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
www.research.microsoft.com/vision/cambridge

ABSTRACT

This paper describes models and algorithms for the real-time segmentation of foreground from background layers in stereo video sequences. Automatic separation of layers from color/contrast or from stereo alone is known to be error-prone. Here, color, 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 color 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 marginalized over disparities to evaluate foreground and background hypotheses, and then fused with a contrast-sensitive color 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 performance, substantially better than either stereo or color/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. The assumption is that the visible scene can be expressed as two, spatially coherent layers, one a "foreground" layer masking the other "background" layer. 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 [2], but with computational efficiency sufficient to attain live streaming speed.

Layer extraction from images has long been an active area of research [3], [4], [5], [6], [7]. The challenge addressed here is to segment the foreground layer both accurately and efficiently. Conventional stereo algorithms e.g. [8], [9] have proven competent at computing depth. Stereo occlusion is a further cue that needs to be accurately computed [10], [11], [12], [13], [14] 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 color/contrast-based segmentation techniques have been demonstrated to be very effective [15], [16], even in the absence of texture. Segmentation based on color/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 color 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 [15], [17]. 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, color and contrast to achieve segmentation quality unattainable from either stereo or color/contrast on their own (see figure 2). This claim is verified by evaluation on stereo videos with respect to ground truth (section VI). Finally, efficient post-processing for matting [18] is applied to obtain good video quality as illustrated in stills in this paper, and companion videos [1].

The paper is organized as follows. In sections II and III we describe the common components of the probabilistic models for LDP and LGC. In sections IV and V we describe LDP and LGC algorithms, respectively. Experimental results and conclusions are presented in sections VI and VII.

II. PROBABILISTIC MODELS FOR BI-LAYER SEGMENTATION OF STEREO IMAGES

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

A. Notation

Pixels in the rectified left and right images are indexed by m and n respectively, so the images are denoted

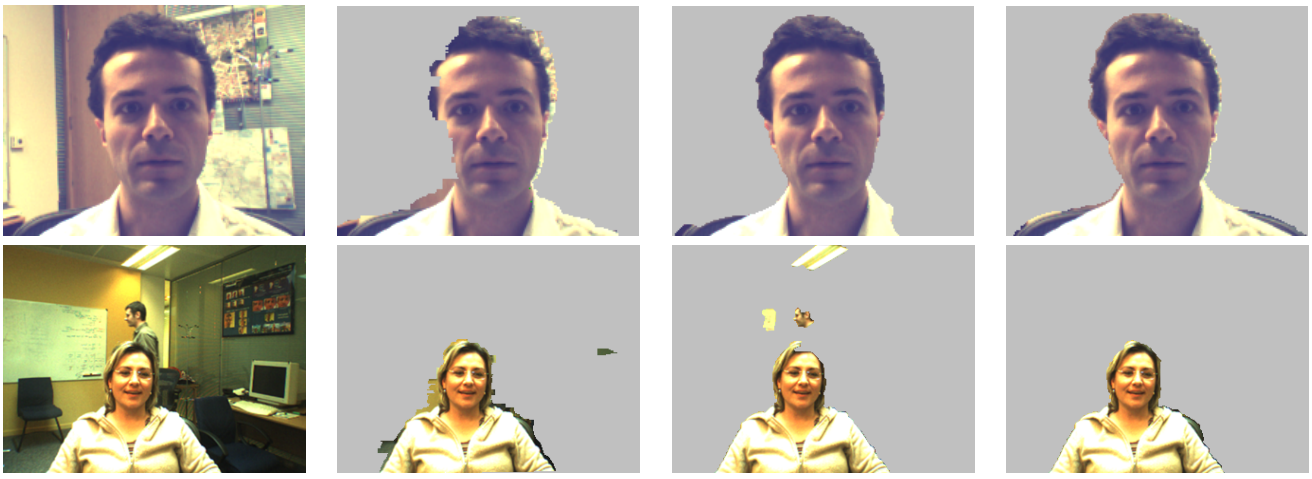
$$\mathbf{L} = \{L_m,\ m = 1, \ldots, N\},\ \mathbf{R} = \{R_n,\ n = 1, \ldots, N\}.$$

We refer jointly to the data as $\mathbf{z} = (\mathbf{L}, \mathbf{R})$. In addition an array $\mathbf{x}$ of state variables is defined, either in left-image coordinates $\mathbf{x} = \{x_m\}$, or, in cyclopean coordinates, as $\mathbf{x} = \{x_k\}$, and 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. Stereo disparity is defined to be $d = m - n$ and the disparity values along one epipolar line are expressed as $\mathbf{d} = \{d_k,\ k = 1, \ldots, 2N - 1\}$. Note this means that

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

input left view *input right view* *automatic layer separation and background substitution*

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.



a) **b)** **c)** **d)**

Fig. 2. **Segmentation by fusing color, 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 [13]; **c)** Segmentation based on color/contrast [16]; **d)** The LGC algorithm proposed here fuses color, contrast and stereo to achieve a more accurate segmentation. The foreground artefacts visible in b) and c) are corrected in d).

so that k, d forms an alternative *cyclopean* coordinate system for the space of epipolar matches, which is well known to be helpful for probabilistic modeling of stereo matching [11]. For good reasons (see later) the two algorithms presented in this paper are each based on different image coordinate systems, one on cyclopean coordinates (k, d), the other on left image-based coordinates (m, d).

This sets up the notation for a complete match of two images as the combined vector $(\mathbf{d}, \mathbf{x})$ of disparities and states. Now a posterior distribution over $(\mathbf{d}, \mathbf{x})$, conditioned on image data, can be defined.

B. Generative model

A Gibbs energy $E(\mathbf{z}, \mathbf{d}, \mathbf{x}; \theta)$ is defined to specify the posterior over the inferred sequence $(\mathbf{d}, \mathbf{x})$, given the image data $\mathbf{z}$, as:

$$p(\mathbf{x}, \mathbf{d} \mid \mathbf{z}) \propto \exp -E(\mathbf{z}, \mathbf{d}, \mathbf{x}; \theta). \tag{2}$$

Here θ is a vector of parameters for the model, which will need to be set according to their relation to physical quantities in the stereo problem, and by learning from labeled data. The posterior could be globally maximised to obtain a segmentation $\mathbf{x}$ and also stereo disparities $\mathbf{d}$. In this paper, the aim is simply to compute a segmentation, in which case the posterior should, in principle, be marginalised with respect to $\mathbf{d}$, and then maximised with respect to $\mathbf{x}$ to estimate a segmentation

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \sum_{\mathbf{d}} p(\mathbf{x}, \mathbf{d} \mid \mathbf{z}). \tag{3}$$

The model (2) can be regarded simply as a Conditional Random Field (CRF) [19], without any generative explanation/decomposition in terms of priors over $(\mathbf{x}, \mathbf{d})$ and data likelihoods. However, simpler forms of the model do admit a generative decomposition, and this is very helpful also in motivating the structure of a fuller CRF model that is not so naturally decomposed. One reasonable generative model has a Gibbs energy with the following decomposition:

$$E(\mathbf{z}, \mathbf{x}, \mathbf{d}; \theta) = V(\mathbf{x}, \mathbf{d}; \theta) + U^{\mathrm{M}}(\mathbf{z}, \mathbf{x}, \mathbf{d}; \theta) + U^{\mathrm{C}}(\mathbf{z} \mid \mathbf{x}; \theta), \tag{4}$$

in which the role of each of the three terms is as follows.

