Dense Motion and Disparity Estimation via Loopy Belief Propagation

Michael Isard and John MacCormick
Microsoft Research Silicon Valley
November 2005

Abstract

We describe a method for computing a dense estimate of motion and disparity, given a stereo video sequence containing moving non-rigid objects. In contrast to previous approaches, motion and disparity are estimated simultaneously from a single coherent probabilistic model that correctly accounts for all occlusions, depth discontinuities, and motion discontinuities. The model is a Markov random field (MRF) whose label space incorporates every possible occlusion status for every pixel. Hence, the MRF's data likelihoods are physically realistic. The results demonstrate that simultaneous estimation of motion and disparity is superior to estimating either in isolation, and show the promise of the technique for accurate, probabilistically justified, scene analysis.

1. Motivation and previous work

The "temporal stereo + motion" problem of estimating the disparity and motion fields in a video sequence of moving objects captured by a calibrated pair of stereo cameras has been studied for at least two decades [16]. It is worthwhile to distinguish between the standard temporal stereo + motion problem, and the more restricted problem of estimating disparity and motion from two consecutive frames in a stereo sequence; we refer to the latter as "two-frame stereo + motion". This paper first introduces a novel solution for two-frame stereo + motion, then explains how to extend the solution to stereo sequences.

Our ultimate objective is to form a reliable, dense 2.5D representation of an image sequence. Acquiring a rectified stereo sequence and running traditional stereo algorithms fills in much of the necessary information, but dense disparity estimation from a single stereo pair is challenging. Matches can be highly ambiguous in non-textured regions; and background regions near foreground object boundaries are only visible in a single camera, meaning their depth must be estimated using only prior information about the shapes of objects in the world.

Exploiting temporal coherence in the stereo sequence can in principle alleviate both of these problems, however as previous work has noted [17], in the absence of explicit motion estimates it is hard to do better than to average out thermal imaging noise in *stationary* regions. We therefore propose to jointly estimate dense motion and disparity in a single coherent probabilistic framework. We show that making use of two-frame motion estimation in conjunction with traditional stereo greatly reduces the regions of the scene which are visible only in a single image. In addition, by filtering over time we are able to propagate information about the depth of scene patches during extended occlusions in the non-reference image.

Our approach to two-frame stereo + motion defines a single Markov random field (MRF) whose nodes are the pixels of the reference image, and whose labels incorporate all possible disparity, fronto-parallel motion, change-in-disparity, and occlusion values. The disparity and motion fields are determined by approximating the MAP estimate for this MRF using loopy belief propagation. As far as we are aware, this is the first work to attempt simultaneous disparity and motion estimation using MRFs. In more abstract terms, however, our approach is distinguished from previous approaches to temporal stereo + motion in three important respects: (i) our estimates are *dense*, in contrast to feature-based approaches such as [4]; (ii) we employ a single *coherent* probabilistic model, in contrast to iterative segmentation approaches such as [15]; and (iii) the likelihoods correctly account for occlusions and discontinuities. We believe this paper presents the first stereo + motion work satisfying all of (i)-(iii).

Item (iii), the modeling of occlusions and discontinuities, can be viewed as a generalization of the occlusion modeling in much previous work on stereo (e.g. [2, 5]). The essential idea is that the likelihood of a particular disparity hypothesis for a particular world point cannot be computed without also specifying whether that point is visible or occluded in each of the images. This "occlusion status" varies in a deterministic fashion near object boundaries.

Figure 1 gives a schematic example of this for the stereo + motion problem. One key contribution of this work is that the data likelihoods in the MRF are computed in the following way. The MRF label at a reference image pixel includes an occlusion status (corresponding to the color rendered in figure 1), and this is used in turn to determine which of the non-reference image patches should contribute to the data

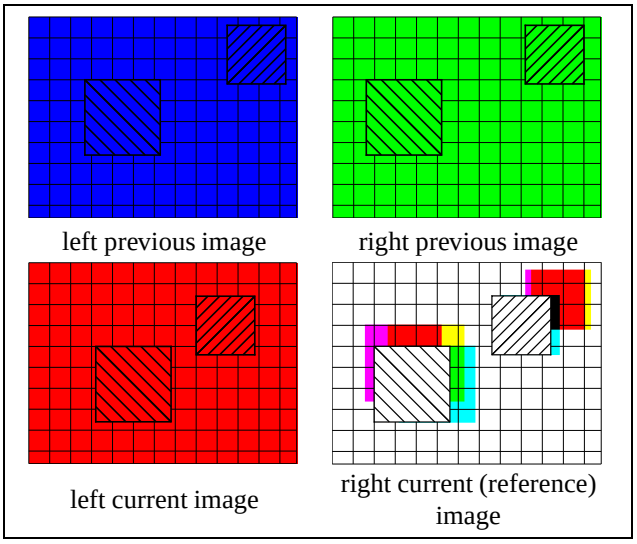


Figure 1: **Motion and disparity determine visibility in non-reference images.** *Two foreground objects with positive disparities are shown moving against a zero-disparity stationary background. Each pixel in the reference image is colored according to which non-reference images it is visible in. For example, a pixel visible in the left and right previous images but not the left current image is colored blue + green = cyan, pixels visible in all three non-reference images are white, and pixels visible nowhere except the reference image are black.*

likelihood. In contrast to much previous work on stereo and motion, patches corresponding to occluded world points are explicitly excluded when they should be.

Our solution to the multi-frame temporal stereo + motion problem amounts to a simple extension of the two-frame MRF. By treating the problem in the context of filtering (as opposed to smoothing), the outputs from previous frames can be incorporated by adding an extra term to the MRF data cost. We believe this is the first published algorithm for performing dense, temporally-filtered stereo + motion estimation.

