Disparity Component Matching for Visual Correspondence

Yuri Boykov
Mathematics Department
Carnegie-Mellon University
Pittsburgh, PA
yura@andrew.cmu.edu

Olga Veksler
Ramin Zabih
Computer Science Department
Cornell University
Ithaca, NY 14853
olga@cs.cornell.edu rdz@cs.cornell.edu

Abstract

We present a method for computing dense visual correspondence based on general assumptions about scene geometry. Our algorithm does not rely on correlation, and uses a variable region of support. We assume that images consist of a number of connected sets of pixels with the same disparity, which we call disparity components. Using maximum likelihood arguments, at each pixel we compute a small set of plausible disparities. A pixel is assigned a disparity d based on connected components of pixels, where each pixel in a component considers d to be plausible. Our implementation chooses the largest plausible disparity component; however, global contextual constraints can also be applied. While the algorithm was originally designed for visual correspondence, it can also be used for other early vision problems such as image restoration. It runs in a few seconds on traditional benchmark images with standard parameter settings, and gives quite promising results.

1 Introduction

Many problems in early vision are ill-posed, i.e. cannot be uniquely solved without additional constraints. Scene geometry can provide constraints that make these problems well-posed [6]. In this paper we describe an approach to early vision which exploits scene geometry in a novel manner. We assume that images consist of a number of connected sets of pixels with the same disparity, which we call *disparity components*. Note that the boundaries of a disparity component are discontinuities, by definition. A disparity component may be rectangular, or of any other shape.

Our method first computes for every pixel a small set of plausible disparities, which are disparities more likely to be the pixel's true disparity than not given the observed intensities. Then for each disparity we compute connected components among those pixels for which that disparity is plausible. Finally, for each pixel we select the plausible disparity with the largest connected component.

In section 2 we describe our approach. In section 3 we show empirical results from our method, on both real and synthetic data. We conclude by discussing the relationship between our work and previous methods, and proposing a number of extensions.

2 Disparity component matching

Let P be a pixel, $I(P)$ its intensity in one image and $I'(P)$ its intensity in the other image. We will denote a disparity by d, and the set of possible disparities by D. In stereo, disparities are typically restricted to lie along a scanline, while motion involves 2D disparities. We will write the statement that pixel P has disparity d by P^d. If P^d holds, then

$$I(P) = I'(P + d) + \nu(P), \tag{1}$$

where ν is the measurement error, which includes such unmodeled phenomena as analog noise and changes in viewing angle and illumination. If S is a connected set of pixels, we define

$$S^d = \bigwedge_{P \in S} P^d. \tag{2}$$

We will call S^d a *disparity component hypothesis*, or a component hypothesis for brevity.

The number of component hypotheses is, of course, exponential. However, almost all hypotheses can be pruned by a maximum likelihood argument. Define P^d to be *plausible* if the likelihood of P^d is greater than the likelihood of $\neg P^d$ for the observed data. A component hypothesis S^d is plausible when P^d is plausible for all P in S. Section 2.3 gives the precise definition of plausibility, along with a simple algorithm for plausibility testing.

2.1 Our method

Our method has two steps. The first step only considers a small number of disparities per pixel.[1] The second step chooses from among these disparities based on the component hypotheses.

In the first step we find the plausible component hypotheses S^d. For each fixed disparity d we first test all P^d's for plausibility, as described in section 2.3. Then for each d we group pixels with plausible P^d's into connected components, thus determining all plausible S^d's for the given d.

In the second step of our method, for each pixel P we pick a disparity from the plausible hypotheses S^d containing P. The simplest method is to rank the component hypotheses S^d by decreasing size of S. Note that other sources of information, including global contextual constraints, may also be used for ranking (see section 5.2.) We thus assign P the disparity d such that S^d is the largest plausible S^d containing P.