Prior:: an MRF prior for $(\mathbf{x}, \mathbf{d})$ has an energy specified as a sum of unary and pairwise potentials:

$$V(\mathbf{x}, \mathbf{d}; \theta) = \sum_{(k,k') \in \mathcal{N}} [F(x_k, x_{k'}, \Delta d_k, \Delta d_{k'})] + \sum_k G_k(x_k, d_k), \tag{5}$$

where $\Delta\mathbf{d}$ is the *disparity gradient* along epipolar lines, that is

$$\Delta d_k = d_k - d_{k-1}. \tag{6}$$

Typically, $F(\ldots)$ discourages excessive disparity gradient within matched regions. Pixel pairs $(k, k') \in \mathcal{N}$ are the ones that are deemed to be neighbouring in the pixel grid. The first component $F(\ldots)$ of the prior Gibbs energy V in (5) should incorporate an Ising component that favours coherence in the segmentation variables $x_k, x_{k'}$. It should also favour continuity of disparity over matched regions, and do so anisotropically — more strongly along epipolar lines than across them. The $G_k(\ldots)$ term implements "disparity-pull", the tendency of foreground elements to have higher disparity than background ones. The specific form of $G_k(\ldots)$ can be set by taking

$$G_k(x_k, d_k) = -\log p(d_k \mid x_k), \tag{7}$$

and determining the conditional density $p(d_k \mid x_k)$ from the observed statistics of some labelled data. Various models could be used here, but in our experiments a simple, constant disparity, separating surface is used, so that $d > d_0$ characterises foreground, with uniform distributions for $p(d_k \mid x_k)$ over each of the possible states $x \in \{\mathrm{F}, \mathrm{B}, \mathrm{O}\}$.

Stereo likelihood: , represented by the U^{M} term, evaluates the stereo-match evidence in the data $\mathbf{z}$, both to distinguish occlusion ($x_k = \mathrm{O}$) from full visibility ($x_k \in \{\mathrm{F}, \mathrm{B}\}$) and, given visibility, to determine disparity d_k.

Color likelihood: , represented by the U^{C} term, uses probability densities in colour space, one density for the background and another for the foreground, to apply evidence from pixel colour to the segmentation x_k of each pixel.

C. Contrast dependence

One further elaboration, due to Boykov and Jolly [15], incorporates the evidence from image contrast for segmentation — see also "line processes" [20], "weak constraints" [21] and "anisotropic diffusion" [22]. It proves important in refining segmentation quality, at the cost of obscuring somewhat the clear generative distinction between prior and likelihood [23]. The Ising component F in (5) is made contrast dependent, disabling the penalty for breaking coherence in $\mathbf{x}$ wherever image contrast is high. Segmentation boundaries tend, as a result, to align with contours of high contrast. The MRF model (4) is extended in this way to a CRF

$$E(\mathbf{z}, \mathbf{x}, \mathbf{d}; \theta) = V(\mathbf{z}, \mathbf{x}, \mathbf{d}; \theta) + U^{\mathrm{M}}(\mathbf{z}, \mathbf{x}, \mathbf{d}; \theta) + U^{\mathrm{C}}(\mathbf{z} \mid \mathbf{x}; \theta), \tag{8}$$

in which dependence on data $\mathbf{z}$ is now incorporated in to the $V(\ldots)$ term.

D. Tractability of inference and learning

The stated inference problem (3) for segmentation, is intractable with the Gibbs energy model (8) above. A related problem,

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \left(\max_{\mathbf{d}} p(\mathbf{x}, \mathbf{d} \mid \mathbf{z}) \right), \tag{9}$$

while not formally tractable, could be regarded as tractable in practice because it can be solved approximately by the α-expansion form of graph-cut [17], over the variables $\mathbf{x}, \mathbf{d}$ jointly (provided the energy function E is chosen to meet the necessary regularity conditions). The approximation (9) to the original problem is likely to be a good one, because the posterior density is likely to be sharply peaked with respect to $\mathbf{d}$, since stereo constraints on disparity are typically strong. However α-expansion over $(\mathbf{x}, \mathbf{d})$ jointly would be rather inefficient, at least an order of magnitude below real-time, for current architectures. This paper proposes two approaches to simplifying the Gibbs energy model, to make inference of segmentation $\mathbf{x}$ practically tractable and efficient.

LDP. In Layered Dynamic Programming, all vertical cliques in V (5) are removed, resulting in a posterior density consisting simply of a set of one-dimensional Hidden Markov Models (HMMs), one HMM along each epipolar line. For the disparity-gradient dependence in V, this means retaining the strong epipolar constraints, but omitting figural continuity constraints, which are weaker. For the segmentation coherence encouraged by V, constraints can be imposed only horizontally, and the vertical constraint is lost. Nonetheless there is some implicit transfer of information vertically via the overlap of the patches used in the stereo match likelihood (see section III-A and also [14]). In exchange for the lost vertical constraint, the max-max (9) form of the problem becomes exactly tractable by dynamic programming. Not only that, but because the prior energy V has become a Markov chain, the parameter learning problem also becomes tractable.

LGC. In Layered Graph Cut, the prior term $F(\ldots)$ in (5) is made independent of disparity $\mathbf{d}$. Now the posterior density can be marginalised exactly over $\mathbf{d}$ in the original inference problem (3). Marginalization gives the posterior density $p(\mathbf{x} \mid \mathbf{z})$ for segmentation only, which can be maximised by ternary graph-cut, using α-expansion. Parameter learning has not been made tractable, but some guidance comes from priors and likelihoods estimated for LDP, transplanted (and simplified) to the LGC model.

In summary, we have two approximate models for the original problem. One, LDP, has the advantage of practical tractability not only for inference but also for parameter learning. It has the disadvantage though that vertical constraints have been neglected. On the other hand LGC retains vertical constraints at least for segmentation, but neglects all direct constraints on continuity of disparity. It has the advantage of solving the original max-sum form of the inference problem, rather than just the max-max approximation, but the disavantage that parameter estimation remains intractable.

III. Probabilistic modelling of stereo, colour and contrast

In this section we describe the likelihood functions for each type of image cue, which are then combined in the model, giving the effect of cue fusion in inference.

A. Likelihood for stereo

The stereo-matching energy $U^{\mathrm{M}}(\mathbf{z},\mathbf{x},\mathbf{d};\theta)$ from (8) is modelled as a sum over pixels

$$U^{\mathrm{M}}(\mathbf{z},\mathbf{x},\mathbf{d}) = \sum_k U_k^{\mathrm{M}}(\mathbf{z},x_k,d_k) \tag{10}$$

where each U_k^{M} is the cost associated with the stereo match at pixel k. Commonly [24] stereo matches are scored using SSD (sum-squared difference), that is the L^2-norm of difference between image patches $L_m^{\mathrm{P}},\ R_n^{\mathrm{P}}$, surrounding hypothetically matched pixels m,n. Following [13] we model U_k^{M} in terms of SSD but with additive and multiplicative normalization for robustness to non-Lambertian effects and photometric calibration error. This is termed NSSD — normalized SSD:

$$U_k^{\mathrm{M}}(\mathbf{z},x_k,d_k) = \begin{cases} M(L_m^{\mathrm{P}},\ R_n^{\mathrm{P}}) & \text{if } x_k \in \{\mathrm{F},\mathrm{B}\} \\ 0 & \text{if } x_k = \mathrm{O}, \end{cases} \tag{11}$$

where $M = \lambda(N - N_0)$ with λ, N_0 being constants, and $m,n = (k \pm d_k)/2$ the left and right image indices, as before (1). The NSSD N is:

$$N(L^{\mathrm{P}},\ R^{\mathrm{P}}) = \frac{1}{2}\frac{\|(L^{\mathrm{P}} - \overline{L^{\mathrm{P}}}) - (R^{\mathrm{P}} - \overline{R^{\mathrm{P}}})\|^2}{\|L^{\mathrm{P}} - \overline{L^{\mathrm{P}}}\|^2 + \|R^{\mathrm{P}} - \overline{R^{\mathrm{P}}}\|^2} \in [0,1], \tag{12}$$

in which $\overline{R^{\mathrm{P}}}$ denotes the mean value over the patch R^{P}. The constant N_0 can be thought of as a penalty for failure to match. Further details of modelling and parameter estimation for the stereo likelihood are given in appendix A., and this leads to statistical estimators for the constants λ and N_0.