Section 2 describes the MRF employed for two-frame stereo + motion, and Section 3 explains the extension to the multi-frame case. Section 4 discusses the use of loopy belief propagation to approximate MAP estimates in these MRFs, and Section 5 describes the results.

1.1. Related work

Work on temporal stereo + motion has generally been based on sparse image features. This sparsity is not directly compatible with the dense reconstruction of the disparity and motion fields, which is the goal of this paper. Examples of the feature-based approach include [4], which uses line correspondences, and [8].

One significant example that uses optical flow rather than features is [15]. However, this approach employs an iterative segmentation of the scene: an initial estimate is obtained assuming a single rigid motion of the entire scene, then objects with distinct motions are segmented in later iterations by detecting outliers. In contrast, the approach of this paper employs a single probabilistic model from which the motions of all objects are inferred coherently.

Our work is closer in spirit to the large literature on dense stereo reconstruction, including those methods that use belief propagation [13], graph cuts [9], or dynamic programming [5, 10]. However, none of these approaches attempt motion estimation.

Other notable temporal stereo + motion contributions include [19], which achieves excellent accuracy using structured light, and [12, 14], both of which describe interesting algorithms which cannot conveniently be placed in a probabilistic framework.

2. The MRF for two-frame stereo + motion

The input to the two-frame stereo + motion algorithm consists of four images: Left_0, Right_0, Left_1, Right_1 (which are, respectively, the left and right stereo views of the previous and current frames of a stereo video sequence). The stereo pairs are assumed to be rectified, so that epipolar lines are horizontal, with corresponding pairs occurring on the same scanline.

The output consists, informally, of a complete reconstruction of the disparity and motion fields implied by these four images. To formalize this, we define a graphical model and compute an approximation to the MAP estimate of the disparity and motion fields. The unknowns in the graphical model form a standard four-connected rectangular lattice of the same size as the input images. The nodes are denoted $g_{x,y}$, $x \in \{0, 1, \ldots, X-1\}$, $y \in \{0, 1, \ldots, Y-1\}$, where X, Y are the width and height, respectively, of the input images. Much previous work on stereo has aimed to estimate disparities in the "cyclopean" image—that of a virtual camera placed halfway between the rectified left and right cameras. Employing the cyclopean image permits a symmetric formulation and enables one to enforce the well-known "ordering constraint" [7] if desired, but it makes the implementation and description of the algorithm more complex. Here, we don't plan to enforce the ordering constraint, so we choose not to use a cyclopean image. Instead, we arbitrarily select the current right-hand image Right_1 to be the reference image. Thus, the state at node $g_{x,y}$, denoted $s_{x,y}$, represents the motion and disparity estimated at pixel (x, y) in Right_1.

The state $s_{x,y}$ at node $g_{x,y}$ models a particular point (or, more realistically, a patch) P on a particular object in the

world. P is found by back-projecting a ray from the pixel (x, y) in the reference camera until the ray intersects a scene object. Note that P is fixed on the object, but the object itself may have moved between the previous and current frames. Note also that P may or may not be visible in each of the three non-reference images. The state $s_{x,y}$ is specified by five components. Omitting the x, y suffices, we write $s = (o, d, u, v, w)$, where:

- o is an "occlusion status", described below
- d is P's disparity in the current frame
- u and v are respectively the horizontal and vertical components of P's motion
- w is the difference between P's disparity in the previous frame and the current frame; w can also be thought of as the "depth" component of the motion.

The occlusion status o comprises three binary flags, specifying whether or not P is visible in the non-reference images. To be precise, $o = (o_{L1}, o_{L0}, o_{R0})$ where each flag takes values in $\{\mathsf{Visible}, \mathsf{Occluded}\}$ according to the following rules:

$$o_{L1} = \begin{cases} \mathsf{Visible} & \text{if } P \text{ is visible in } \mathsf{Left}_1 \\ \mathsf{Occluded} & \text{otherwise} \end{cases}$$

$$o_{L0} = \begin{cases} \mathsf{Visible} & \text{if } P \text{ is visible in } \mathsf{Left}_0 \\ \mathsf{Occluded} & \text{otherwise} \end{cases}$$

$$o_{R0} = \begin{cases} \mathsf{Visible} & \text{if } P \text{ is visible in } \mathsf{Right}_0 \\ \mathsf{Occluded} & \text{otherwise} \end{cases}$$

A formal definition of the remaining state variables — d, u, v, w — consists of describing where P projects to in each non-reference image, assuming that it is visible. The definitions adopted are that P projects to

$$\begin{aligned} (x + d, y) \quad & \text{in } \mathsf{Left}_1 \\ (x - u + d - w, y - v) \quad & \text{in } \mathsf{Left}_0 \\ (x - u, y - v) \quad & \text{in } \mathsf{Right}_0 \end{aligned} \tag{1}$$

Figure 2 gives a schematic example.

The posterior probability of the graphical model with states $\{s_{x,y}\}$ is (by definition) the product of some one- and two-node potentials:

$$\mathcal{L} = \prod_{(x,y)} \Phi(s_{x,y}) \prod_{(x,y)\sim(x',y')} \Psi(s_{x,y}, s_{x',y'}), \tag{2}$$

where the second product is over pairs of neighboring nodes.

If the model contains no cycles, Φ and Ψ can be correctly interpreted as data likelihoods and prior probability densities, respectively. When cycles are present, they are arbitrary potentials. Our rectangular lattice contains cycles, of course, but in Sections 2.1 and 2.2 we will nevertheless view the potentials as likelihoods and densities: this approximation will allow us to choose sensible forms for Φ and Ψ without resorting to the black magic of parameter-tweaking.

