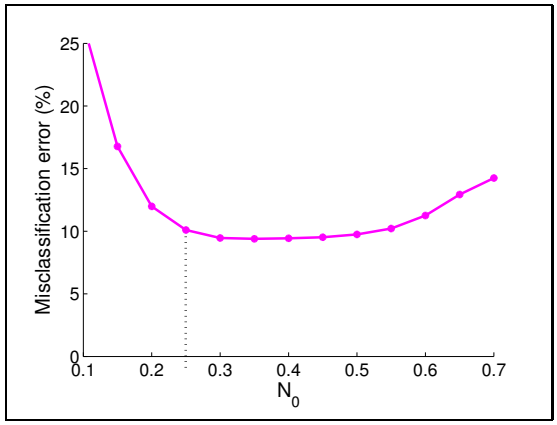


Fig. 5. **Sensitivity of the likelihood ratio offset parameter:** the value of the parameter N_0 affects the error rate in classification of occlusions, for the 4-state DP algorithm described below in section III. Results are shown here for the "cones" data set of figure 4. The value $N_0 = 0.25$, estimated by linear fitting of the likelihood ratio, gives performance that is close to optimal.

λ, which turns out to be quite consistent at around $\lambda = 10$. [2] For the parameter N_0, the data analysis yields a value of approximately 0.3. However, we found the discriminatively optimal N_0 is usually a little larger: a typical value is $N_0 = 0.4$, and that value gives better error rates in practice. An example of the sensitivity of the N_0 parameter is shown in figure 5.

As it has been more conventional [29] in stereo to use SSD as a match-cost rather than NSSD, results are included also for U^{M} modelled as a function of SSD, in fig. 4b). Two issues arise from this. The first is that an effect of normalisation is that the U^{M}-characteristic is more consistent across data sets for NSSD than for SSD. Hence it is reasonable to fix the function used to model the $\mathcal{L}$-characteristic in the NSSD case, whereas for SSD adaptivity would be necessary. The second is that the linearity apparent for NSSD is absent for SSD. Therefore the statistical evidence does not support the conventional modelling of match-cost as proportional to SSD. In fact the data fits an inverse power model $U^{\mathrm{M}} \propto 1/S^r$, with r varying in the range $0.6 \leq r \leq 1.7$ over the Middlebury data set. With this non-linear likelihood, we have found DP stereo based on SSD to perform at comparable error rates to NSSD, or slightly worse. On balance the linearity and consistency of the likelihood for NSSD are reasons why we prefer to assume NSSD as the sufficient statistic for discriminating matches from mismatches.

D. Likelihood for colour

Following previous approaches to two-layer segmentation [10], [27] we model likelihoods for colour in foreground and background using Gaussian mixtures in RGB colour space, learned from image frames labelled (automatically)

[2]From monochrome components of the 8 images in the Middlebury set, we obtain $\lambda = 10.5 \pm 1.5$ for 5×5 patches as used in LGC, and $\lambda = 10.1 \pm 1.4$ for 3×7 patches as used in LDP.

from earlier in the sequence. The foreground colour model $p^{\mathrm{F}}(z)$ is simply a spatially global Gaussian mixture learned from foreground pixels, and similarly for the background model $p^{\mathrm{B}}(z)$. The combined colour model is then given by an energy U_k^C:

$$U_k^C(z_k, x_k) = \begin{cases} -\log p^{\mathrm{F}}(z_k) & \text{if } x = \mathrm{F} \\ -\log p^{\mathrm{B}}(z_k) & \text{if } x = \mathrm{B} \text{ or } x = \mathrm{O} \end{cases} \tag{8}$$

Learning of the global foreground and background colour models p^{F} and p^{B} proceeds as follows. Each is a mixture of $N_C = 20$ full covariance Gaussian components in RGB colour-space, and is learned, at each video timestep, using 10 iterations of EM [17], initialised from the mixture in the previous frame. The data is taken from the previous timestep, labeled as foreground/background from the output of the segmentation process. In the case of LGC, the algorithm will be defined with respect to one (the left) image only, so colour models are built from that one image. In the case of the LDP algorithm, models are maintained independently for each of the left and right images. The total energy for colour is taken as:

$$U^{\mathrm{C}}(\mathbf{z}, \mathbf{x}; \Theta) = \rho \sum_k U_k^{\mathrm{C}}(z_k, x_k) \tag{9}$$

where the *colour discount* constant ρ (typical value $\rho = 1/2$) is included to tune the balance of influence between the stereo model and the colour model. In principle, the generative derivation of the energies should have balanced them already. In practice, the pixelwise independence assumptions built in to the colour model renders the influence of colour excessively strong, and choosing a value $\rho < 1$ discounts for that. Colour models are initialised by switching them off at time $t = 0$ by setting the weight $\rho = 0$, and then switching it to its final value at time $t = 1$. (A more progressive strategy might seem reasonable, but is found in practice to be unnecessary.)

E. Contrast model and figural continuity

There is a natural tendency for segmentation boundaries in images to align with contours of high contrast and it is desirable to represent this as a constraint in stereo matching. This can be achieved by adjusting the prior penalties $F_{k,k'}$ associated with segmentation boundaries, abating them where there is evidence from image contrast. This is related to the very well known themes in image-segmentation of "line processes" [21], "weak constraints" [9] and anisotropic diffusion [26]. In a recent, particularly effective model for binary segmentation [10] a penalty is associated with boundaries, and abated by a discount factor that depends monotonically on image contrast. Simpler versions of such contrast models have been used previously in stereo algorithms [11], [23] to favour figural continuity. From the probabilistic point of view, the combined penalty and discount seems to obscure the separation between prior distribution and likelihood. However it has been shown, at least for binary segmentation, that a consistent interpretation of segmentation-prior and contrast-likelihood does exist [8].