B. Likelihood for color

Following previous approaches to two-layer segmentation [15], [16] we model likelihoods for color in foreground and background using Gaussian mixtures in RGB color space, learned from image frames, labeled (automatically), from earlier in the sequence. The foreground color 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 color model is then given by an energy U_k^C:

$$U_k^C(z_k,x_k) = \begin{cases} -\log p^{\mathrm{F}}(z_k) & \text{if } x_k = \mathrm{F} \\ -\log p^{\mathrm{B}}(z_k) & \text{if } x_k = \mathrm{B} \text{ or } x = \mathrm{O} \end{cases} \tag{13}$$

Learning of the global foreground and background color models p^{F} and p^{B} proceeds as follows. Each is a mixture of $N_C = 20$ full covariance Gaussian components in RGB color-space, and is learned, at each video timestep, using 10 iterations of EM [25], initialized 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 color 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 color is taken as:

$$U^{\mathrm{C}}(\mathbf{z},\mathbf{x};\theta) = \rho \sum_k U_k^{\mathrm{C}}(z_k,x_k) \tag{14}$$

where the *color discount* constant ρ (typical value $\rho = 1/2$) is included to tune the balance of influence between the stereo model and the color model. In principle, the generative derivation of the energies should have balanced them already. In practice, the pixelwise independence assumptions built in to the color model render the influence of color excessively strong, and choosing a value $\rho < 1$ discounts for that. Color models are initialized at time $t = 0$, by setting $\rho = 0$, estimating segmentation without using color, and using the labelled segments to learn the foreground and background color models for $t = 0$. Note that for working in the cyclopean frame, separate foreground and background color models are maintained for each of the left and right images.

C. Contrast dependence

To implement the contrast dependence described above, a soft switch for the Ising penalty is defined, replacing $F(\ldots)$ in (5) by

$$F(x_k,x_{k'},\Delta d_k,\Delta d_{k'},V^*_{k,k'}), \tag{15}$$

where $V^*_{k,k'}$ is the soft contrast switch applying across sites k,k':

$$V^*_{k,k'} = \frac{1}{1+\epsilon}\left(\epsilon + \exp - \frac{\|g_k - g_{k'}\|^2}{2\sigma^2 {d_{k,k'}}^2}\right). \tag{16}$$

Here $\mathbf{g}$ is the image-data, Gaussian smoothed at a scale of 0.7 pixels and with components g_k at each pixel; $d_{k,k'}$ is the Euclidean distance between pixels k,k' and $\sigma^2 = \left\langle \|g_k - g_{k'}\|^2 / {d_{k,k'}}^2 \right\rangle$, a mean contrast over all neighboring pairs of image pixels. The factor $V^*_{k,k'}$ acts as a soft contrast switch, and is typically allowed to multiply certain of the costs in $F(\ldots)$, so that built-in tendencies to coherence are abated in the presence of high contrast. Details are given in the next two sections. [The constant ϵ is a "dilution" constant for contrast, empirically found to be best set to $\epsilon = 1$.]

To summarise, the final CRF model is as in (8), with

$$V(\mathbf{z},\mathbf{x},\mathbf{d};\theta) = \sum_{(k,k')\in\mathcal{N}} F(x_k,x_{k'},\Delta d_k,\Delta d_{k'},V^*_{k,k'}) + \sum_k G_k(x_k,d_k), \tag{17}$$

incorporating the contrast sensitivity, via V^*, as required.

D. Choice of image coordinate frame

Finally, we promised at the start to comment on the choice of coordinate frame, cyclopean for LDP and left for LGC. The reason is that cyclopean is intrinsically preferable, not only for reasons of symmetry and elegance, but also because occlusions occur on *both* sides of a foreground object (not just one side, as with left image coordinates), and this gives additional constraint for segmenting the foreground. In LGC however, the cyclopean image is not accessible because marginalisation hides the disparities. Thus image contrast, for the contrast-sensitivity term, has to be computed from a physical image, eg the left.

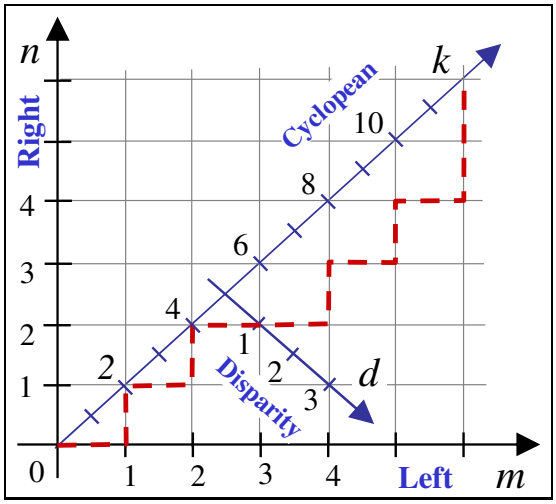


Fig. 3. **Stereo match-space.** Notation conventions for left and right epipolar lines with pixel coordinates m, n, cyclopean coordinates k and stereo disparity $d = m - n$. Hypothetical matching path shown dashed (*cf.* [11], [9]).

IV. Layered Dynamic Programming (LDP)

The model used in LDP, as mentioned earlier, is the general stereo CRF model (8) with energy $E(\mathbf{z}, \mathbf{x}, \mathbf{d}; \theta)$ from section II, but with all vertical constraints removed. All optimization therefore takes place independently, within individual scanlines. In this one-dimensional situation, the Gibbs energy specification is equivalent to specifying a Hidden Markov Model (HMM) on $(\mathbf{x}, \mathbf{d})$ along each scanline. As usual for an HMM, the prior energy V (here also switched softly by contrast) is expressed as a Markov chain over x_k, d_k given x_{k-1}, d_{k-1}. Observation likelihoods, U^{M} for stereo and U^{C} for colour, are expressed as emission costs, as is standard for an HMM. In this section we first set out the notation for the HMM on a scanline, and then give details of how the various energies are represented in the model, all finally summarised in a state-transition diagram for the HMM.

A. Optimal matching path along a scanline

Left pixels L_m and right pixels R_n, on a given scanline of length N_{S} pixels, are ordered by any particular matching path (figure 3), giving $2N$ cyclopean pixels

$$\mathbf{z} = \{z_k,\ k = 1, \ldots, 2N_{\mathrm{S}}\},$$

where $k = m + n$. The k-axis is the so-called cyclopean[1] coordinate axis. Conventionally in DP stereo matching the "ordering constraint" [26], [8] is imposed, and this means that each move in figure 3 is allowed only in the positive (North-to-East) quadrant of the diagram. Stereo disparity along the cyclopean epipolar line is $\mathbf{d} = \{d_k,\ k = 1, \ldots, 2N_{\mathrm{S}} - 1\}$ where $d_k = m - n$.

