Fusion of stereo, colour and contrast

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

Microsoft Research Cambridge, 7 JJ Thomson Avenue, Cambridge, UK.
www.research.microsoft.com/vision/cambridge

1 Introduction

Stereo vision has numerous applications in robotics, graphics, inspection and other areas. A prime application, one which has driven work on stereo in our laboratory, is teleconferencing in which the use of a stereo webcam already makes possible various transformations of the video stream. These include digital camera control, insertion of virtual objects, background substitution, and eye-gaze correction [9, 8].

Digital camera control: Here the foreground part of a scene is isolated using stereo, and used to drive the digital pan/zoom/tilt controls of a camera, to keep the subject well framed in the virtual view.

Insertion of virtual objects: Knowing the depth structure of a scene, virtual objects can be inserted live into the video stream, in a way that respects the space occupancy of existing, real objects.

Background substitution: Having isolated the background of a scene using stereo, it can be manipulated — for example blurred, re-colored or replaced entirely, without touching foreground elements. This demands foreground layer separation to near Computer Graphics quality, including α-channel determination as in video-matting [6], but with computational efficiency sufficient to attain live streaming speed.

Eye-gaze correction: A particularly challenging application is to combine video streams from a pair of cameras, stereoscopically, to generate a virtual camera in locations that are inaccessible to a real physical camera. In this way, a virtual camera can be placed in the centre of the screen of each of two computers, so that a pair of subjects in conversation can gaze at one another directly, eye to eye. This problem is particularly hard in practice because the baseline separating the left and right cameras has to be large (fig. 1), resulting in more substantial differences to be resolved between the two images.

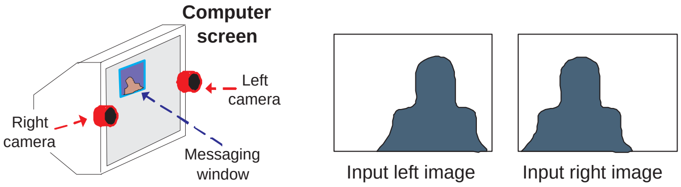


Fig. 1. Stereo cameras for teleconferencing. Two cameras are placed on the frame of a computer monitor. A virtual camera can be sythesised right over the window for viewing the remote participant is (marked in blue) on the middle of the computer screen. Viewing the subject through the virtual camera, rather than the left or right, ensures direct eye contact.

Stereo algorithms have been developed over the past 25 years that are competent at computing depth "densely" — that is at all pixels — in 3D scenes. Earlier algorithms used Dynamic Programming (DP) to compute optimal matches [15, 7] but lately two new algorithms — *Belief Propagation* and *Graph Cut* — have come to head the league table of stereo performance [18]. Stereo "occlusion" is a further cue, arising for those parts of a scene that are visible in one eye (or camera) but not the other. Occlusion needs to be accurately detected, as it is a strong cue for discontinuity of surfaces, and some modern algorithms are capable of doing this [10, 1, 13, 9, 8]. However, some problems remain. In particular, the strength of stereo cues degrades over low-texture regions such as blank walls, sky or saturated image areas. In general, it is difficult to deal with this problem, but in the particular application of stereo to foreground/background segmentation, a powerful remedy is at hand in the form of cue fusion. Recently color and contrast have been shown to be powerful cues for segmentation [4, 17], 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 for segmentation. Here we propose a model and algorithms for fusion of stereo with color and contrast, and a prior for intra-layer spatial coherence.

2 Probabilistic models for stereo matching

First we outline the probabilistic structure of the stereo matching and color/contrast models. A notation is set out for state variables and observables. Then an energy E or cost-function is defined to characterise well matched images. The energy E also defines a probabilistic model, by acting as the Gibbs energy in a *Conditional Random Field* (CRF) [14].