Here we define a discounted penalty for the stereo matching problem as an image energy of the form

$$V_{k,k'} = F_{k,k'} V^*(z_k, z_{k'}), \tag{10}$$

where k, k' are neighbouring pixel-pairs in the cyclopean image. The function $F_{k,k'}$ is the clique potential coefficient defined earlier in section II-B. The exact form of $F_{k,k'}$ is different for LDP and LGC, and it is given later in corresponding sections. Generally, it has the effect of applying a penalty at boundaries, where the state changes between $x = \mathrm{F}, \mathrm{B}, \mathrm{O}$. The term V^* is the contrast sensitive discount to the boundary penalty (10):

$$V^*_{k,k'}(z, z') = \frac{1}{1+\epsilon}\left(\epsilon + \exp -\frac{\|G(z) - G(z')\|^2}{2\sigma^2_{k,k'} {d_{k,k'}}^2}\right), \tag{11}$$

where $G(\ldots)$ is a Gaussian smoothing filter at the (approximately) Nyquist scale of 0.7 pixels, $d_{k,k'}$ is the Euclidean distance between pixels k, k' and $\sigma^2_{k,k'} = \left\langle \|G(z_k) - G(z_{k'})\|^2 / {d_{k,k'}}^2 \right\rangle$, a mean contrast over all neighbouring pairs of image pixels. The constant ϵ is a "dilution" constant for contrast, previously [10] set to $\epsilon = 0$ for pure colour segmentation. Here, $\epsilon = 1$ seems more appropriate — diluting the influence of contrast in recognition of the increased diversity of segmentation cues, and mild supporting evidence for this is given later.

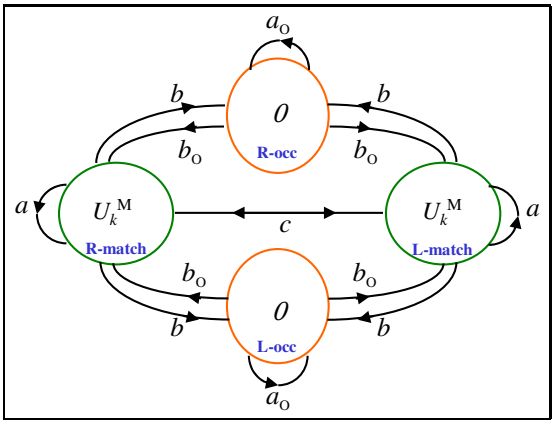


Fig. 6. **State space for stereo matching with occlusion.** Matched and occluded states (each in left and right forms) make up a 4-state system. Successive pixels along a cyclopean epipolar line (fig. 3) incur a cost increment (*e.g.* b) for the arc $k-1 \rightarrow k$ traversed, plus an increment (*e.g.* U_k^{M}) for the new node k.

III. 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 [13], [25], together with augmentation of the state space to handle occlusion by means of the "4-state" model [15]. As a general comment, it is worth acknowledging at this point that DP restricted to scanlines obviously cannot perform exact inference on the model as set out in the previous section, as there is no explicit imposition of constraints between epipolar lines. This is one of the advantages of the alternative LGC algorithm, described in the next section, which does fully integrate constraints. Nonetheless, in DP there is some implicit transfer of information across scanlines in that the patches used to define the stereo-matching likelihood (section II-C) in adjacent epipolar lines do overlap, so that the evaluated likelihoods will be somewhat correlated in adjacent locations on adjacent epipolar lines. As previous studies have shown [15] this implicit imposition of constraints is quite successful in reducing stereo labelling artefacts. Further reduction of epipolar artefacts is encouraged by imposing the figural continuity constraint described earlier in section II-E, given that edge features tend naturally to be coherent.

This section sets out the Markov model underlying the LDP algorithm. First the 4-state model for stereo is reviewed in section III-A. To achieve segmentation, foreground/background states are then added to the 4-state model, together with colour/contrast energy, to arrive at a new 6-state model, which is described in section III-B. In summary, it is defined by an energy function composed of four terms:

$$E(\mathbf{z}, \mathbf{x}; \Phi, \Theta) = V(\mathbf{z}, \mathbf{x}; \Theta) + U^{\mathrm{M}}(\mathbf{z}, \mathbf{x}, \Phi) + U^{\mathrm{D}}(\mathbf{z}, \mathbf{x}, \Phi) + U^{\mathrm{C}}(\mathbf{z}, \mathbf{x}; \Theta), \tag{12}$$

representing energies for spatial coherence/contrast, stereo likelihood, disparity-pull and colour-likelihood respectively.

A. 4-state stereo with occlusions

The 4-state model for stereo matching is reviewed in this section; its basic structure is summarised in fig. 6. The 4-state system and its transitions has associated energy terms that define a global energy

$$E(\mathbf{z}, \mathbf{d}, \mathbf{x}; \Theta, \Phi) = \sum_k E_k(d_k, d_{k-1}, x_k, x_{k-1}) \tag{13}$$

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

$$E_k = V_{k-1,k} + U_k^{\mathrm{M}} \tag{14}$$

of a state cost U_k^{M}, inside nodes on the diagram of fig. 6, and a cost $V_{k-1,k}$ of transition $k-1 \rightarrow k$ (on arcs). The occluding state $x_k = \mathrm{O}$ is split into two sub-states (red circles in fig. 6), left-occluding and right-occluding (which do not intercommunicate, reflecting geometric constraints). The matching state $x_k = \mathrm{M}$ also has left and

right substates (green circles in fig. 6). The typical progress of a matching path then alternates between left and right, as in figure 3. In both cases, matched and occluding, handedness h_k can be computed directly from disparity as follows:

$$h_k = \begin{cases} \textbf{Left} & \text{if } d_k = d_{k-1} + 1 \\ \textbf{Right} & \text{if } d_k = d_{k-1} - 1. \end{cases} \tag{15}$$

There are a total, then, of 4 possible states: $x_k \in \{\text{L-match, R-match, L-occ, R-occ}\}$. Match costs inside nodes are defined in terms of match likelihood energy defined earlier (6), so that:

$$U_k^{\mathrm{M}} = M(L_m, R_n), \tag{16}$$

with m, n calculated from disparity as in (1).