Stepwise restriction for LDP: Previous matching algorithms, e.g. [9], [27], have allowed multiple and/or diagonal moves on the stereo matching paths (fig 3). Here the problem differs significantly. In [9], [27] 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 figure 3, 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 figure 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 a z_k in a $p(z_k \mid \ldots)$ term, in the joint likelihood $\prod_k p(z_k \mid x_k, d_k, z_1, \ldots, z_{k-1})$ for the scanline. This makes for a consistent definition of the likelihood.

B. LDP: stereo with occlusion and layers

The three possible states $x_k \in \{\mathrm{F}, \mathrm{B}, \mathrm{O}\}$ are doubled up, for convenience, to reflect the existence of *left* and *right* variants, respectively the horizontal and vertical moves in figure 3. This gives 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 HMM for the Gibbs model is then reflected in the state-space diagram of figure 4, which represents Markov chain transitions $k-1 \rightarrow k$, in terms of costs (*ie* energy increments) on arcs, and these capture the contrast-modified prior energy V. Observation likelihood energies are represented by the costs U^{M}_k and U^{C}_k on nodes. [Note that left-occluding and right-occluding states cannot directly intercommunicate, reflecting constraints of stereo geometry.]

Prior and contrast: Transition energies between occluding and foreground states represent the component $F(\ldots)$ of the prior energy V (17), and incorporate the soft contrast switch V^* defined earlier (16). (In this cyclopean setting, V^* must be computed from contrast in the left or the right image, according to whether the state is left-foreground or right-foreground.)

The model has a number of parameters $\{a_F, a_B, a_O, b_F, b_B, b_O, c_F, c_B\}$. It might seem problematic that so many parameters need to be set, but in fact they can be learned from labeled training frames as follows:

$$b_{\mathrm{O}} = \log(2W_{\mathrm{O}}) \quad b_{\mathrm{F}} = \log(W_{\mathrm{F}}) \quad b_{\mathrm{B}} = \log(W_{\mathrm{B}}) \qquad (18)$$

where W_{O}, W_{F} and W_{B} are the mean widths of occluded, foreground and background regions respectively. This follows simply from the fact that $2\exp -b_0$ is the probability of escape from an occluded state, and so on. Then consideration of viewing geometry together with an assumption about typical

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

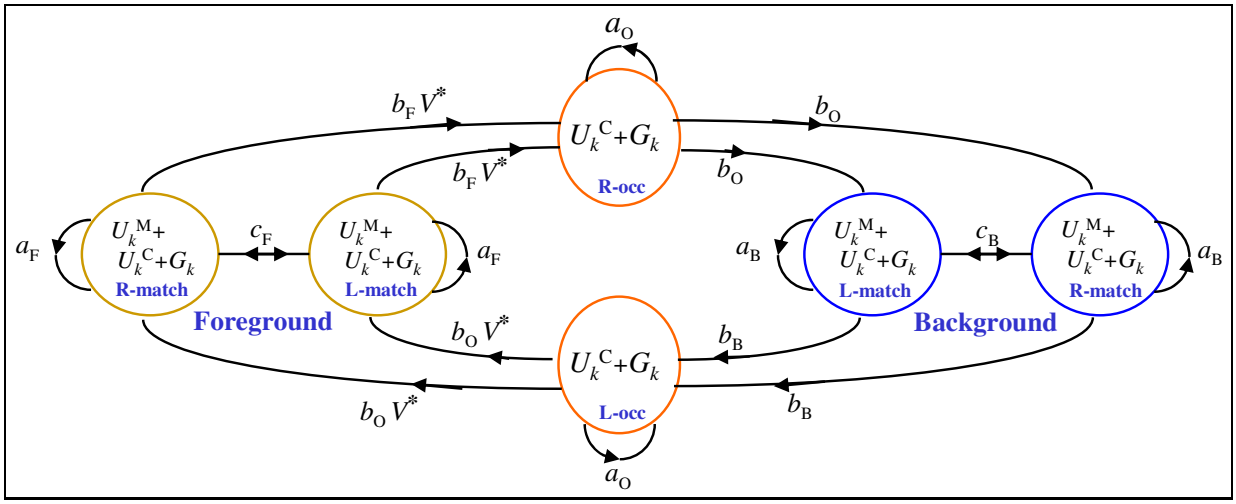


Fig. 4. **State space for foreground/background segmentation.** The segmentation state space $x_k \in \{\mathrm{F}, \mathrm{B}, \mathrm{O}\}$ is doubled to take account of left and right variants; hence matched and occluded states together form a 6-state system. Note that from the foreground states (yellow circles), only the *right* occluding state is accessible, and from background (blue circles) only the *left* occluding state; this reflects a simplification of the model to exclude the possibility of foreground/foreground occlusion. Match costs incorporate disparity-pull and contrast effects — see text for details.

slopes of visible surfaces (see appendix B. for details) indicates that:

$$a_{\mathrm{F}} = \log(1 + D/B) - \log(1 - 1/W_{\mathrm{F}}), \tag{19}$$

where D is a nominal distance to objects in the scene and B is the interocular distance (camera baseline), and similarly for a^{B}. Lastly, probabilistic normalization demands that

$$a_{\mathrm{O}} = -\log(1 - 2e^{-b_{\mathrm{O}}}), \tag{20}$$

and that

$$c_{\mathrm{F}} = -\log(1 - e^{-b_{\mathrm{F}}} - e^{-a_{\mathrm{F}}}) \tag{21}$$

and similarly for c_{B}, and so all the parameters are fixed.

Disparity-pull: The disparity-pull term $G_k(\ldots)$ in the prior (17) is implemented in the transition-diagram as a cost applied at each node, as shown.

Stereo and color fusion: Likelihood costs for stereo and for color are U_k^{M} and U_k^{C}, as described earlier in section III. They appear as nodal costs on the state transition diagram.

The 6-state HMM can be optimized straightforwardly by dynamic programming and this gives a solution to the alternative "max-max" estimation problem (9) described at the end of section II. Results are given later in section VI.

V. 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 stereo disparity $\mathbf{d}$ does not appear explicitly. The energy function is defined from the CRF (8) for the full stereo problem in section II, by marginalizing over disparity, to give a posterior distribution $p(\mathbf{x} \mid \mathbf{z})$. Segmentation becomes a ternary optimization problem, over the three labels $\mathrm{O}, \mathrm{F}, \mathrm{B}$ on the x-values at each pixel, which can be solved (approximately) by iterative application of a binary graph-cut algorithm — so-called α-expansion [17].

As explained earlier, LGC is expressed in the coordinate frame of one (*e.g.* left) image, rather than in the cyclopean frame as in LDP. Hence image related variables such as x_m carry the left image index m, rather than the cyclopean k used earlier.

A. Marginalized energy

In order for the marginalization to be tractable, the energy V (17) is simplified by neglecting explicit disparity dependence in $F(\ldots)$, that is, assuming that:

$$F(x_m, x_{m'}, \Delta d_m, \Delta d_{m'}, V^*_{m,m'}) = F(x_m, x_{m'}, V^*_{m,m'}). \tag{22}$$

Now the marginalized posterior is defined by its energy

$$E(\mathbf{z}, \mathbf{x}; \theta) = V(\mathbf{z}, \mathbf{x}; \theta) + H(\mathbf{z}, \mathbf{x}, \theta) + U^{\mathrm{C}}(\mathbf{z}, \mathbf{x}; \theta), \tag{23}$$

where

$$V(\mathbf{z}, \mathbf{x}; \theta) = \sum_{(m,m') \in \mathcal{N}} F(x_m, x_{m'}, V^*_{m,m'}) \tag{24}$$

is derived from a simplified prior, with added soft contrast switching as earlier (15). The color likelihood U^{C} is unchanged from the earlier discussion, except now referred entirely to the left image. Finally, the new term H in (23) is a sum over pixels:

$$H(\mathbf{z}, \mathbf{x}) = \sum_m H_m(x_m) \tag{25}$$

where H_m is defined by marginalization over disparity to be:

$$H_m(x_m) = \tag{26}$$
$$-\log\left[\sum_{d_m} \exp -\left\{G_m(x_m, d_m) + U_m^{\mathrm{M}}(\mathbf{z}, x_m, d_m)\right\}\right].$$

Note that, from (7) and (11), this definition has the property that H_m is normalised such that $H_m(\mathrm{O}) = 0$.