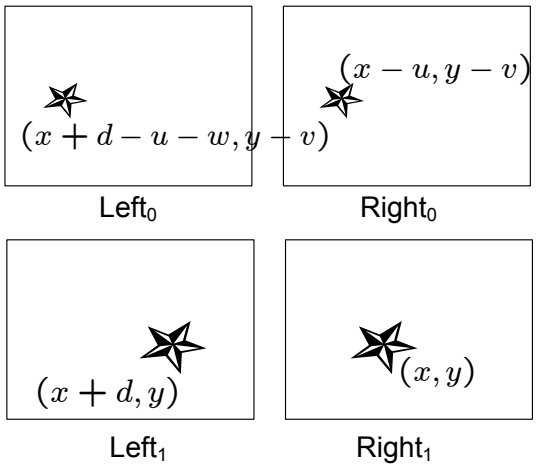


Figure 2: **State variable definitions.** *This schematic example shows a foreground object moving down, right, and towards the cameras. The object's current location in the reference image is (x, y); its current disparity is d; its motion perpendicular to the optical axis is (u, v); and its change in disparity is w.*

Maximizing $\mathcal{L}$ is the same as minimizing its negative log, so writing $\phi = -\log \Phi, \psi = -\log \Psi$ we can cast the final objective as minimizing the log posterior:

$$L = \sum_{(x,y)} \phi(s_{x,y}) + \sum_{(x,y)\sim(x',y')} \psi(s_{x,y}, s_{x',y'}).$$

The first term here is the *data cost*, discussed next in section 2.1. The second term is the *continuity cost*, discussed in section 2.2.

2.1. Data cost

The *normalized sum of squares difference* (NSSD) [11] between patches centered at (x, y) in image I and (x', y') in image I' is defined as

$$\mathrm{NSSD}(I, x, y; I', x', y') = \frac{\sum_{dx,dy} \|(I_{x+dx,y+dy} - \overline{I}_{x,y}) - (I'_{x'+dx,y'+dy} - \overline{I}'_{x',y'})\|^2}{2 \sum_{dx,dy} \left(\|I_{x+dx,y+dy} - \overline{I}_{x,y}\|^2 + \|I'_{x'+dx,y'+dy} - \overline{I}'_{x',y'}\|^2 \right)} \tag{3}$$

Here, (dx, dy) ranges over an origin-centered $K \times K$ patch of integers in $\mathbb{Z}^2$; $\|\cdot\|$ is the Euclidean norm in RGB space

(i.e. $\mathbb{R}^3$); $I_{x,y}$ is the RGB value (in $\mathbb{R}^3$) of the image I at pixel location (x, y); $\overline{I}_{x,y}$ is the average RGB value of the image I over a $K \times K$ patch centered on (x, y).

Experience has shown that the discriminatory power of the NSSD (3) is improved by changing it in two ways. First, the means $\overline{I}_{x,y}$ are computed with a Gaussian weighting centered on the relevant patch, with a relatively small standard deviation of 0.75 pixels. Second, the NSSD is redefined to be the minimum of (3) over all 2-D sub-pixel shifts of the patch centered at (x, y). This sub-pixel shift is extremely important when dealing with coarse-scale pixels, as one must in the initial stages of a multi-resolution algorithm, for instance. The sub-pixel shift can be computed analytically from the image and gradient values within the patch, using the Lucas-Kanade formulas [1].

Obviously, the NSSD is expected to be small for patches derived from different views of the same world point, and arbitrary otherwise. This intuition is captured here by assuming the NSSD is distributed according to some probability law $\Pi(\cdot)$ when the patches correspond, and a distinct probability law $\tilde{\Pi}(\cdot)$ otherwise. The negative log probabilities for these distributions will be written $\pi = -\log \Pi$, $\tilde{\pi} = -\log \tilde{\Pi}$. Numerical values for $\Pi, \tilde{\Pi}$ can be learned from training data or derived from physical assumptions, as described in the appendix.

The data cost associated with graph node $g_{x,y}$ in state $s = (o, d, u, v, w)$ can now be defined. First, let

$$\begin{aligned}
\mathsf{NSSD}_{L1} &= \mathrm{NSSD}(\mathsf{Right}_1, x, y; \mathsf{Left}_1, x + d, y) \\
\mathsf{NSSD}_{L0} &= \mathrm{NSSD}(\mathsf{Right}_1, x, y; \mathsf{Left}_0, x + d - u - w, y - v) \\
\mathsf{NSSD}_{R0} &= \mathrm{NSSD}(\mathsf{Right}_1, x, y; \mathsf{Right}_0, x - u, y - v)
\end{aligned} \tag{4}$$

These definitions have a simple intuitive interpretation. The node $g_{x,y}$ models a world point P. Each of the NSSDs in (4) computes the similarity of two patches that are projections of P: one in the reference image, centered at (x, y), and one in a non-reference image, centered at the location implied by d, u, v, w, as defined by equation (1). However, there is no guarantee that P is actually visible in the non-reference images. In the cases when P is visible, the NSSD will be distributed according to $\Pi(\cdot)$; but when it is occluded, the NSSD is distributed according to $\tilde{\Pi}(\cdot)$. Recalling the definitions of $\pi, \tilde{\pi}$ above, this motivates the following further definitions:

$$\mathsf{Cost}_{L1} = \begin{cases} \pi(\mathsf{NSSD}_{L1}) & \text{if } o_{L1} = \mathsf{Visible}, \\ \tilde{\pi}(\mathsf{NSSD}_{L1}) & \text{otherwise.} \end{cases}$$

$$\mathsf{Cost}_{L0} = \begin{cases} \pi(\mathsf{NSSD}_{L0}) & \text{if } o_{L0} = \mathsf{Visible}, \\ \tilde{\pi}(\mathsf{NSSD}_{L0}) & \text{otherwise.} \end{cases}$$

$$\mathsf{Cost}_{R0} = \begin{cases} \pi(\mathsf{NSSD}_{R0}) & \text{if } o_{R0} = \mathsf{Visible}, \\ \tilde{\pi}(\mathsf{NSSD}_{R0}) & \text{otherwise.} \end{cases}$$