The prior model over matching paths $F_{k,k'}$ (section II-B) is expressed in terms of a number of parameters $\Phi = \{a_{\mathrm{O}}, b_{\mathrm{O}}, a, b, c\}$ (figure 6). 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:

$$b_{\mathrm{O}} = \log(2W_{\mathrm{O}}) \quad b = \log(2W_{\mathrm{M}}) \tag{17}$$

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 $2\exp -b_{\mathrm{O}}$ from an occluded state. Then consideration of viewing geometry (details omitted) indicates:

$$a = \log(1 + D/B) - \log(1 - 1/W_{\mathrm{M}}), \tag{18}$$

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

$$c = -\log(1 - 2e^{-b} - e^{-a}) \quad \text{and} \quad a_{\mathrm{O}} = -\log(1 - 2e^{-b_{\mathrm{O}}}),$$

so the number of independent parameters in Φ is reduced to three: a, b_{O} and b.

B. 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 from the 4-state system are split into foreground and background substates. The diagram of fig. 6 is cut by the splitting of the matched states to give a total of 6 possible states: $x_k \in \{\text{L-match-F, R-match-F, L-match-B, R-match-B, L-occ, R-occ}\}$. This is reflected in the topology of the extended state-space diagram of fig. 7 which has 6 possible states: $x_k \in \{\text{L-match-F, R-match-F, L-match-B, R-match-B, L-occ, R-occ}\}$, with costs $V_{k-1,k}$ of transition $k-1 \to k$ on arcs and state costs U_k^{M} inside nodes, as before. 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 both for the $x_k = \mathrm{F}$ and for the $x_k = \mathrm{B}$ conditions.

C. Adding disparity-pull and colour/contrast fusion

It remains to add in energies for the colour and contrast likelihoods. The full energy for stereo matching, per cyclopean pixel, is now

$$E_k = V_{k-1,k} + U_k^{\mathrm{M}} + U_k^{\mathrm{D}} + U_k^{\mathrm{C}} \tag{19}$$

where U_k^{M} and $V_{k-1,k}$ are respectively the node and transition energies from section III-B. The nodal energy has been 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 maintained as described earlier in section II-D, and with one foreground/background model pair for each of the left and right images. The constant $\rho < 1$ to discount the strength of the colour model is included as before. This gives a colour energy term of the form

$$U_k^{\mathrm{C}} = U_k^{\mathrm{C}}(z_k, x_k, h_k), \tag{20}$$

where h_k takes the value $\mathbf{Left}$ or $\mathbf{Right}$, and is computed as in (15). The disparity-pull energy

$$U_k^{\mathrm{D}}(z_k, x_k) = -\log p(d_k | x_k) \tag{21}$$

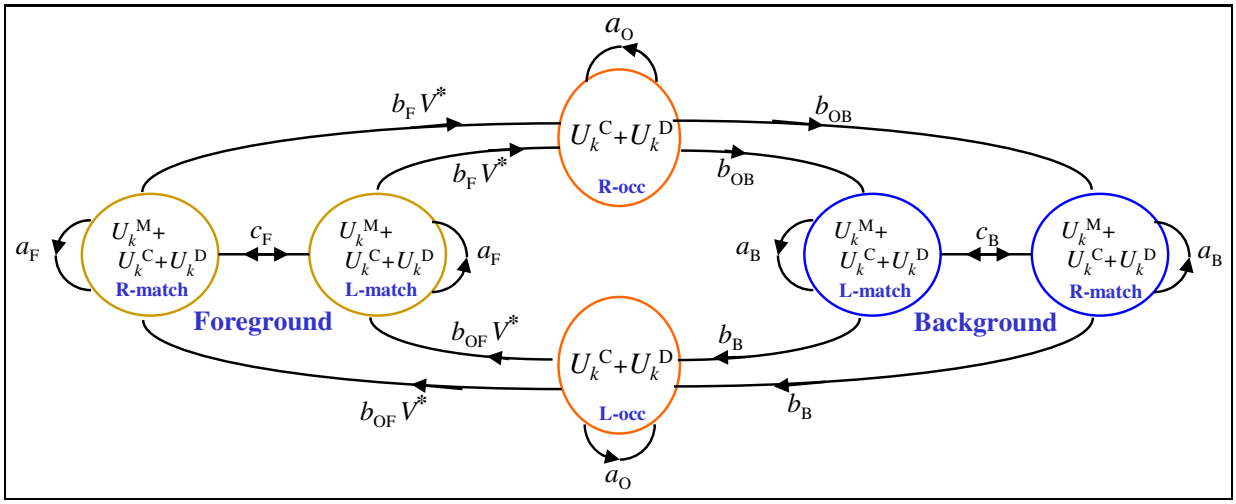


Fig. 7. **Extended state space for foreground/background segmentation.** The matched state of fig. 6 is split into a foregound and a background substate. Note that from the foreground states (yellow circles), only the *right* occluding state is accessible, and from background (blue circles) only the *left* occluding state — reflecting a neglect, in the interests of simplifying our model, of the possibility of foreground/foreground occlusion. Modified match costs now incoprorate disparity-pull and contrast effects — see text for details.

represents the pull of each layer towards certain disparities, as determined by the pull-densities $p(d_k|x_k = \mathrm{F}, \mathrm{B}, \mathrm{O})$. Typically this term pulls the foreground/background layers towards larger/smaller values of disparity respectively. From the point of view of Bayesian modelling, the U^{D} term should be considered as a modification of the matching path prior, to take account of foreground/background influence.

Finally, the transition component $V_{k-1,k}$ from the 6-state model is further modified to take account of contrast (10). This is done by modifying each transition energy between occluding and foreground states (fig. 7) as follows:

$$b_F \rightarrow b_F V_k^* \quad \text{and} \quad b_{OF} \rightarrow b_{OF} V_k^*, \tag{22}$$

where contrast discount V^* is defined as before (11), but applying to the left or right image as appropriate:

$$V_k^* = \begin{cases} V^*(L_m, L_{m-1}) & \text{if } h_k = \textbf{Left} \\ V^*(R_n, R_{n-1}) & \text{if } h_k = \textbf{Right}. \end{cases} \tag{23}$$