B. Coherence and contrast

The coherence/contrast costs (24) for the LGC model are defined to be

$$F(x_m, x_{m'}, V^*_{m,m'}) = F_{m,m'} V^*_{m,m'} \tag{27}$$

where again $V^*_{m,m'}$ is the soft contrast switch. Anisotropic coherence costs $F_{m,m'}$ are defined as follows. Cliques consist of horizontal, vertical and diagonal neighbors on the square grid of pixels. For vertical and diagonal cliques $F_{m,m'}$ acts as a switch triggered by transitions 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 (epipolar) 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_F and b_O in the LDP model, so a reasonable working value for γ is

$$\gamma = \frac{1}{2}(b_F + b_\mathrm{O}) = \log(2\sqrt{W_\mathrm{F} W_\mathrm{O}}), \tag{28}$$

where width parameters W_F and W_O were defined earlier (18).

C. Expansion move algorithm

Currently, graph cut based stereo algorithms techniques such as [15], [12] are not suited for real-time implementation. The main reason is that they perform $O(d_{max})$ α-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 figure 5). 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. Furthermore, 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 (figure 5b).

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

VI. Results

Performance of the LGC and LDP algorithms was evaluated with respect to ground-truth segmentations on every fifth or tenth frame (left view), in each of six test stereo sequences[2]. The data was labeled 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 IV, equations (18) and (19), with the same values for foreground and background parameters, *i.e.* a_F and a_B *etc*. Region widths in equations (19) and (18) are set to $W_\mathrm{O} = 10$ pixels and $W_\mathrm{F} = W_\mathrm{B} = 100$ 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

Experiments are shown here on The first set of experiments, with the LGC algorithm, are shown in figure 6. 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 summarized for each parameter in turn.

Likelihood offset parameter N_0, introduced in section III-A, 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 III-A. The value $N_0 = 0.35$ is very slightly superior discriminatively — *i.e.* it gives lower error rate in figure 6.

Coherence constant γ for LGC, defined in section V, 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 color/contrast only [16]. Presumably the presence of the additional cue from stereo to some extent takes over the role of coherence. The default value, from equation (28) in section V, and taking $W_\mathrm{O} = 10$ pixels and $W_\mathrm{M} = 100$ pixels as before, gives $\gamma = 3.8$ which is satisfactorily consistent with the experimental results.

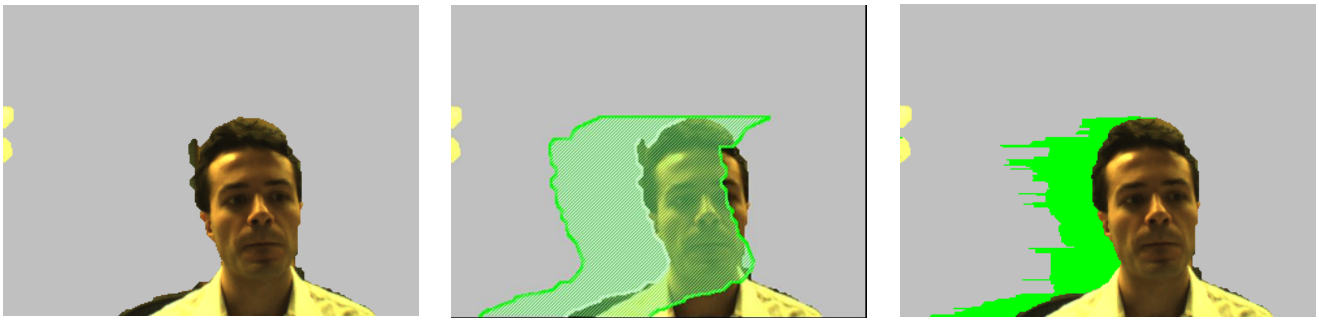
Color discount constant ρ, defined in section III-B equation (14), gives best error rates around $\rho = 0.5$. Without a discount ($\rho = 1$) error rates are appreciably higher, and this confirms the need for a discount to modify the generative assumption of independence of color at neighboring pixels.

Contrast parameter ϵ, defined in section III-C, equation (16) 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* [16].

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 III-B, mixing in a probability density learned, for each pixel, by pixelwise background maintenance [28], [29], [30]. The learned pixelwise densities $p^\mathrm{B}_k(z_k)$ are typically strongly peaked, and

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



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

Fig. 5. **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.)