These costs are genuine log probabilities, based on the distribution of NSSDs for matched and unmatched patches. Assuming independence between the different NSSD outcomes is equivalent to summing these log probabilities, leading to a total data cost given by

$$\phi_{x,y}(s) = \mathsf{Cost}_{L1} + \mathsf{Cost}_{L0} + \mathsf{Cost}_{R0}. \tag{5}$$

Previous work [3] by others using a similar data cost has shown empirically that the log likelihood ratio of NSSDs, $\pi/\tilde{\pi}$, is well-approximated by a linear function in the region of interest. We take advantage of this here by noting that (5) can be expressed in terms of this log likelihood ratio, and adopt a learnt linear function for $\pi/\tilde{\pi}$.

2.2. Continuity cost

Consider two neighboring nodes g, g' in the graphical model. They are in states $s = (o, d, u, v, w)$ and $s' = (o', d', u', v', w')$ respectively. We would like to derive the continuity cost $\psi(s, s')$. For the sake of simplicity (and somewhat in the spirit of variational methods), we assume the five components of the state are probabilistically independent, given the image data. Neglecting these dependencies is equivalent to adopting the following functional form for the continuity cost:

$$\begin{aligned}
\psi(s, s') = \psi_m(o, o') + \psi_d(d, d') + \psi_u(u, u') \\
+ \psi_v(v, v') + \psi_w(w, w').
\end{aligned} \tag{6}$$

Reasonable choices for each of these terms can be determined based on expected scene characteristics and the physics of image formation in a calibrated stereo camera rig. For ψ_m, we choose a Potts model with temperature T:

$$\psi_m = \begin{cases} 0 & \text{if } o = o', \\ 1/T & \text{if } o \neq o'. \end{cases} \tag{7}$$

An appropriate value of the temperature T can be determined by simulating the Potts model, adjusting T to achieve the desired proportion of o-discontinuities in the image.

For each of the remaining terms in (6), we assume the absolute difference is distributed such that the negative log of its distribution function has a truncated linear form, for example:

$$\psi_d(d, d') = \min\left(a, b\,|d - d'|\right). \tag{8}$$

The appendix describes how to choose sensible values for a, b based on physical reasoning.

In fact, a need not be constant over the graphical model. Observe that disparity and motion fields are often discontinuous at object boundaries, and object boundaries often occur at locations with high image gradients. This intuition can be incorporated by setting $a = a_0 \exp(-\|\nabla I\|/\alpha)$,

where $\|\nabla I\|$ is the gradient of the reference image at the location corresponding to the nodes g, g'. We follow the judicious choice of [3] in setting α to be the average value of the image gradient over the whole reference image. However, note that the authors of [3] switch on this so-called "contrast model" only between nodes whose occlusion status differs: this is because [3] deals with 1-D horizontal MRFs, in which a change of occlusion status is guaranteed to correspond to an object boundary. In contrast, when using 2-D or 3-D MRFs, object boundaries can occur between two neighboring MRF nodes with the same occlusion status. (The simplest example is two vertical neighbors straddling a horizontal object boundary—in this case, both relevant world points are visible in all three non-reference images, so the nodes' occlusion statuses are the same.) Hence, our contrast model is switched on for all pairs of neighboring nodes.

3. Temporal filtering of stereo+motion

The previous section described a model for computing disparity and motion fields from two consecutive frames of a stereo video sequence. Clearly, this model could be applied separately to each pair of consecutive frames in a sequence, to obtain disparity and motion fields for the entire sequence. However, we would like to do better: it should be possible to obtain improved estimates by exploiting temporal coherence. This can be achieved with very little extra computational cost, by adopting a *filtering* (as opposed to *smoothing*) model in which inferences at time t are influenced by the past — specifically, the output at time $t-1$ — but are independent of the future.

To explain the details of this, some more general notation is needed. Let $\mathcal{G}^{(t)}$ be the MRF for time t, with nodes $g_{x,y}^{(t)}$ and labels $s_{x,y}^{(t)}$. The output of the filtering algorithm at time t is a set of estimated labels $\hat{\mathbf{s}}^{(t)} = \{\hat{s}_{x,y}^{(t)}\}$.

It can be shown [17] that this filtering model is equivalent to adding an extra term to the data cost (5), consisting of a *temporal compatibility function* $\gamma(s_{x,y}^{(t)}; \hat{\mathbf{s}}^{(t-1)})$. A plausible form of this temporal compatibility function can be derived as follows. As usual, write the label in terms of its occlusion status, disparity, and motion as $s_{x,y}^{(t)} = (o, d, u, v, w)$, with the occlusion status further broken out into three bits expressing the visibility in the non-reference images: $o = (o_{L,t}, o_{L,t-1}, o_{R,t-1})$. Let P be the world point visible at location (x, y) in the reference image. Then $s_{x,y}$ expresses certain physical facts about P, including the following: if $o_{R,t-1} = \mathsf{Visible}$, then P is visible in image Right_{t-1} at location $x' = x - u, y' = y - v$, with disparity $d' = d - w$. Adopting a constant velocity motion model, we may also assume that P's velocity at time $t-1$ is given by $u' = u$, $v' = v$, $w' = w$.