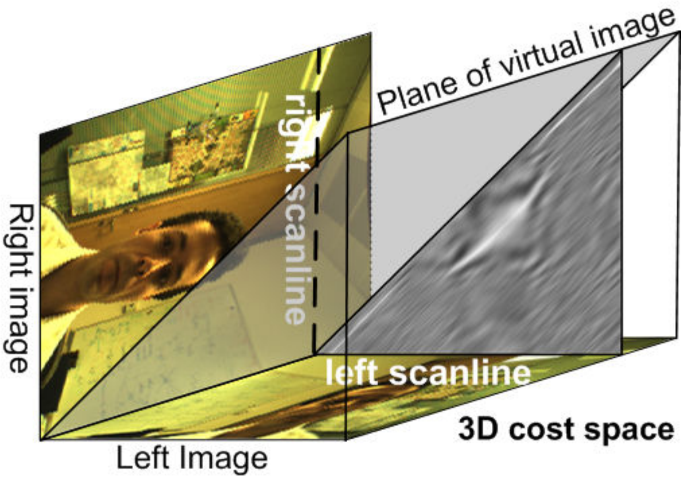


Fig. 2. Matching a pair of rectified stereo images. *Rectification means that pixels in a left scanline all match to pixels in the corresponding right scanline. A match function over a triangular domain is shown that scores the likelihood of all feasible pixel pairs over a particular pair of corresponding epipolar lines — bright means high match likelihood.*

2.1 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 $\mathbf{z} = (\mathbf{L}, \mathbf{R})$. Rectification is a projective warping transformation applied to left and right images that brings their respective scanlines into direct correspondence. Thus, in rectified images, all pixels on a horizontal ("epipolar") line in the left image match to pixels in the corresponding epipolar line in the right image. This geometrical normalisation greatly simplifies the complexity of matching pixels. A pair of rectified images is illustrated in figure 2 and the stereo problem is to establish a match between pixels in the left image and corresponding pixels in the right image. Typically, most pixels in each of the images are matched in this way. Those that remain unmatched are the *occluded* pixels, arising for instance where a particular point in the background of a scene is masked by a foreground object in the left view, but visible in the right view.

The mapping between left and right images is expressed in terms of state variables $\mathbf{x}$ and *disparities* $\mathbf{d}$. The array $\mathbf{x}$ of state variables can be defined symmetrically with respect to left and right image coordinate frames, in so-called *cyclopean* coordinates k. The array then coprises components $\mathbf{x} = \{x_k\}$ which take values $x_k \in \{\mathrm{M}, \mathrm{O}\}$, according to whether the pixel is matched or occluded. A further elaboration of the state space, employs 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 subdivision of the scene into foreground and background layers offers the opportunity for imposing further constraints, both prior and driven by data.

Stereo *disparity* is an inverse measure of three-dimensional depth, and is defined to be $d = m - n$. The disparity values along one epipolar line are expressed as $\mathbf{d} = \{d_k,\ k = 1, \ldots, 2N - 1\}$. Note this means that

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

so that k, d forms the cyclopean coordinate system for the space of epipolar matches, which is symmetric and this is well known to be helpful for probabilistic modeling of stereo matching [1].

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 for $(\mathbf{d}, \mathbf{x})$, conditioned on image data, can be defined.

2.2 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, for instance, be globally maximised to obtain an estimated segmentation $\hat{\mathbf{x}}$ and estimated stereo disparities $\hat{\mathbf{d}}$.

The model (2) can be regarded simply as a Conditional Random Field (CRF) [14], 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:

$$E(\mathbf{z}, \mathbf{x}, \mathbf{d}; \theta) = V(\mathbf{x}, \mathbf{d}; \theta) + U^{\mathrm{M}}(\mathbf{z}, \mathbf{x}, \mathbf{d}; \theta) + U^{\mathrm{C}}(\mathbf{z} \mid \mathbf{x}; \theta), \tag{3}$$

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:

$$V(\mathbf{x},\mathbf{d};\theta)=\sum_{(k,k')\in\mathcal{N}}[F(x_k,x_{k'},\Delta d_k,\Delta d_{k'})]+\sum_k G_k(x_k,d_k), \tag{4}$$

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

$$\Delta d_k = d_k - d_{k-1}. \tag{5}$$

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 (4) 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.

Optionally, when the extended state-space $x_k\in\{\mathrm{F},\mathrm{B},\mathrm{O}\}$ is used, the $G_k(\ldots)$ term is included to implement "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

$$G_k(x_k,d_k)=-\log p(d_k\mid x_k), \tag{6}$$

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. This term is optional, used only with the extended state-space $x_k\in\{\mathrm{F},\mathrm{B},\mathrm{O}\}$.