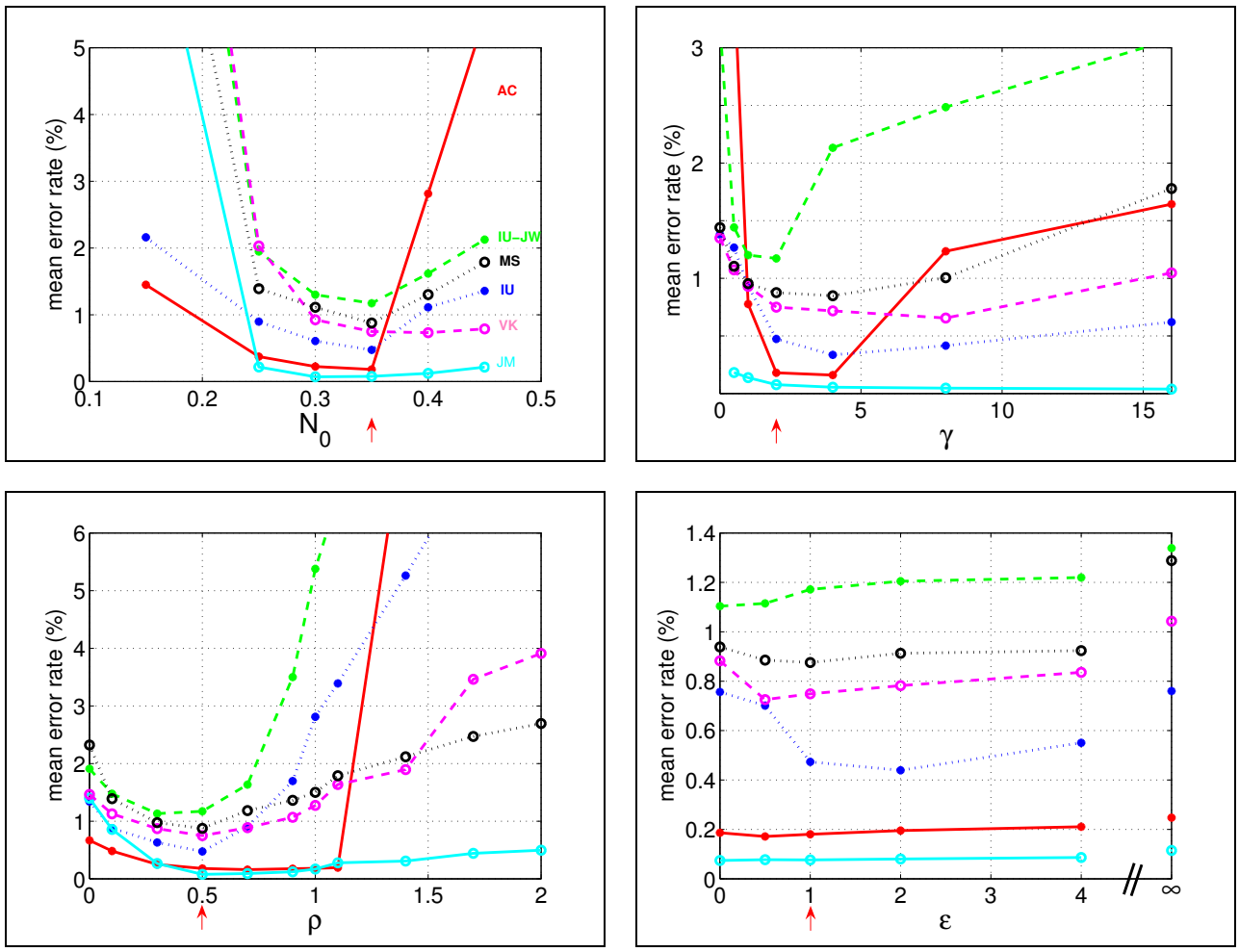


Fig. 6. **Effect of values of LGC parameters** N_0, γ, ρ and ϵ on segmentation error rate, for each of 6 test-data sets — see text for detailed discussion. The default value of each parameter is indicated by an arrow on the abscissa axis.

hence very informative, but sensitive to movement in the background. That sensitivity is robustified by adding in the general background distribution $p^{\mathrm{B}}(z_k)$ 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 $p_k^{\mathrm{B}}(z_k)$ and $p^{\mathrm{B}}(z_k)$ decreased error rates by just 0.03–0.3% across the 6 data sets tested (see section VI), compared with using $p^{\mathrm{B}}(z_k)$ alone. Note however that using the pixelwise $p_k^{\mathrm{B}}(z_k)$ alone, without any $p^{\mathrm{B}}(z_k)$ component, increased error rates by a disastrous $0.5 - 8.1\%$. That is in addition to the disadvantage that pixelwise background models are sensitive to camera shake.

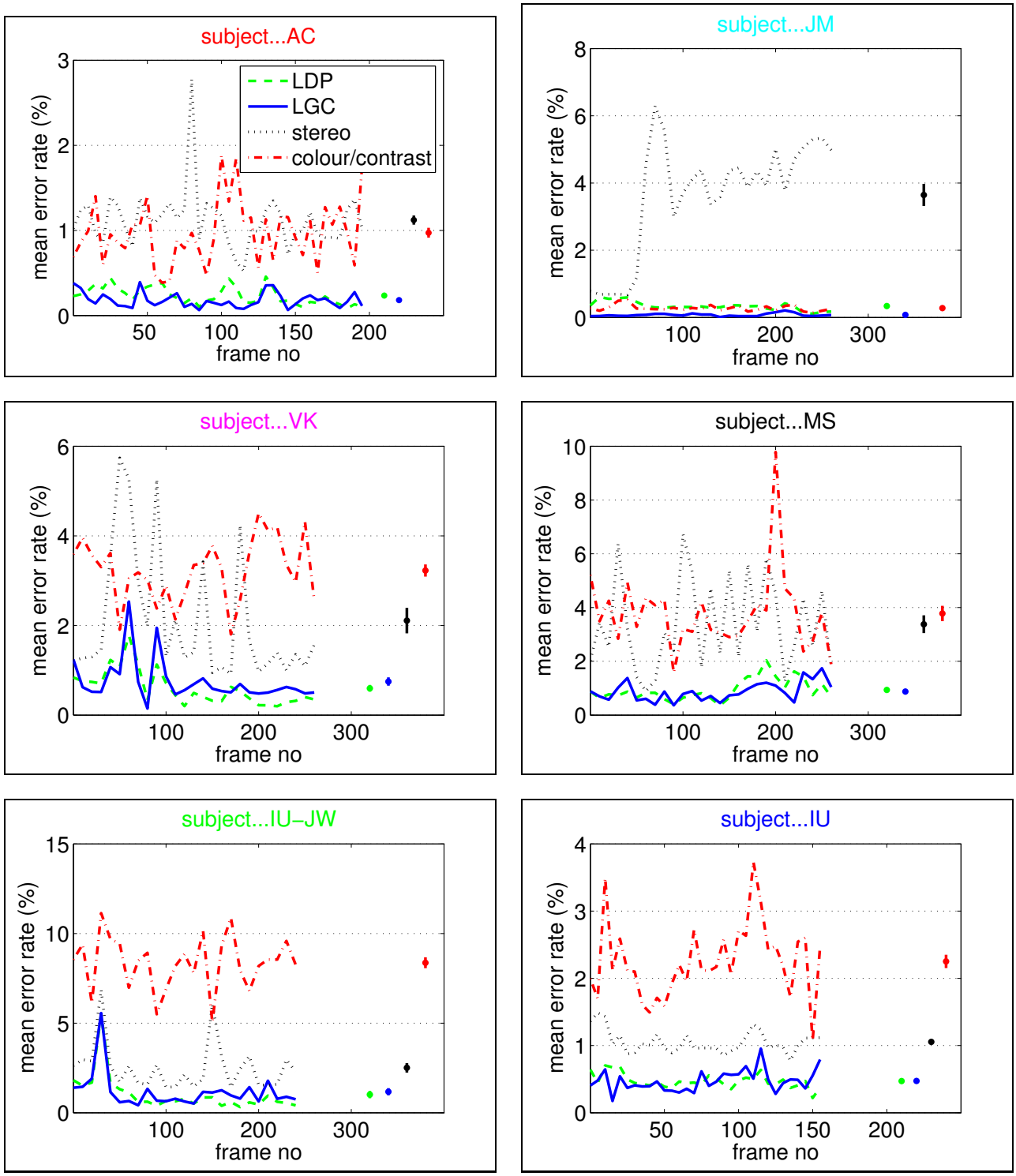


Fig. 7. **Segmentation performance advantage from fusion.** Segmentation error (percentage of misclassified pixels) is computed on all six sequences, frame by frame, for LDP, LGC, color only and stereo only. Error bars are also shown, on the right of each plot, for temporal mean and standard error. Note that fused stereo and color/contrast (LGC and LDP) perform substantially better than either stereo or color/contrast alone.

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

Segmentation performance for the various stereo test-sequences, including the AC sequence of figure 1 and five others, is compared for color/contrast, for stereo alone, and for color/contrast with stereo fused together (figure 7). The color/contrast algorithm here is simply LGC in which the stereo component is switched off. The stereo-only algorithm is 4-state DP [13]. Fusion of color/contrast and stereo by the LGC and LDP algorithms both show similarly enhanced performance compared with color/contrast or stereo alone.