However, note that the image Right_{t-1} is the reference image for the stereo + motion computation on $\mathcal{G}^{(t-1)}$. Thus (still assuming that $o_{R,t-1} = \mathsf{Visible}$), the MAP estimate for $\mathcal{G}^{(t-1)}$ also has an opinion about P's state: specifically, its opinion is equal to $\hat{s}_{x',y'}^{(t-1)}$, which we write more explicitly as $\hat{s}_{x',y'}^{(t-1)} = (\hat{o}, \hat{d}, \hat{u}, \hat{v}, \hat{w})$.

The temporal compatibility function γ expresses the fact that P's disparity and motion is expected to vary slowly, so this cost should be small when $s_{x,y}$ is close to $\hat{s}_{x',y'}$. A standard choice is to interpret γ as the negative log of a robust distribution function whose components are independent. This is equivalent to taking $\gamma(s_{x,y}; \hat{\mathbf{s}}^{(t)}) = \gamma_d(s_{x,y}, \hat{s}_{x',y'}) + \gamma_u(s_{x,y}, \hat{s}_{x',y'}) + \gamma_v(s_{x,y}, \hat{s}_{x',y'}) + \gamma_w(s_{x,y}, \hat{s}_{x',y'})$, with a robust cost function such as the truncated linear for each component e.g. $\gamma_d(s_{x,y}, \hat{s}_{x',y'}) = \min(a, b\,|d' - \hat{d}|)$ for constants a, b.

However, the previous discussion assumed that point P was visible in Right_{t-1} (i.e. $o_{R,t-1} = \mathsf{Visible}$). If P is not visible, the temporal compatibility function should be uniform. Therefore, the final form adopted for the components of γ is:

$$\gamma_d(s_{x,y}, \hat{s}_{x',y'}) = \begin{cases} \min(a, b\,|d' - \hat{d}|) & \text{if } o_{R,t-1} = \mathsf{Visible}, \\ a & \text{otherwise,} \end{cases}$$

and similarly for $\gamma_u, \gamma_v, \gamma_w$. The similarities with our continuity cost model in equations (6) and (8) should be obvious, and again, the appendix explains how to make sensible choices for a, b.

4 Inference for stereo + motion

We propose to estimate the MAP of the MRF described in the previous section using the min-sum formulation of loopy belief propagation (BP) [18]. This well-known technique computes messages m_{ij} from each node i to each of its MRF neighbors j. Denoting the state space at a single node by $\mathcal{S}$, each message is a function $m_{ij} : \mathcal{S} \to \mathcal{S}$. The messages are initialized to 0 and updated according to

$$m_{ij}(s) = \min_{s' \in \mathcal{S}} \left(\phi(s) + \psi(s, s') + \sum_{l \sim i \backslash j} m_{li}(s') \right), \quad s \in \mathcal{S}, \tag{9}$$

where the notation $l \sim i \backslash j$ means all neighbors l of i except for j. The messages are updated according to some schedule until convergence, at which point the MAP is estimated as

$$m_i = \mathrm{argmax}_{s \in \mathcal{S}} \left(\phi(s) + \sum_{l \sim i} m_{li}(s) \right). \tag{10}$$

A key difficulty for us here is that the state space $\mathcal{S}$ is enormous. For a typical moderate-baseline stereo sequence

with significant motion, we may have to consider, say, 50 d-values, 50^2 (u, v)-values, and 10 w-values. Throwing in the 8 o-values for the occlusion status leads to state space size $|\mathcal{S}| \approx 1 \times 10^7$. Moreover, this must be computed at, say, 8×10^4 pixels. At first glance, the situation appears hopeless, since (9) demands $|\mathcal{S}|$ calculations, each of complexity $O(|\mathcal{S}|)$, for a total complexity of $O(|\mathcal{S}|^2)$.

It turns out that for certain classes of continuity cost ψ, a so-called *distance transform* can be applied, permitting all $|\mathcal{S}|$ calculations in (9) to be accomplished in only $O(|\mathcal{S}|)$ operations. The use of distance transforms in low-level vision problems was pioneered by Felzenszwalb and Huttenlocher [6]. Naturally, we have deliberately chosen suitable cost functions: the truncated linear cost defined in (8), and the Potts model (7), are both amenable to distance transforms, and this remains true even when the costs are combined component-wise as in (6).

Notwithstanding the enormous computation savings derived from the distance transform, our computational worries are not over yet: belief propagation on large images with large disparities and motions remains expensive. It is clear that a multi-resolution approach would help to ameliorate the expense. But note that approaches such as [6], which employ coarser resolutions of the *pixel* (or graph node) space, while retaining the full *state space* resolution, are insufficient: the multiscale algorithm must reduce the number of states considered at each node, from the naïve 10^7 to something much more reasonable, perhaps 10^4. We believe it is possible to do this, but the design of such a multiscale algorithm is not at all trivial, and must be postponed to a future paper. Hence, the results presented in the next section employ small, coarsely-subsampled images in order to demonstrate the effect of our stereo+motion algorithm while keeping computational requirements within acceptable limits.

5. Results

We tested our algorithm on 21 frames of the "Geoff" stereo sequence obtained from the public database at `http://www.research.microsoft.com/vision/cambridge/i2i/DSWeb.htm`. For reasons of computational cost, we focus on a 100×80 pixel region in the top corner of the sequence, subsampled by a factor of 2 to give 50×40 pixels per frame. For the full stereo+motion computation we use a label space with maximum values of $|o| = 8, |d| = 8, |u| = 8, |v| = 3, |w| = 1$, giving 1536 labels per node. This corresponds for example to a maximum disparity in the original image resolution of 16 pixels, and horizontal translations in the range $[-8, 7]$ pixels, again in the original image resolution. The small image size and restricted range of disparity and motion are clearly chosen for computational convenience, however the power of the approach is demonstrated even on this limited example. The full sequence is shown in the accompanying video submission.