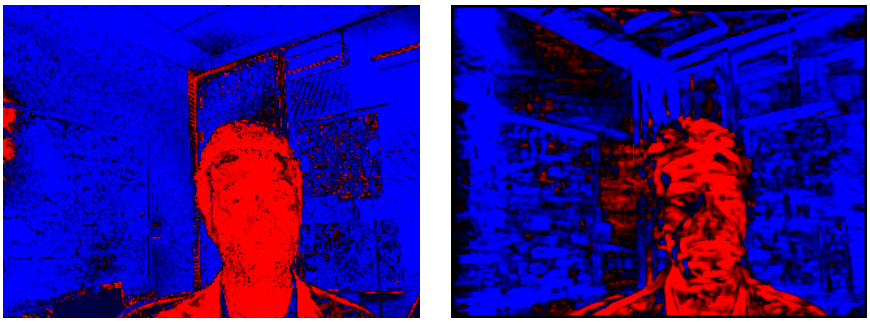
Now the full 6-state system, augmented both for bi-layer inference and for fusion of colour/contrast with stereo can be optimised by dynamic programming as before. Results of this approach are shown below in section V, but in the meantime the alternative LGC algorithm is described. This effectively defines the $F_{k,k'}$ terms from section II-E. At this point all prior and likelihood parameters for the LGC model have been defined.

IV. 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, the stereo match distribution (4) in section II-C is marginalised over disparity, aggregating support from each putative match, to give a likelihood $p(\mathbf{L} \mid \mathbf{x}, \mathbf{R})$ for each of the three label-types in $\mathbf{x}$: foreground, background and occlusion $(\mathrm{F}, \mathrm{B}, \mathrm{O})$. Segmentation is therefore a ternary problem, and it can be solved (approximately) by iterative application of a binary graph-cut algorithm, augmented for a multi-label problem by so-called α-expansion [11]. Thus the LGC algorithm is an alternative way of implementing the colour-stereo fusion idea, that turns out to be very effective. A particular difference between LDP and LGC is that, given that it does not explicitly solve for stereo disparity, LGC is most conveniently specified with respect to one (*e.g.* left) image, rather than in the cyclopean frame as in LDP.

The energy function for LGC is composed of three terms (*cf.* 4-term energy (12) in section III for LDP):

$$E(\mathbf{z}, \mathbf{x}; \Theta, \Phi) = V(\mathbf{z}, \mathbf{x}; \Theta) + U^{\mathrm{S}}(\mathbf{z}, \mathbf{x}, \Phi) + U^{\mathrm{C}}(\mathbf{z}, \mathbf{x}; \Theta), \tag{24}$$



a) *colour* **b)** *stereo (F vs. B)*

Fig. 8. **Colour and stereo log-likelihood ratios in LGC.** If a value is positive, it is shown in the red channel, otherwise it is shown in the blue channel. (a) $-U_m^C(L_m, F) + U_m^C(L_m, B)$. (b) $-U_m^S(L_m, F) + U_m^S(L_m, B)$. Results for sequence AC at frame 0.

representing energies for spatial coherence/contrast, stereo likelihood and colour-likelihood respectively. The colour energy is simply a sum over pixels, as before (9), but now over the left image only:

$$U^{\mathrm{C}}(\mathbf{z}, \mathbf{x}; \Theta) = \rho \sum_m U_m^{\mathrm{C}}(L_m, x_m) \tag{25}$$

of the colour energy defined earlier (8), with the adjustment factor ρ as before. Typical color likelihoods are shown in Fig. 8a.

The coherence/contrast energy $V(\mathbf{z}, \mathbf{x}; \Theta)$ is a sum, over cliques, of pairwise energies of the form (10) in section II-E, but with the potential coefficients $F_{m,m'}$ now defined as follows. Cliques consist of horizontal, vertical and diagonal neighbours on the square grid of pixels. 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. Horizontal cliques, along epipolar lines, inherit the same cost structure, except that certain transitions are disallowed on geometric grounds. These constraints are imposed via infinite cost penalties:

$$F_{m,m'}[x = \mathrm{F}, x' = \mathrm{O}] = \infty; \quad F_{m,m'}[x = \mathrm{O}, x' = \mathrm{B}] = \infty.$$

The constant γ is broadly related to b and b_{O} in the LDP model, so a reasonable working value for γ is

$$\gamma = \frac{1}{2}(b + b_{\mathrm{O}}) = \log(2\sqrt{W_{\mathrm{M}} W_{\mathrm{O}}}), \tag{26}$$

where width parameters W_{M} and W_{O} were defined earlier (17).

A. Marginalisation of stereo likelihood

The remaining term in (24) 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

$$U^{\mathrm{S}}(\mathbf{z}, \mathbf{x}) = \sum_m U_m^{\mathrm{S}}(x_m) \tag{27}$$

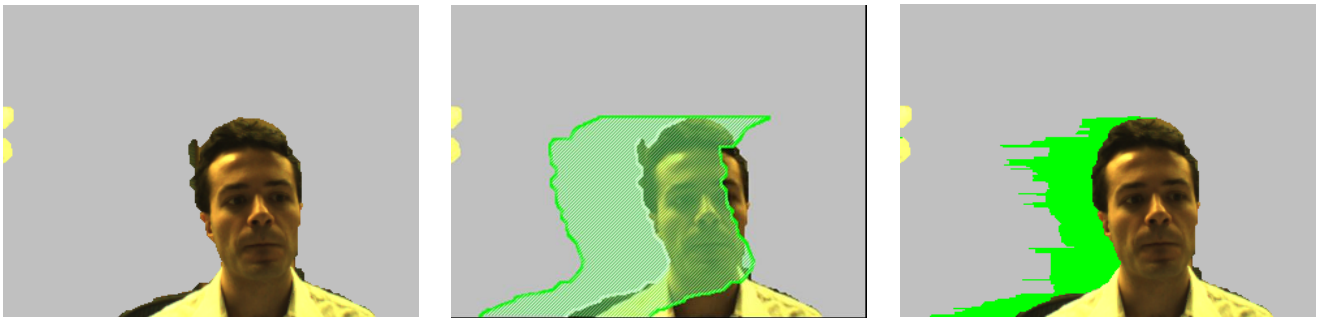
$$\text{where } U_m^{\mathrm{S}}(x_m) = -\log p(L_m | x_m, \mathbf{R}) + \mathrm{const}, \tag{28}$$

$$p(L_m | x_m, \mathbf{R}) = \sum_d p(L_m | x_m, d_m = d, \mathbf{R}) p(d_m = d | x_m) \tag{29}$$