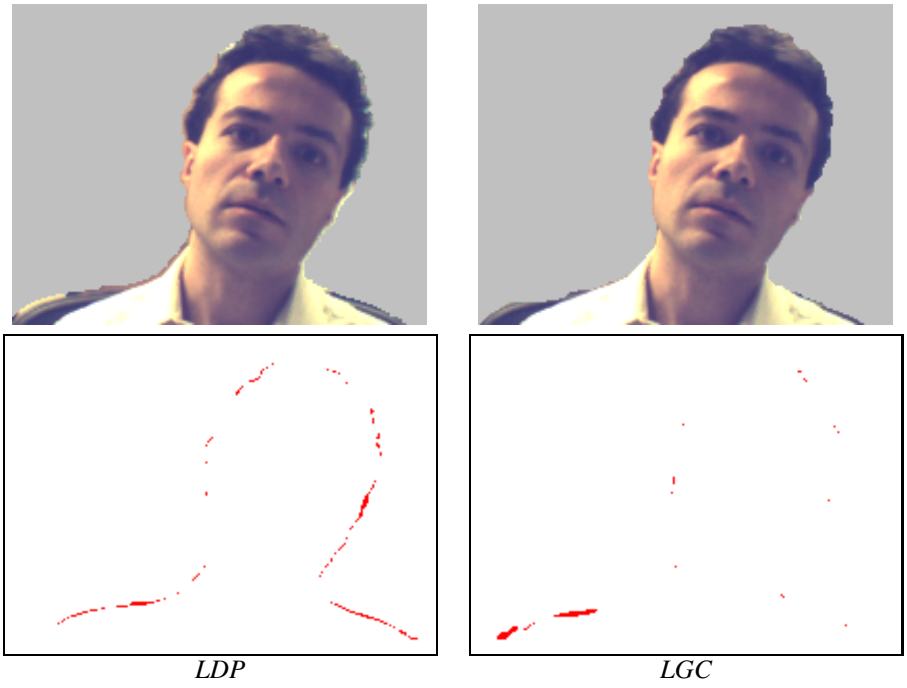
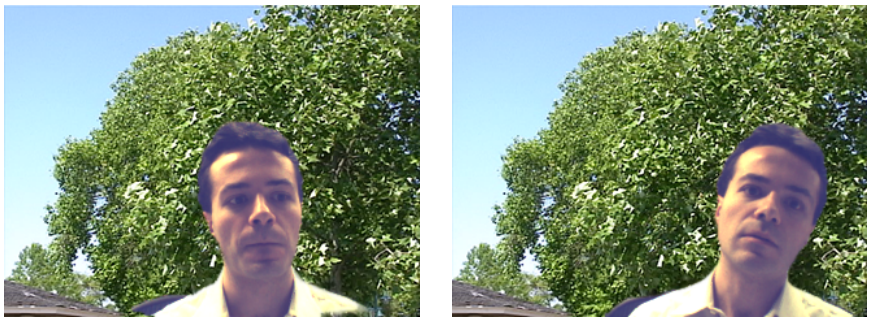


Fig. 8. **Extracted foreground layer** (top) for the left view of sequence AC, frame 100, for LGC and LDP. Segmentation error maps (bottom).

Fig. 9. **LGC Segmentation error illustrations.** We show here two results corresponding to high error rates in the test data of figure 7. Segmented foreground is shown against a grey background.

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 color/contrast and stereo is immediately apparent. In fact, the error rates shown for color/contrast alone are even optimistic, in that color 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 color/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 figure 8 together with the spatial distribution of segmentation errors: the errors tend to cluster closely around object boundaries. Finally figure 9 shows two results corresponding to high error rates in the test data of figure 7. The first (VK) aparently arises where the subjects hand saturates the intensity range of one of the cameras, disturbing the stereo matching. The second (IU−JW), more interesting, shows slightly over-aggressive action of the coherence constraint momentarily gluing two subjects together.

Background substitution in sequences.: Finally, figs. 10-12 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 practice in foreground/background segmentation, α-matting has been computed by border matting [16] as a (real time) post-process, though patch based priors can alternatively be used [31], [18]. The LGC algorithm gives good results, with blended boundaries and little visible flicker [1, Background substitution demo]; LDP gives subjectively similar results.



LGC, frame 0 *LGC, frame 100*

Fig. 10. **Segmentation and background substitution.** Here we show background substitution (using LGC) for two frames of the sequence `AC`.

Fig. 11. **Segmentation with non-stationary background.** (**Top**) Four frames of the input left 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. 12. **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.

VII. Conclusion

This paper has addressed the important problem of segmenting stereo sequences. Disparity-based segmentation and color/contrast-based segmentation alone are prone to failure. We have demonstrated properties of the LDP and LGC algorithms and underlying models, 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 color and contrast can be captured in a probabilistic model, in which parameters can mostly be learned, or are otherwise stable.
- Fusion of stereo with color and contrast makes for more powerful segmentation than for stereo or color/contrast alone.
- Good quality segmentation of temporal sequences (stereo) can be achieved, without imposing any explicit temporal consistency between neighboring 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 modeling and inference [33], but currently this is too expensive computationally for a real time system.

Tradeoff between LDP and LGC: 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 marginalization 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.

There are some other important differences between the algorithms. First, the LDP algorithm produces the entire stereo disparity map as a by-product of segmentation, whereas LGC delivers the segmentation alone. This favors LDP in applications such as cyclopean view generation, for which the full disparity map is needed in addition to the occlusion map. Quality of the disparity map computed by LDP, within segmented regions, is as for 4-state DP [13]. Another interesting difference is that whereas the figural continuity constraint, captured by the contrast term of section III-C, makes only a marginal difference to LGC performance (figure 6), 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 [15] overcomes that limitation by providing an indirect but effective linkage between nearby epipolar lines.

Computation times: Both the LDP and the LGC algorithms are capable of real time operation — in both cases, around 10 fps at 320×240 resolution, with 60 disparity levels on a conventional (3 GHz) processor. For LDP, execution times scale linearly with image area and with number of disparity levels. LGC consists of NSSD evaluation and graph cut, each of which take roughly equal time with the parameters above. [Ternary graph cut has been applied, in our laboratory, at around 1.5 M-pixels/second on a 3GHz Pentium desktop machine.] The NSSD evaluation then scales linearly with image area, and number of disparity levels. Graph cut scales approximately linearly with image area, but is, of course, independent of the number of disparity levels.

A still faster algorithm?: Relative to the full segmentation model (8), we saw that one set of simplifications leads to the LDP model, and another leads to the LGC model. It is reasonable to ask the question, what sort of performance would result in making both sets of simplifications at once? The resulting algorithm would require only ternary computation (like LGC) and be restricted to scan lines (like LDP). Estimation would simply require DP on a 3-state Markov chain, potentially very efficient. Experiments with this model gives results which, in all but one case (AC), show a clear improvement over colour segmentation alone, for the 6 datasets of figure 7. Typically error rates are reduced by around a factor of 2. Clearly, stereo under this reduced model has an effect in improving accuracy. However, the error rates are between approximately 2 and 5 times greater than for LGC, so a considerable degree of accuracy is sacrificed in the extra simplification of the model. Thus the relative computational expense of the LGC and LDP models brings clear benefits.

Future work: Future work will address several outstanding issues. One is the solution of the full problem (8), without any simplifying neglect of disparity constraints or any restriction to epipolar lines. Possible approaches are being considered both to the max-sum problem (3) and the max-max variant (9). Another important issue is the imposition of a restriction of match-cost computation to a limited range or "Panum-band". If this can be achieved without too great a loss of quality there is a considerable potential gain in efficiency, and ongoing experiments are producing promising looking results.

Acknowledgements

The authors gratefully acknowledge helpful discussions with M. Isard, J. MacCormick, O. Williams, R. Szeliski and R. Zabih.

References

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

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

[3] J. Bergen, P. Burt, R. Hingorani, and S. Peleg, "A three-frame algorithm for estimating two-component image motion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 9, pp. 886–896, 1992.

[4] S. Baker, R. Szeliski, and P. Anandan, "A layered approach to stereo reconstruction," in *Proc. Conf. Comp. Vision Pattern Rec.*, 1998, pp. 434–441.

[5] N. Jojic and B. Frey, "Learning flexible sprites in video layers," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2001, pp. 199–206.