2.2 Efficiency

If there are n pixels and m disparities, the running time of our method is $O(nm)$. We will show below that plausibility testing can be done in $O(nm)$ time. In step 1 we perform $O(m)$ connected component computations, which are $O(n)$ time. Step 2 maximizes over m quantities at each of n pixels. In practice, our initial implementation takes a few seconds per image to compute depth. For example, the tree image shown in section 3.2.1 is 256 by 240, and took 3 seconds to process (with 10 disparities, on a 50-MHz sparc).

2.3 Plausibility testing

Consider some fixed disparity $\hat{d}$ for pixel P. We need to choose between the two hypotheses:

$$\begin{aligned} H_0 &: P^{\hat{d}} \\ H_1 &: \neg P^{\hat{d}}, \end{aligned}$$

$P^{\hat{d}}$ is plausible if H_0 is more likely than H_1. The statement that the pixel P is occluded will be represented by P^o.

Assume that the function $f(\ \cdot \mid i'\ )$ specifies the noise model, that is the distribution of intensity of a pixel in the first image given intensity i' of the corresponding pixel in the second image. For any event E we define $\Pr'(E) = \Pr(E|I'(P_1), \ldots, I'(P_n)) = \Pr(E|I')$, which is the probability of E conditioned on all the observed intensities from the image I'. Similarly we define $\Pr'(E|F) = \Pr(E|F, I')$. Then

$$\Pr'(\ I(P) \mid P^d\ ) = f(\ I(P) \mid I'(P+d)\ ).$$

We choose between H_0 and H_1 by comparing the likelihoods $\Pr'(I(P)|H_0)$ and $\Pr'(I(P)|H_1)$. Obviously, we have

$$\Pr'(I(P)|H_0) = f(I(P)|I'(P+\hat{d})).$$

To compute $\Pr'(I(P)|H_1)$ we proceed as follows:

$$\Pr'(I(P)|H_1) = \frac{\Pr'(I(P), H_1)}{\Pr'(H_1)}$$

$$= \frac{\sum_{d \neq \hat{d}} \Pr'(I(P), P^d) + \Pr'(I(P), P^o)}{\Pr'(H_1)}$$

$$= \frac{\sum_{d \neq \hat{d}} f(I(P)|I'(P+d))\Pr'(P^d)}{\Pr'(H_1)}$$

$$+ \frac{\Pr'(I(P)|P^o)\Pr'(P^o)}{\Pr'(H_1)}$$

To prefer H_0 over H_1 we should have

$$f(I(P)|I'(P+\hat{d}))$$

$$> \frac{\sum_{d \neq \hat{d}} f(I(P)|I'(P+d))\Pr'(P^d)}{\Pr'(H_1)}$$

$$+ \frac{\Pr'(I(P)|P^o)\Pr'(P^o)}{\Pr'(H_1)}.$$

Multiplying both sides by $\Pr'(H_1)$ and then adding $f(I(P)|I'(P+\hat{d})) \cdot \Pr'(H_0)$ gives

$$f(I(P)|I'(P+\hat{d}))$$

$$> \sum_d f(I(P)|I'(P+d))\Pr'(P^d)$$

$$+ \Pr'(I(P)|P^o)\Pr'(P^o),$$

where the summation is over all possible values of disparity d.

Assuming that the probability of occlusion $\Pr'(P^o)$ is given by some constant q and that $\Pr'(I(P)|P^o) = \frac{1}{|R|}$ where $|R|$ is the number of all possible intensities, we have the inequality

$$f(I(P)|I'(P+\hat{d}))$$

[1] A somewhat similar model is used by Spoerri and Ullman [7] for motion segmentation.

$$> \sum_d f(I(P)|I'(P+d))\Pr'(P^d) + \frac{q}{|R|}.$$

We assume that the prior probabilities of all disparities are equal. This implies that $\Pr'(P^d)$ does not depend on d. Consequently,

$$q + |D|\Pr'(P^d) = 1 \quad \forall d,$$

where $|D|$ denotes the number of all possible disparities. Finally, the comparison test can be equivalently rewritten as

$$f(I(P)|I'(P+\hat{d}))$$
$$> \frac{1-q}{|D|} \cdot \sum_d f(I(P)|I'(P+d)) + \frac{q}{|R|}, \qquad (3)$$

so that we accept hypothesis H_0 if the likelihood of disparity $\hat{d}$ is larger than the weighted sum of the *average likelihood* of all possible disparities and a cut off constant $\frac{q}{|R|}$.