— marginalizing over disparity, and the distributions $p(d_m = d | x_m)$ for $x_m \in \{\mathrm{F}, \mathrm{B}\}$ are fixed to disjoint uniform distributions, and $p(d_m = d | x_m = \mathrm{B}) = p(d_m = d | x_m = \mathrm{O})$. (Alternatively, at least for LDP, the distributions could be learned adaptively using labelled data from previous frames.) The const term in (28) allows us to make use of the likelihood-ratio model of section II-C for stereo matches, giving

$$U_m^{\mathrm{S}}(x_m) = \\ -\log \left[\textstyle\sum_d p(d_m = d | x_m) \exp -\lambda N(L_m^{\mathrm{P}},\ R_n^{\mathrm{P}}) \right] - \lambda N_0. \tag{30}$$

Typical stereo likelihoods are shown in Fig. 8b.



a) *F-expansion* **b)** *O-expansion region* **c)** *O-expansion*

Fig. 9. **One iteration of the expansion move algorithm in LGC.** Configuration is initialized with $x_m = B$ for all pixels, then subjected to F-expansion to give (a). (b) O-expansion is restricted to a region close to B-F transitions, shown shaded, to give the final result (c), in which the O-label is shown in green. (Results for sequence AC at frame 0.)

B. Expansion move algorithm

Currently, graph cut based stereo algorithms techniques such as [10], [23] are not suited for real-time implementation. The main reason is that they perform $O(d_{max})$ alpha expansion operations (binary graph cuts), where d_{max} is the number of possible disparities. Having marginalized over disparities, we are left with just three labels which is a substantial saving. In addition, the ternary expansion move algorithm can be implemented practically at a cost of a single graph computation by taking advantage of the structure of our problem.

First, we have observed that results after one iteration of the expansion move algorithm are very close to the results achieved at convergence. This is not surprising considering that the number of labels is small. Therefore, only one iteration, involving two graph cut computations, is needed. We initialize the segmentation with $x_m = B$ for all pixels and then run F-expansion and O-expansion (see fig. 9). Second, in the O-expansion operation it suffices to add nodes only for a small fraction of all pixels. Indeed, due to the geometric constraints O-expansion cannot change pixels in scanlines that do not contain B-F type transitions. Furthemore, it happens that the segmentation boundary found after F-expansion normally lies in the real occluded region located to the left of foreground object. Therefore, it is reasonable to perform O-expansion operation only for pixels within distance d_{max} from B-F transitions (fig. 9b).

Results of segmentation using LGC and LDP are given in the next section.

V. RESULTS

Performance of the LGC and LDP algorithms was evaluated with respect to ground-truth segmentations on 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.

Prior parameters for LDP: Prior parameters for LDP are set as in section III-A, equations (17) and (18), with the same values for foreground and background parameters, *i.e.* a_{F} and a_{B} *etc.*. Regions widths in equations (18) and (17) are set to $W_{\mathrm{O}} = 10$ pixels and $W_{\mathrm{M}} = 100$ pixels, and typical values for object distance and baseline are $D = 1000$ mm and $B = 50$ mm.

A. Determination of LGC parameters and their sensitivity

The first set of experiments, with the LGC algorithm, are shown in figure 10. Parameters N_0, γ, ρ and ϵ are varied, one at a time, around their default values $N_0 = 0.35$, $\gamma = 2$, $\rho = 0.5$ and $\epsilon = 1$. Results are summarised for each parameter in turn.

Likelihood offset parameter N_0, introduced in section II-C, gives low error rates over a range $0.25 \leq N_0 \leq 0.35$. Note that $N_0 = 0.25$ is the value obtained generatively, *i.e.* from likelihood fitting in section