Figure 3 demonstrates resistance to fast-moving occluders. When a nearby foreground object moves in from the left the stereo computation alone is unable to accurately estimate the foreground disparity in the newly-occluded region. The filtered stereo + motion algorithm correctly uses information from previous timesteps to recover a reasonable disparity estimate. The two-frame stereo+motion algorithm, not shown, has a slightly noisier output but avoids the gross artifact.

Figure 4 shows an additional benefit of temporal filtering. The right hand edge of the image is textureless and the foreground person is almost stationary, hence neither the disparity alone nor two-frame stereo+motion can accurately estimate the disparity where the wall is occluded in the left image. Since the foreground person was previously further to the left, there was a reliable disparity estimate on the wall at an earlier frame, and the filtering algorithm has propagated this estimate in the absence of new information. In principle a filtered disparity estimate with no motion might also have correctly estimated this region but in practice this is not found to be the case. This is not surprising given the result from [17] that filtering on disparity alone is not powerful near the boundaries of moving objects.

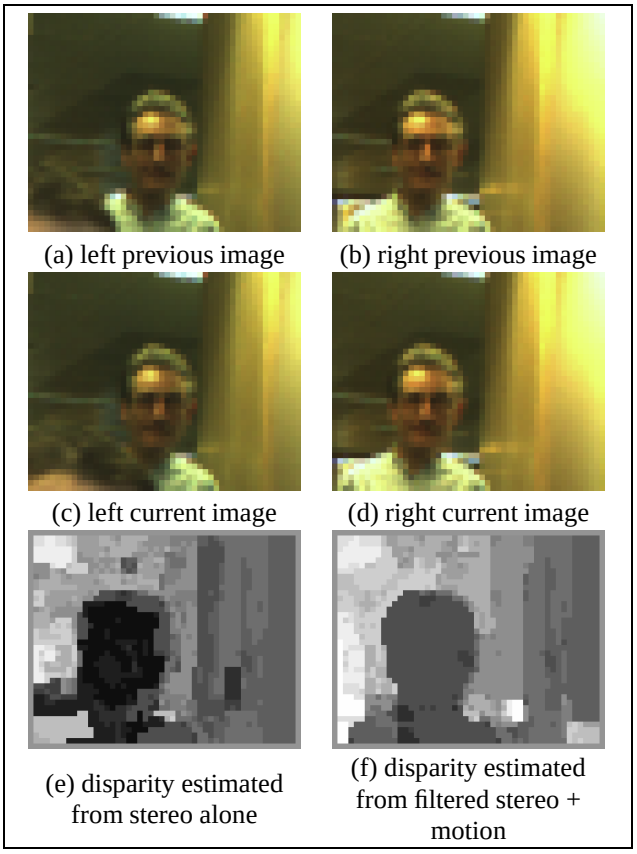
The estimated motion vectors are generally somewhat noisy, but nevertheless produce visually appealing results. Figure 5 shows a typical motion field recovered from another sequence.

The full filtering algorithm for the examples shown takes around 5 s per frame in a C++ implementation running on a 2.2GHz Intel Xeon workstation. For comparison, the disparity-only computation on this small image patch takes 330 ms per frame; comparing with the state of the art suggests there is substantial room for improvement if performance were critical.

6. Conclusions

An algorithm was presented to solve the temporal stereo + motion problem. We believe this is the first such algorithm to obtain dense disparity and motion estimates using a coherent probabilistic framework with physically correct occlusion labels. The approach models a two-frame stereo + motion problem as a single MRF, and extends to the multi-frame case by using temporal filtering in the same MRF framework.

The results confirm that dense stereo + motion produces superior results to stereo alone. The estimates for both stationary and moving objects are stabilized, exhibiting less flicker. Additionally, there are certain image regions in which stereo alone has no information, but stereo + motion



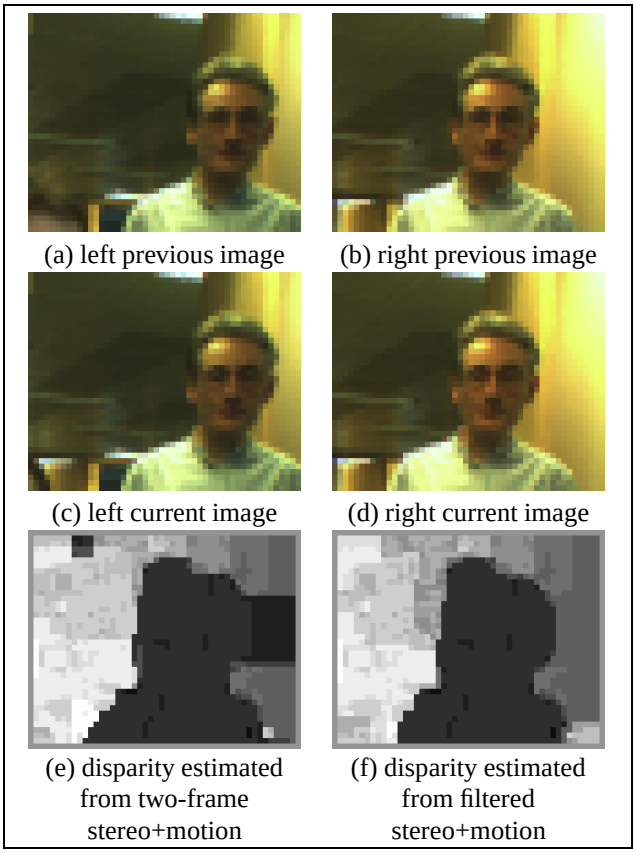
(a) left previous image (b) right previous image