We can use any noise model $f(\ \cdot \mid i'\ )$ in formula (3). If f satisfies[2] $f(\ x \mid i'\ ) = f(x - i')$ and if ΔP^d denotes $I(P) - I'(P+d)$ then test (3) is equivalent to

$$f(\Delta P^{\hat{d}}) > \frac{1-q}{|D|} \cdot \sum_d f(\Delta P^d) + \frac{q}{|R|}. \qquad (4)$$

This test is equivalent to

$$|\Delta P^{\hat{d}}| < \epsilon(P)$$

where ϵ depends on the noise model f. This provides a computationally trivial way to test plausibility. If the noise model f and the parameters $|R|$ and q are specified in advance, then $\epsilon(P)$ can be computed in $O(m)$ time at each pixel.

3 Experimental results

To run the algorithm on real images, we need to select a noise model f, as well as a percentage of occlusion q. We use $q = 4\%$ for all the experiments. For the noise model we chose $f(\ i \mid i'\ ) \propto e^{\frac{-(i-i')^2}{2\sigma^2}}$. We experimented with different values of σ as shown below.

Theoretically, f corresponds to errors due to camera noise. To handle other measurement errors, like errors due to changes in illumination conditions and in viewing angle, we introduce gain and bias parameters g and b that adjust pixel intensities in the right image. Equation 1 becomes:

$$I(P) = (g \times I'(P+d) + b) + \nu(P). \qquad (5)$$

To implement the new measurement error model, connected components are computed for all disparities d in D, for all values of g in G, and for all b in B, where G and B are some fixed ranges. We also need to change formula (4) to take the new variation parameters g and b into account. Similarly to the derivation in section 2.3, we obtain the following plausibility test:

$$f(\Delta P^{\hat{d},\hat{g},\hat{b}}) >$$
$$\frac{1-q}{|D||G||B|} \cdot \sum_{d,g,b} f(\Delta P^{d,g,b}) + \frac{q}{|R|}. \qquad (6)$$

Here $\Delta P^{d,g,b}$ denotes $I(P) - (g \cdot I'(P+d) + b)$. Using equation 6 instead of 4 slows our algorithm by approximately a factor of 10; however, we believe that this can be significantly reduced.

Finally, for each pixel P we chose disparity d, gain g, and bias b corresponding to the largest connected component containing P. This is a simple way to deal with g and b parameters (see [1] for a different solution). For all real images shown here, we vary g from 0.9 to 1.1 in intervals of 0.1 and b from -14 to 14 in intervals of 1.

To produce a feasible disparity map, we check the output for double assignments in the following way. If two pixels in the left image are mapped to the same pixel in the right image, then the pixel which belongs to a bigger connected component gets assigned to that pixel in the right image. This approach not only produces feasible disparity assignments, but it also automatically handles occlusions.

On our disparity images the brighter colors correspond to larger disparities, and the very brightest color to the pixels for which no matches were found. The disparity images are speckled with white pixels, indicating that for these pixels no matches were found. But the majority of white points is to the left of objects, since these points are occluded in the right image. We also clean up isolated pixels by forcing pixels that are surrounded by a single disparity to take on their neighbor's disparity.