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

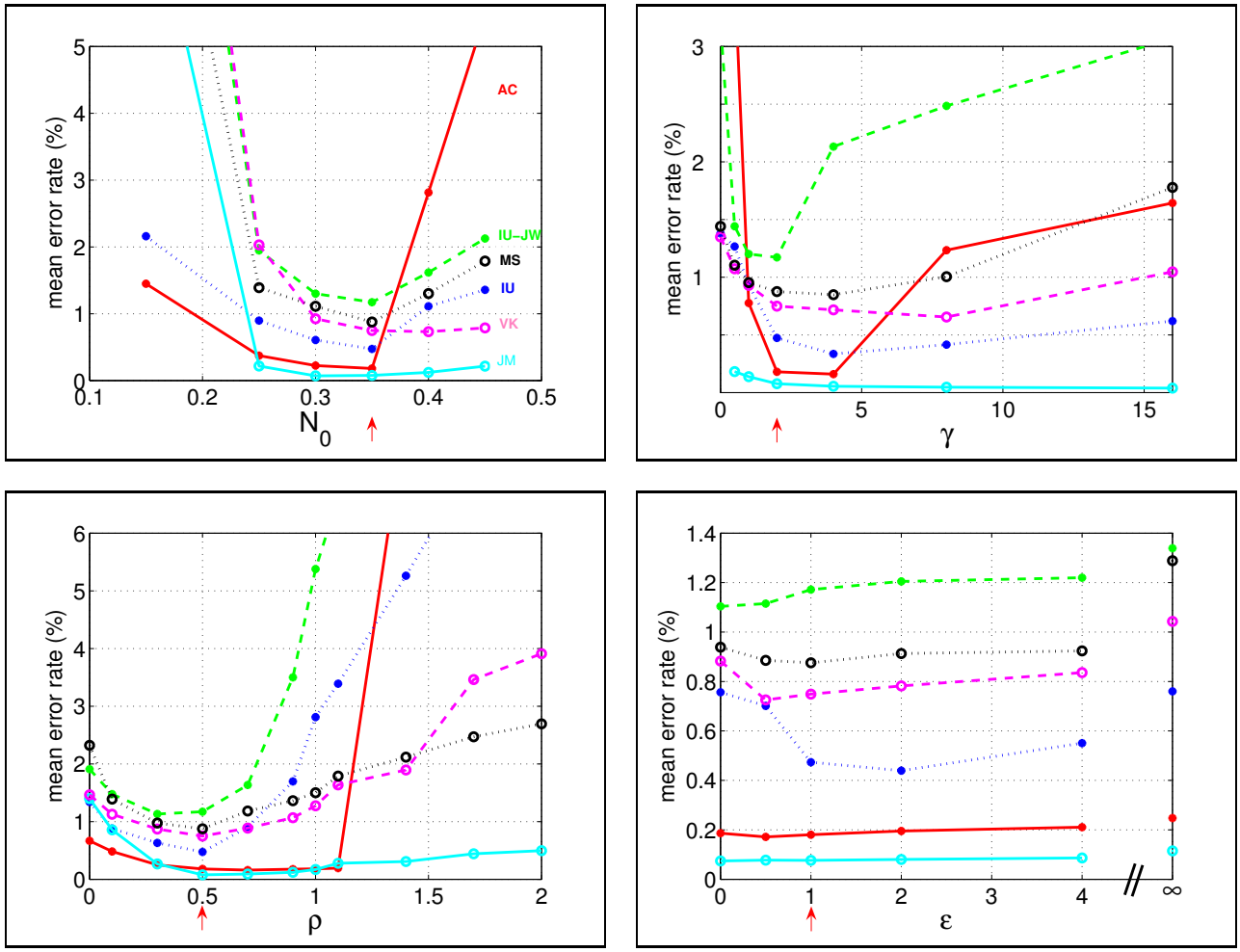


Fig. 10. **Effect of values of LGC parameters** N_0, γ, ρ and ϵ on segmentation error rate, for each of 6 test-data sets — see text for detailed discussion. The default value of each parameter is indicated by an arrow on the abscissa axis.

II-C. The value $N_0 = 0.35$ is very slightly superior discriminatively — *i.e.* it gives lower error rate in figure 10.

Coherence constant γ for LGC, defined in section IV, gives low error rates for $2 \leq \gamma \leq 4$. Notably this is far smaller than the optimal value $\gamma \approx 25$ for segmentation using colour/contrast only [27]. Presumably the presence of the additional cue from stereo to some extent takes over the role of coherence. The default value, from equation (26) in section IV, and taking $W_{\rm O} = 10$ pixels and $W_{\rm M} = 100$ pixels as before, gives $\gamma = 4.1$ which is entirely consistent with the experimental results.

Colour discount constant ρ, defined in section II-D equation (9), gives best error rates around $\rho = 0.5$. Without a discount ($\rho = 1$) error rates are appreciably higher, and this confirms the need for a discount to modify the generative assumption of independence of colour at neighbouring pixels.

Contrast parameter ϵ, defined in section II-E, equation (11) to impose figural continuity, has a mild effect on error rate performance. Our default $\epsilon = 1$ performs a little better than either removing the contrast term altogether ($\epsilon = \infty$), or setting it at full strength ($\epsilon = 0$) as done in *GrabCut* [27].

In all four cases, error rate performance is seen to be quite robust as parameters vary around their default values.

Pixelwise background model: We further experimented with an extension to the background model of section II-D, mixing in a probability density learned, for each pixel, by pixelwise background maintenance [28], [30], [32]. The learned pixelwise densities $p_k(z_k)$ 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 $p^{\rm B}(z_k)$ as the contamination component in the mixture. However, rather surprisingly, experiments showed negligible improvement from the extended background model, presumably because of the strength of the other cues. A density

equally weighted between $p_k(z_k)$ and $p^{\mathrm{B}}(z_k)$ decreased error rates by just 0.03–0.3% across the 6 data sets tested (see section V), compared with using $p^{\mathrm{B}}(z_k)$ alone. Note however that using the pixelwise $p_k(z_k)$ alone, without any $p^{\mathrm{B}}(z_k)$ component, increased error rates by a disastrous $0.5 - 8.1\%$.

B. Error rate reduction due to fusion of stereo/colour/contrast

Segmentation performance for the various stereo test-sequences, including the AC sequence of fig.1 and five others, is compared for colour/contrast, for stereo alone, and for colour/contrast with stereo fused together (fig. 11). 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 III-A. Fusion of colour/contrast and stereo by the LGC and LDP algorithms both show similarly enhanced performance compared with colour/contrast or stereo alone. The six test sequences include one with two subjects in the foreground (IU–JW) and another with people moving in the background (IU). Even in those difficult cases, the power of fusing colour/contrast and stereo is immediately apparent. In fact, the error rates shown for colour/contrast alone are even optimistic, in that colour maps are trained from ground truth segmentations whereas practically they would have to be trained adaptively from the imperfect segmentations obtained online. Note that while LDP and LGC conclusively achieve better performance than either colour/contrast or stereo alone, neither of LDP or LGC performs conclusively better than the other. An example of a segmented image from the AC sequence is shown in fig. 12 together with the spatial distribution of segmentation errors: the errors tend to cluster closely around object boundaries. Finally figure 13 shows two results corresponding to high error rates in the test data of figure 11. The first (VK) arises where the subjects hand approaches the frame of the image where stereo no longer operates because of occlusion by the image frame. The second (IU–JW), more interesting, shows slightly over-agressive action of the coherence constraint momentarily gluing two subjects together.

Background substitution in sequences.: Finally, figs. 14-16 demonstrate the application of segmentation to background replacement in video sequences (further results are available at [1]). 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 by border matting [27] as a post-process, though patch based priors could alternatively be used [19], [14]. The LGC algorithm gives good results, with blended boundaries and little visible flicker; LDP (not shown) gives very similar looking results.

VI. 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. We have demonstrated properties of the LDP and LGC algorithms and underlying model as follows.

- LDP and LGC are algorithms capable of fusing the two kinds of information, together with a coherence prior, with a substantial consequent improvement in segmentation accuracy.
- 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. The subjective effect of temporal artefacts is visible but not too obtrusive — see results movies [1]. Temporal artefacts in stereo can be alleviated by explicit temporal modelling and inference [34], but currently this is too expensive computationally for a real time system.