[6] 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.*, vol. 23, no. 3, pp. 297–303, 2001.

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

[8] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scan line search using dynamic programming," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 7, no. 2, pp. 139–154, 1985.
[9] I. Cox, S. Hingorani, and S. Rao, "A maximum likelihood stereo algorithm," *Computer vision and image understanding*, vol. 63, no. 3, pp. 542–567, 1996.
[10] D. Geiger, B. Ladendorf, and A.Yuille, "Occlusions and binocular stereo," *Int. J. Computer Vision*, vol. 14, pp. 211–226, 1995.
[11] P. Belhumeur, "A Bayesian approach to binocular stereopsis," *Int. J. Computer Vision*, vol. 19, no. 3, pp. 237–260, 1996.
[12] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proc. European Conf. Computer Vision*, 2002, pp. 82–96.
[13] A. Criminisi, J. Shotton, A. Blake, and P. Torr, "Gaze manipulation for one to one teleconferencing," in *Proc. Int. Conf. on Computer Vision*, 2003, pp. 191–198.
[14] ——, "Efficient dense stereo and novel view synthesis for gaze manipulation in one-to-one teleconferencing," Microsoft Research Cambridge, Tech. Rep. MSR-TR-2003-59, 2003.
[15] 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, pp. 105–112.
[16] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[17] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, 2001.
[18] 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*, 2004, pp. 721–728.
[19] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML '01: Proc. Int. Conf. Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.
[20] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.
[21] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, USA: MIT Press, 1987.
[22] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.
[23] 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*. Springer-Verlag, 2004, pp. 428–441.
[24] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Computer Vision*, vol. 47, no. 1–3, pp. 7–42, 2002.
[25] A. Dempster, M. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B.*, vol. 39, pp. 1–38, 1977.
[26] H. Baker and T. Binford, "Depth from edge and intensity based stereo," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1981, pp. 631–636.
[27] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*. Cambridge University Press, 1998.
[28] S. Rowe and A. Blake, "Statistical mosaics for tracking," *J. Image and Vision Computing*, vol. 14, pp. 549–564, 1996.
[29] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1999, pp. 246–252.
[30] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. Int. Conf. on Computer Vision*, 1999, pp. 255–261.
[31] A. Fitzgibbon, Y. Wexler, and A. Zisserman, "Image-based rendering using image-based priors," in *Proc. Int. Conf. on Computer Vision*, 2003, pp. 279–290.
[32] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Bi-layer segmentation of binocular stereo video," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2005.
[33] O. Williams, M. Isard, and J. MacCormick, "Estimating disparity and occlusions in stereo video sequences," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2005, pp. CD–ROM.
[34] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1981, pp. 674–679.
[35] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 401–406, 1998.
[36] http://cat.middlebury.edu/stereo/.

APPENDIX

A. DETAILS OF STEREO LIKELIHOOD MODELLING

In section III-A the likelihood for evaluation of stereo matches was defined, but there remained the issue of using the statistics of labelled, matched data sets to justify the detailed model, and fix the values of parameters λ and N_0. We start from the matching cost U_k^{M} defined in (11), in terms of the NSSD N. As a refinement, we further allow for subpixel offset by parabolic interpolation, along epipolar lines, of the NSSD values

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

at successive pixels, and take the minimum value of the parabola to replace the value of $N(L_m^{\mathrm{P}},\ R_n^{\mathrm{P}})$. This subpixel refinement was found to improve error rates mildly, and was similar in effect to alternative interpolation schemes [34], [35].

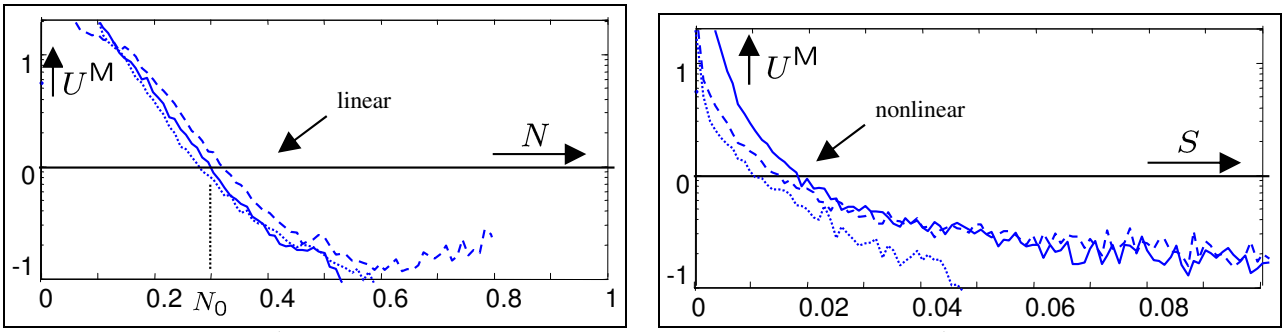
This stereo likelihood model, based on NSSD with subpixel interpolation, has been tested against the Middlebury data-sets [36] and found to be reasonable — examples of results are given in figure 13a). Importantly, linear regression analysis on U^{M} as a function of N yields $-\lambda$ as the slope and N_0 as the intercept, from (10). This gives useful working values for λ, which turns out to be quite consistent, across data sets, at around $\lambda = 10$. [3] For the parameter N_0, the data analysis yields a value of approximately 0.3, compared with the discriminatively optimal value $N_0 = 0.35$ from section VI.

As it has been more conventional [24] in stereo to use SSD as a match-cost rather than NSSD, results are included also for U^{M} modeled as a function of SSD, in figure 13b). Two issues arise from this. The first is that an effect of normalization is that the U^{M}-characteristic is more consistent across data sets for NSSD than for SSD. Hence it is reasonable to fix the parameters used to model the log-likelihood-ratio in the NSSD case, whereas for SSD, the parameters would need to be allowed to adapt — an added system complexity. The second is that the linearity apparent for NSSD is absent for SSD. Therefore the statistical evidence does not support the conventional modeling of match-cost as linear in SSD. Given a non-linear likelihood based on SSD, we have found DP stereo 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.

B. VIEWING GEOMETRY AND ITS INFLUENCE ON TRANSITION ENERGIES

This brief section explains the formula (19) for the LDP energy coefficient a_{F}, and its claimed dependence on viewing geometry.

[3] 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.



a) NSSD **b)** SSD

Fig. 13. **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 [36] ("cones", "teddy", and "sawtooth"). Note the linearity in the region of $U^{\mathrm{M}} = 0$, where the matched/unmatched hypothesis switches, and hence 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.

Assume an average slope magnitude of 1, for a visible surface, in 3D viewer-centred coordinates. In cyclopean match space coordinates d, k, this slope scales to a slope of B/D where B is the stereo baseline and D is the nominal distance from object to viewer. From figure 4, this implies:

$$\frac{\exp -a_{\mathrm{F}}}{\exp -c_{\mathrm{F}}} = \frac{B}{D}, \tag{29}$$

the ratio of probabilities for following an R-match foreground transition with another of the same, vs. switching to an L-match transition. This is simply because a strictly alternating sequence of L-match and R-match corresponds to a constant disparity trajectory, a 45^o line in match space (figure 3), and hence a line of gradient 0 in cyclopean coordinates. Allowing one repeated step, out of M otherwise alternating steps generates, by straightforward trigonometry, a line in match-space of gradient $1/M$, in cyclopean d, k coordinates. Now set $M = D/B$ to arrive at (29).

Finally, combining (29) with (21) and $b_{\mathrm{F}} = \log W_{\mathrm{F}}$ (18), gives the result (19).