2.3 Contrast dependence

One further elaboration, due to Boykov and Jolly [4], incorporates the evidence from image contrast for segmentation — see also 'line processes" [11], "weak constraints" [3] and "anisotropic diffusion" [16]. It proves important in refining segmentation quality, at the cost of obscuring somewhat the clear generative distinction between prior and likelihood [2]. The Ising component F in (4) 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 (3) is extended in this way to a CRF

$$E(\mathbf{z},\mathbf{x},\mathbf{d};\theta) = V(\mathbf{z},\mathbf{x},\mathbf{d};\theta) + U^{\mathrm{M}}(\mathbf{z},\mathbf{x},\mathbf{d};\theta) + U^{\mathrm{C}}(\mathbf{z} \mid \mathbf{x};\theta), \tag{7}$$

in which dependence on data $\mathbf{z}$ is now incorporated in to the $V(\ldots)$ term, so that this no longer represents a pure prior distribution.

3 Inference

Two inference problems are considered for the model of the previous section. The first is the full inference of disparity $\mathbf{d}$ and state $\mathbf{x}$, necessary when the three-dimensional structure of a scene is required explicitly. This would be the case in many robotics applications, and for the gaze-correction function in the teleconferencing application described in section 1. The other three teleconferencing applications however, require only segmentation — the inference of $\mathbf{x}$ but not of $\mathbf{d}$.

3.1 Inferring disparity

To compute both disparity and state, the posterior is maximised with respect to $\mathbf{d}$ and $\mathbf{x}$:

$$(\hat{\mathbf{x}}, \hat{\mathbf{d}}) = \arg\max_{\mathbf{x},\mathbf{d}} p(\mathbf{x},\mathbf{d} \mid \mathbf{z}). \tag{8}$$

Here we take the short form of the state vector $x_k \in \{\mathrm{M},\mathrm{O}\}$, and use only stereo cues, without colour. This problem is not formally tractable but could be regarded as tractable in practice because it can be solved approximately by the α-expansion form of graph-cut [5], over the variables $\mathbf{x},\mathbf{d}$ jointly (provided the energy function E is chosen to meet the necessary regularity conditions). In practice α-expansion over $(\mathbf{x},\mathbf{d})$ jointly is computationally burdensome, one or two orders of magnitude slower than real-time, for a conventional video stream on a current single processor architecture. A faster solution can be computed by neglecting vertical constraints in the model. All vertical cliques in V (4) 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 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 [8]. In exchange for the lost vertical constraint, the problem becomes exactly tractable by dynamic programming and DP can be performed along scanlines, jointly with respect to disparities and state variables [9, 8]. This can be achieved in real time.

3.2 Inferring segmentation

An alternative aim to computing full disparity and state, is to compute only the state, and this useful with the extended state $x_k \in \{\mathrm{F},\mathrm{B},\mathrm{O}\}$, so that the image is segmented into foreground, background and occluded regions. Then colour distributions, associated with background and foreground, can be brought into the model. For this problem, the posterior should, in principle, be marginalised with respect to **d**, and then maximised with respect to **x** to estimate a segmentation:

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

This problem is intractable with the Gibbs energy model (7) above. This paper proposes two approaches to simplifying the Gibbs energy model, to make inference of segmentation **x** practically tractable and efficient.

LDP. In Layered Dynamic Programming [12], vertical constraints are neglected as above for full stereo. The marginalised form of the problem (9) is not tractable even without vertical constraints, so it is necessary to stick with the full problem (8), and simply discard the unwanted disparities. This is not ideal because, in principle, statistical information is wasted on the computation of disparities.
The difference then between LDP and full DP stereo is simply that the extended state $x_k \in \{\mathrm{F}, \mathrm{B}, \mathrm{O}\}$ is used, with appropriate energy terms to represent foreground and background constraints, both prior and from data, in terms of colour properties of the foreground and background layers of the scene.

LGC. In Layered Graph Cut [12], the prior term $F(\ldots)$ in (4) is made independent of disparity **d**. Now the posterior density can be marginalised exactly over **d** in the original inference problem (9). Marginalization gives the posterior density $p(\mathbf{x} \mid \mathbf{z})$ for segmentation only, which can be maximised by graph-cut with 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 in fact also for parameter learning [12]. 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 disadvantage that parameter estimation remains intractable. In terms of practical efficiency and efficacy, the two algoriths, LDP and LGC, perform remarkably similarly [12], despite having very different structures.