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 $U_k^{\mathrm{M}}(x_k, d_k)$ 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. Both the LDP and the LGC algorithms are capable of real time operation on a conventional processor. Fast implementations of DP techniques are well known [13], [16]. Ternary graph cut has been applied, in our laboratory, at around 1.5 M-pixels/second on a 3GHz Pentium desktop machine.

There are some other important differences between the algorithms. First, the LDP algorithm produces the entire stereo disparity map as a bi-product of segmentation, whereas LGC delivers the segmentation alone. This favours LDP in applications such as cyclopean view generation, for which the full disparity map is needed in addition to the occlusion map. Another interesting difference is that where the constraint figural continuity, captured by the contrast term of section II-E, makes only a marginal difference to LGC performance (figure 10), it profoundly improves the performance of LDP (details of experiments omitted). This may be because Dynamic Programming deals independently with each epipolar line, and the figural continuity constraint of [10] overcomes that limitation by providing an indirect but effective linkage between nearby epipolar lines.

Acknowledgements

The authors gratefully acknowledge helpful discussions with M. Isard, R. Szeliski and R. Zabih.

REFERENCES

[1] http://research.microsoft.com/vision/cambridge/i2i.
[2] http://cat.middlebury.edu/stereo/.
[3] H.H. Baker and T.O. Binford. Depth from edge and intensity based stereo. In *Proc. Int. Joint Conf. Artificial Intelligence*, pages 631–636, 1981.
[4] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 434–441, 1998.
[5] P.N. Belhumeur. A Bayesian approach to binocular stereopsis. *Int. J. Computer Vision*, 19(3):237–260, 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] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):401–406, 1998.
[8] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *Proc. European Conf. Computer Vision*, pages 428–441. Springer-Verlag, 2004.
[9] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, Cambridge, USA, 1987.
[10] Y.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*, pages CD–ROM, 2001.
[11] Y.Y. Boykov, O. Veksler, and R.D. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11), 2001.
[12] Y-Y Chuang, A. Agarwala, B. Curless, D.H. Salesin, and R. Szeliski. Video matting of complex scenes. In *Proc. Conf. Computer graphics and interactive techniques*, pages 243–248. ACM Press, 2002.
[13] I.J. Cox, S.L. Hingorani, and S.B. Rao. A maximum likelihood stereo algorithm. *Computer vision and image understanding*, 63(3):542–567, 1996.
[14] 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.
[15] A. Criminisi, J. Shotton, A. Blake, 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 Research Cambridge, 2003.
[16] A. Criminisi, J. Shotton, A. Blake, and P.H.S. Torr. Gaze manipulation for one to one teleconferencing. In *Proc. Int. Conf. on Computer Vision*, pages 191–198, 2003.
[17] A. Dempster, M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B.*, 39:1–38, 1977.
[18] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge University Press, 1998.
[19] A.W. Fitzgibbon, Y. Wexler, and A. Zisserman. Image-based rendering using image-based priors. In *Proc. Int. Conf. on Computer Vision*, 2003.
[20] D. Geiger, B. Ladendorf, and A.Yuille. Occlusions and binocular stereo. *Int. J. Computer Vision*, 14:211–226, 1995.
[21] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.
[22] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2001.
[23] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proc. European Conf. Computer Vision*, pages CD–ROM, 2002.
[24] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. Int. Joint Conf. Artificial Intelligence*, pages 674–679, 1981.
[25] Y. Ohta and T. Kanade. Stereo by intra- and inter-scan line search using dynamic programming. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 7(2):139–154, 1985.
[26] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.
[27] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.
[28] S.M. Rowe and A. Blake. Statistical mosaics for tracking. *J. Image and Vision Computing*, 14:549–564, 1996.

[29] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vision*, 47(1–3):7–42, 2002.

[30] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 246–252, 1999.

[31] 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, 2001.

[32] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proc. Int. Conf. on Computer Vision*, pages 255–261, 1999.

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

[34] O. Williams, M. Isard, and J. MacCormick. Estimating disparity and occlusions in stereo video sequences. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages CD–ROM, 2005.