(c) left current image (d) right current image

(e) disparity estimated from stereo alone

(f) disparity estimated from filtered stereo + motion

Figure 3: **Stereo+motion estimates disparity through transient occlusions.** *Darker shades of disparity indicate nearer objects. A fast-moving occlusion has appeared in the bottom corner of the left current image (c) but not yet in the right (d). The stereo computation alone (e) does not have enough information to accurately estimate the foreground disparity in this region causing a large artifact, but the filtered stereo+motion algorithm (f) correctly uses information from previous timesteps to recover a reasonable disparity estimate. Although a dense motion field has been estimated in (f), it is omitted for clarity. Note that the far right column of the right sequence is permanently occluded and disparity estimates there are based purely on spatial smoothness constraints.*



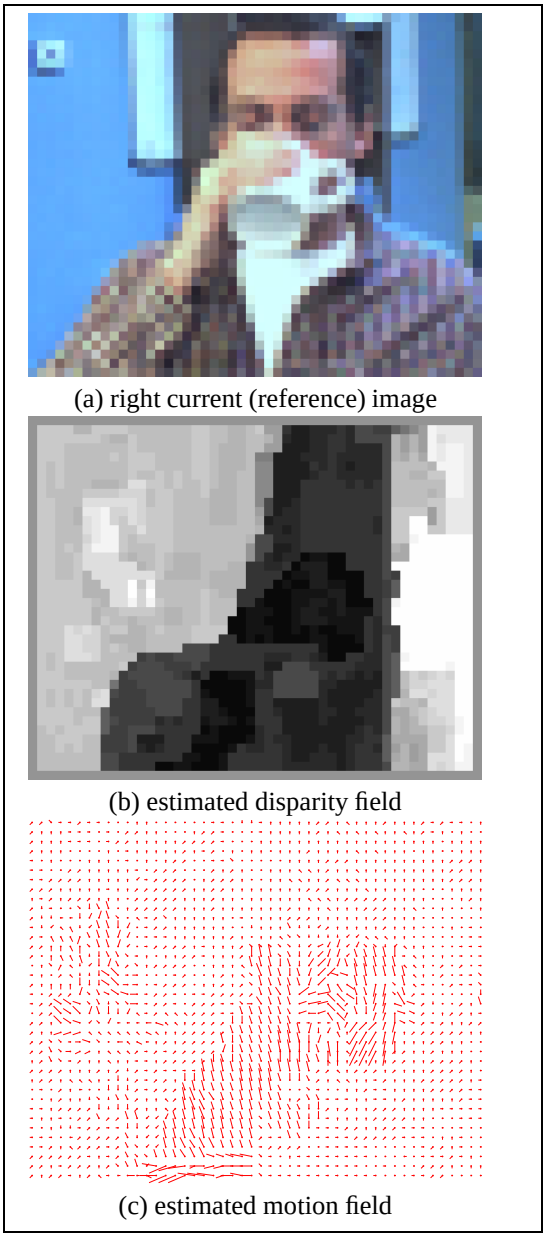
(a) left previous image (b) right previous image

(c) left current image (d) right current image

(e) disparity estimated from two-frame stereo+motion

(f) disparity estimated from filtered stereo+motion

Figure 4: **Stereo+motion propagates disparity estimates through multiple frames.** *Darker shades of disparity indicate nearer objects. The foreground person has stopped moving, and there is a large left occlusion in the textureless area on the right hand side of the image. The two-frame stereo computation (e) has no information about the disparities in this occluded region and the lack of texture causes a large artifact. The filtered stereo+motion estimate (f) correctly propagates disparity estimates from previous frames to stabilise the difficult region. Note that the far right column of the right sequence is permanently occluded and disparity estimates there are based purely on spatial smoothness constraints.*



(a) right current (reference) image

(b) estimated disparity field

(c) estimated motion field

Figure 5: **Disparity and motion are estimated simultaneously.** Darker shades of disparity indicate nearer objects. The right hand (reference) image is shown (a) along with the estimates of the disparity (b) and motion (c) fields generated by the filtered stero+motion algorithm. The person is lifting the cup to his lips. Note that the person's left shoulder is entirely occluded in the left hand sequence, hence the disparity is mis-estimated.

does have information in (the majority of) those regions, and can therefore infer correct disparity and motion fields there.

The clearest opportunity for future work is in decreasing the computational expense of the algorithm, and the most obvious avenue for this is a multi-scale approach. This is presently an object of active research.

A. Probabilistic interpretation of truncated linear energy function

Truncated linear energy functions of the form

$$E(z, z') = \min(C, K\|z - z'\|) \tag{11}$$

are popular because they are both intuitively plausible and convenient for computation. But it's not clear how the parameters C, k should be chosen. Here we describe a probabilistic model that is approximated by truncated linear energy functions. The parameters can then be set by choosing a physically reasonable probabilistic model.

Suppose a random non-negative integer s is either an outlier, with probability $\beta \ll 1$, or an inlier, with probability $1 - \beta$. Outliers are distributed uniformly on $0, 1, \ldots, M-1$, for some maximum value M. Inliers have a geometric distribution with mean λ, so the probability of an inlier taking the value s is

$$\frac{1}{(1+\lambda)}\left(\frac{\lambda}{1+\lambda}\right)^s.$$

We assume $\lambda \ll M$, so only negligible errors are introduced if the inlier distribution is restricted to $0, 1, \ldots, M-1$ by assigning zero probability for $s \geq M$. The mixture distribution of inliers and outliers is distributed as

$$P(s) = \frac{\beta}{M} + \frac{1-\beta}{(1+\lambda)}\left(\frac{\lambda}{1+\lambda}\right)^s \tag{12}$$

for $s \in 0, 1, \ldots, M-1$, and $P(s) = 0$ otherwise. Call $P(s)$ the *robust geometric distribution* with *maximum deviation* M, *outlier fraction* β, and *inlier mean* λ. When necessary, the more explicit notation $P(s; M, \beta, \lambda)$ will be used.