4 Some results from stereo matching and segmentation

Results of full stereo matching, used to generate virtual camera views, are illustrated in figure 3. The top line show real left and right cameras, used

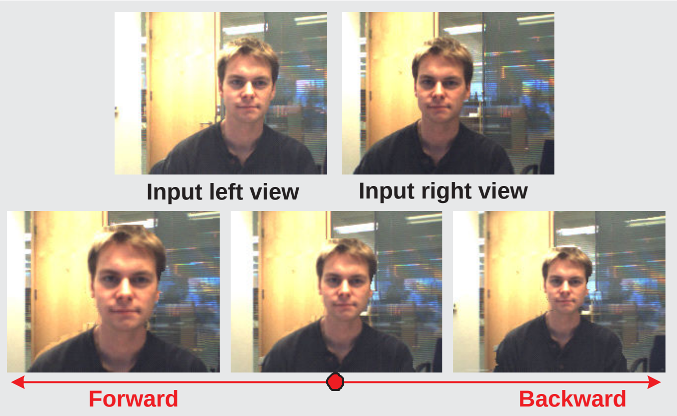
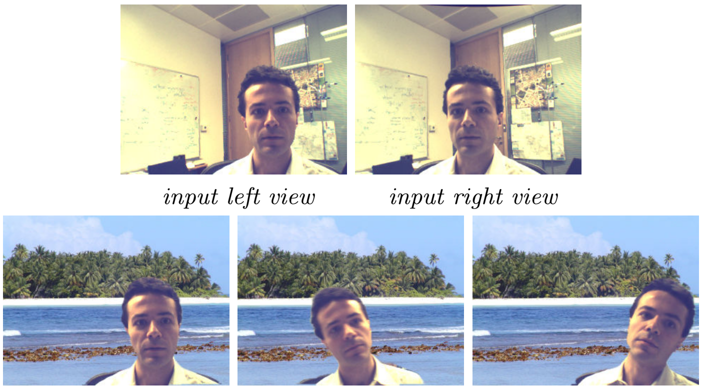


Fig. 3. Virtual camera.

as input to stereo matching. Computed disparities are then used, either implicitly or explicity, to recover the shape of the scene. The scene can then be reprojected onto the image plane of a virtual camera. In the bottom centre, an interpolated cyclopean view is shown, of the sort that can be used for gaze correction — the subject is looking directly forwards in this view. It is critically important for the quality of the virtual image, that not only disparity but also the occlusion information in $\mathbf{x}$ is available [9, 8]. Bottom left and right images in the figure show the views when the virtual camera is moved respectively backward and forward in space.

Results of stereo segmentation, fusing stereo, colour and contrast, are shown in figure 4. Left and right images are processed using the LGC algorithm above, to separate the foreground subject from its background [12]. The extracted foreground can then be applied to a new background and this is illustrated in the figure for three frames of a test video. Special measures — so-called *border matting* [17] — are taken so that he extracted foreground sequence can be composited, free of “aliasing” artefacts, onto the background. Border matting deals with mixed pixels — pixels that contain both foreground and background colour, occurring typically around the boundary of an object. If this is neglected, discolouration occurs around boundaries, where traces of the original background colour remains stuck to the foreground subject.

The paper has made a rapid tour of some recent progress in algorithms for stereo vision. Highlights include: a probabilistic framework; the full treatment of occlusion via an appropriate representation of state; fusion of cues, specifically stereo, colour and contrast. Many details have been omitted in this account, and the reader is directed to [8, 12] for full details.



input left view *input right view*

automatic layer separation and background substitution

Fig. 4. Background substitution.

References

1. P.N. Belhumeur. A Bayesian approach to binocular stereopsis. *Int. J. Computer Vision*, 19(3):237–260, 1996.
2. 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.
3. A. Blake and A. Zisserman. *Visual Reconstruction.* MIT Press, Cambridge, USA, 1987.
4. 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 105–112, 2001.
5. 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.
6. Y-Y Chuang, B. Curless, D.H. Salesin, and R. Szeliski. A Bayesian approach to digital matting. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages CD–ROM, 2001.
7. 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.
8. 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.
9. 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.
10. D. Geiger, B. Ladendorf, and A.Yuille. Occlusions and binocular stereo. *Int. J. Computer Vision*, 14:211–226, 1995.
11. 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.

12. 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.
13. V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proc. European Conf. Computer Vision*, pages 82–96, 2002.
14. 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*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
15. 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.
16. 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.
17. C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.
18. 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.