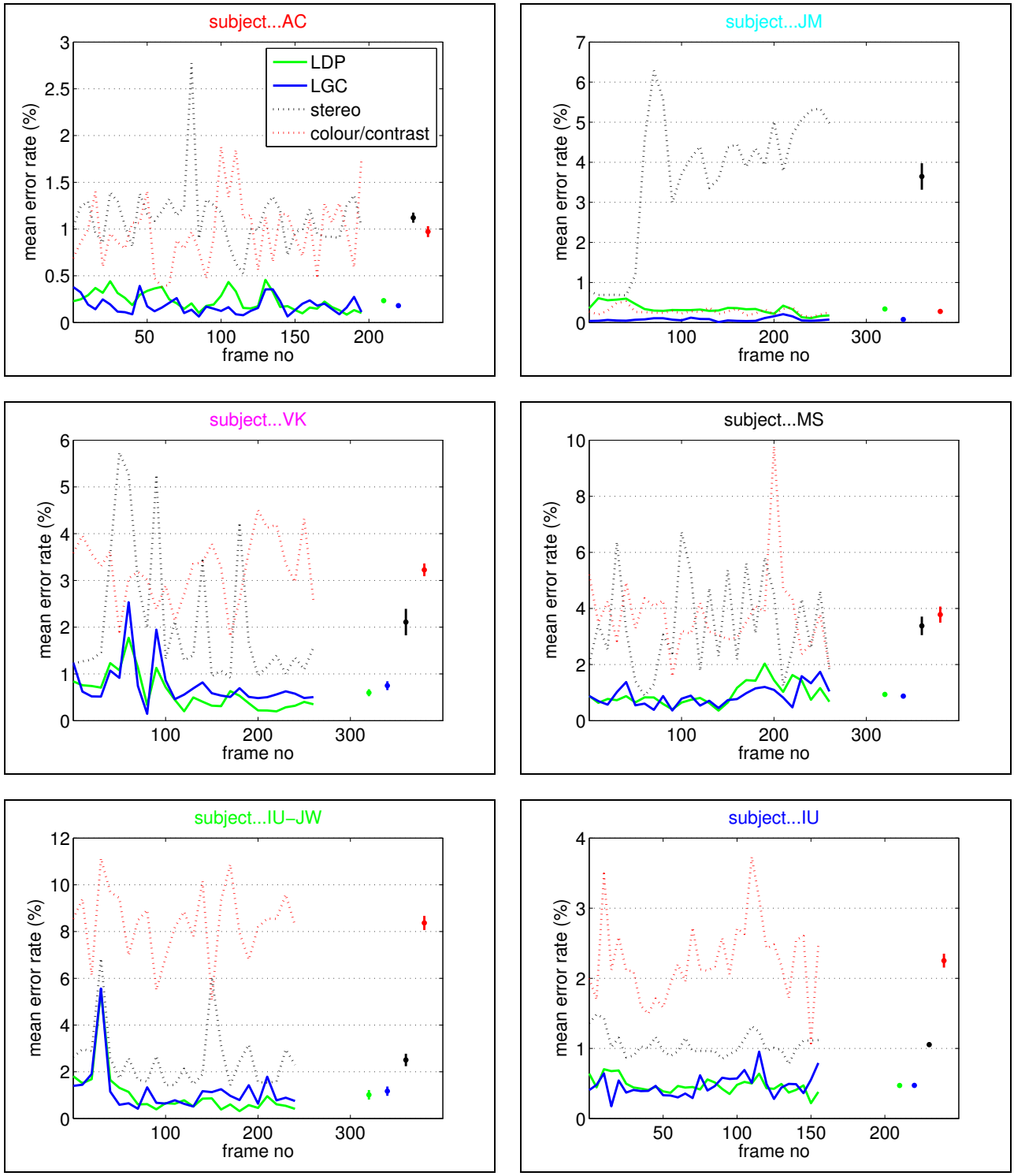
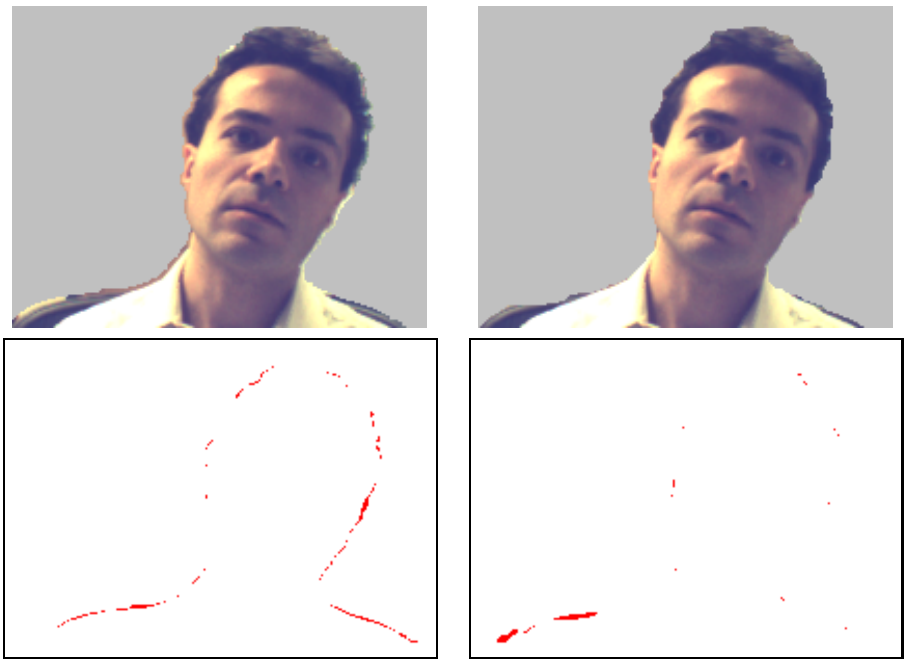


Fig. 11. **Segmentation performance advantage from fusion.** Segmentation error (percentage of misclassified pixels) is computed on all six sequences, frame by frame, for LDP, LGC, colour only and stereo only. Error bars are also shown, on the right of each plot, for temporal mean and standard error. Note that fused stereo and colour/contrast (LGC and LDP) perform substantially better than either stereo or colour/contrast alone.

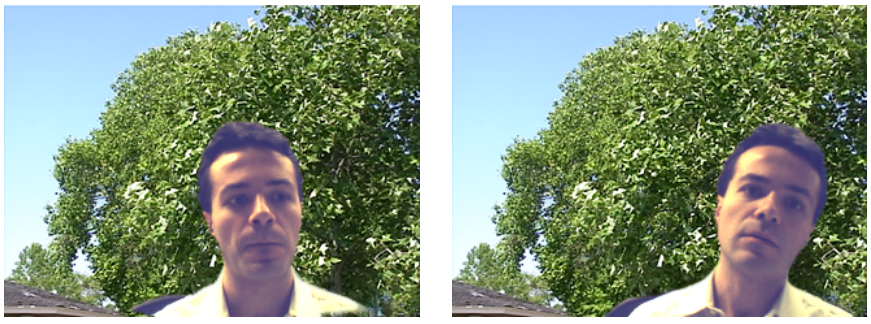


LDP *LGC*

Fig. 12. **Extracted foreground layer** (top) for the left view of sequence AC, frame 100, for LGC and LDP. Segmentation error maps (bottom).

subject VK frame 61 subject IU–JW frame 31

Fig. 13. **LGC Segmentation error illustrations.** We show here two results corresponding to high error rates in the test data of figure 11. Segmented foreground is shown highlighted.



LGC, frame 0 *LGC, frame 100*

Fig. 14. **Segmentation and background substitution.** Here we show background substitution (using LGC) for two frames of the sequence AC.

Fig. 15. **Segmentation with non-stationary background.** (**Top**) Four frames of the input left sequence sequence `IU` (right frame not shown here). (**Bottom**) Corresponding LGC segmentation and background substitution. LDP performs similarly. Note the robustness of the segmentation to motion in the original background.

Fig. 16. **Non-stationary background with more complex foreground.** A final example of segmentation and background substitution (test sequence `S3`). (**Top**) Input left images. A third person is moving in the original background. (**Bottom**) LGC background-substitution.