We now make a rather brutal approximation:

$$P(s) \approx \max\left(\frac{\beta}{M}, \frac{1-\beta}{(1+\lambda)}\left(\frac{\lambda}{1+\lambda}\right)^s\right), \quad s \in 0, 1, \ldots, M-1. \tag{13}$$

The approximation has a small absolute maximum error of β/M; the relative error is also low for most values of s, but rises to 100% at the value for which the two components of the max are equal. Nevertheless, this is no cause for concern: our probabilistic model was only approximate to begin with, and (13) retains the qualitative shape of (12).

Finally, (13) can be converted to an energy function as follows:

$$\begin{aligned}\mathsf{Energy}(s) &= -\log(p(s)) \\ &= \min\left(-\log\left(\frac{\beta}{M}\right), \right. \\ &\qquad \left. -\log\left(\frac{1-\beta}{1+\lambda}\right) - s\log\left(\frac{\lambda}{1+\lambda}\right)\right) \\ &= \text{const}+ \\ &\qquad \min\left(\log\left(\frac{M(1-\beta)}{\beta(1+\lambda)}\right), s\log\left(1+1/\lambda\right)\right)\end{aligned}$$

Note that as energy functions are defined only up to an arbitrary constant, we can ignore the constant in this final expression. Comparing with (11), we see that we should take $s = \|z - z'\|$, $C = \log\left(\frac{M(1-\beta)}{\beta(1+\lambda)}\right)$, and $K = \log(1 + 1/\lambda)$. The problem has been reduced to choosing physically reasonable values for the maximum deviation M, the outlier fraction β, and the inlier mean λ.

Another useful way of thinking of this analysis is to allow M to grow very large, while scaling λ and s to remain fixed as a proportion of M. Writing $\mu = \lambda/M$, $t = s/M$, and using standard approximations for large M, results in

$$\mathsf{Energy}(t) = \log M + \min\left(\log\frac{1}{\beta}, -\log\frac{1}{\mu} + \frac{t}{\mu}\right). \quad (14)$$

This is a particularly convenient expression when t is augmented by a binary flag, indicating whether or not t was generated by the uniform distribution on $\{0, 1, \ldots M-1\}$ or the robust geometric distribution with parameters $M, \beta, \mu M$. The energy function for the uniform distribution is just $\log M$; subtracting this constant from both energy functions results in

$$\mathsf{Energy}(t|\text{uniform}) = 0, \quad (15)$$

$$\mathsf{Energy}(t|\text{robust geometric}) = \min\left(\log\frac{1}{\beta}, -\log\frac{1}{\mu} + \frac{t}{\mu}\right). \quad (16)$$

The function $\mathsf{Energy}(t|\text{robust geometric})$ is plotted in figure 6.

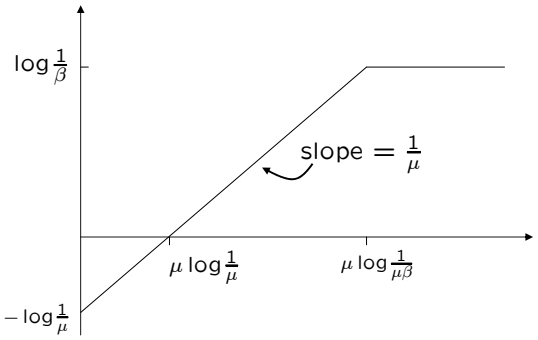


Figure 6: The robust geometric energy function (16), plotted as a function of t.

References

[1] Simon Baker and Iain Matthews. Lucas-Kanade 20 years on: A unifying framework. *Int. J. Comput. Vision*, 56(3):221–255, 2004.

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

[3] A. Blake and et al. Bi-layer segmentation of binocular stereo video. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2005.

[4] Y.L. Chang and J.K. Aggarwal. Line correspondences from cooperating spatial and temporal grouping processes for a sequence of images. *Computer Vision and Image Understanding*, 67(2):186–201, 1997.

[5] 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*, 2003.

[6] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient belief propagation for early vision. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2004.

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

[8] A.Y.K. Ho and T.C. Pong. Cooperative fusion of stereo and motion. *Pattern Recognition*, 29(1):121–130, 1996.

[9] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proc. Int. Conf. on Computer Vision*, 2001.

[10] C. Leung, B. Appleton, B. C. Lovell, and C. Sun. An energy minimisation approach to stereo-temporal dense reconstruction. In *Proc. Int. Conf. on Pattern Recognition*, 2004.

[11] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vision*, 2002.

[12] J. Shao. Generation of temporally consistent multiple virtual camera views from stereoscopic image sequences. *Int. J. Comput. Vision*, 47(1-3):171–180, 2002.

[13] J. Sun, H.-Y. Shum, and N.-N. Zheng. Stereo matching using belief propagation. In *Proc. European Conf. on Computer Vision*, pages 510–524, 2002.

[14] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proc. Int. Conf. on Computer Vision*, volume 2, pages 722–729, 1999.

[15] W. Wang and J.H. Duncan. Recovering the three dimensional motion and structure of multiple moving objects from binocular image flows. *Computer Vision and Image Understanding*, 63(3):430–446, 1996.

[16] A. Waxman and J. Duncan. Binocular image flows: Steps towards stereo-motion fusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):715–729, 1986.

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

[18] J. Yedidia, W. Freeman, and Y. Weiss. *Exploring Artificial Intelligence in the New Millennium*, chapter Understanding Belief Propagation and its Generalizations. Elsevier Science, 2003.

[19] L. Zhang, B. Curless, and S. M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume 2, pages 367–374, 2003.