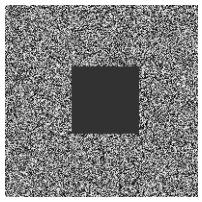

Figure 1. Random dot motion image illustrating aperture problem

3.1 Results on Synthetic Images

3.1.1 The aperture problem

Figure 1 shows the left image of a random dot motion sequence. In the right image, the entire image in translated

[2] This holds if $f(\ \cdot \mid i'\ )$ is uniform, Gaussian, or any other symmetric distribution function centered at i'.

uniformly. Note the large textureless region in the center, which creates problems for existing algorithms. Our method performs correctly for this image, assigning all pixels the correct translation. Our adaptive window scheme ensures that the entire image is treated as a single window. This effectively propagates information from the high-texture regions at the outskirts of the image into the low-texture regions in the middle of the image. Kanade and Okotumi's algorithm [5] may also succeed in this situation, as long as the textureless region is rectangular.

3.2 Results on Real Images

The digital imagery shown below, including both the original images and the results from various algorithms, can be accessed from the web. The address is http://www.cs.cornell.edu/home/rdz/adaptive.

3.2.1 The "Tree" Pair, $\sigma = 1.7$

Figure 2 shows the left image of the tree pair from SRI. Figure 3 shows disparity maps obtained from normalized correlation with window sizes 5 and 8. One can clearly see the problems of correlation using a fixed window: with a small window there are many wrong matches, while large windows perform poorly at the disparity discontinuities.

Figure 4 shows the disparity map obtained by our algorithm. The edges of the tree trunk, branches, and stump are significantly sharp and they correspond closely to the actual tree shapes. Many of the finer details of the scene are also found accurately. For example, consider the leaves at the top of the closest tree.

3.2.2 The "Meter" Pair, $\sigma = 0.8$

Figure 5 shows the left image of the meter pair from CMU. Figures 6 and 7 show the results of normalized correlation and our algorithm. On the wall of the building there are many areas with low intensity variation. Normalized correlation clearly has trouble producing correct answers in these areas. Even if we use a large window of 20×20 pixels, there are still a lot of large bright spots on the wall, which are obviously not matched correctly.

Results produced by our algorithm are by far more coherent, and most of the edges are very accurate. Especially interesting is the long thin pole located between the last 2 meters of the picture. It's clear that a rectangular window algorithm has slim chances of assigning the right disparity to this pole, unless the window width is not much bigger than the width of the pole.

4 Related work

The fundamental distinction between our method and previous work lies in the manner in which we exploit scene geometry. Most methods for computing visual correspondence make assumptions about the underlying scene. For example, one popular method for computing motion [3] assumes that the underlying motion varies smoothly (in fact, a large number of methods based on regularization make such an assumption [6]). These methods tend to produce poor results near discontinuities. In the last decade, a number of methods have been developed that also handle discontinuities. Markov Random Fields [2], for example, explicitly model discontinuities using methods from statistical physics. Our method is comparable to MRF's, in that we avoid crossing discontinuities. MRF's, however, require exponential running time, while our method is linear in the number of pixels and disparities. MRF's also fundamentally handle spatially local interactions, while our method is best suited for global constraints. For example, our method can exploit global contextual constraints that do not fit naturally into an MRF framework.

Most correspondence methods compare rectangular windows, which implicitly assumes that images consist of fronto-parallel planes undergoing translational motion. A recent paper by Stewart *et al.* [8] justifies this assumption under very general assumptions concerning scene geometry. Several researchers have designed adaptive methods that iteratively compute disparity, attempting to avoid crossing discontinuities. Jones and Malik [4] use linear spatial filters, and compute the largest scale that does not straddle a discontinuity. Kanade and Okotumi [5] model local disparity variation, in order to handle sloped surfaces and discontinuities. Their method grows the window locally using a greedy method that minimizes the uncertainty of their estimate. Once the correct window size is found at each point, correspondence is computed using SSD correlation.

Our method is like existing adaptive techniques [4, 5] in that we choose a different window at each pixel. Previous methods effectively search for the best rectangular window at each point, over some limited range. In contrast, our method efficiently handles windows of arbitrary shape and size, without performing significant search. A more fundamental distinction is that the previous adaptive window methods are based on correlation, while our approach is not.

5 Extensions

We are primarily interested in two extensions to our basic method. The first involves generalizing our work to handle other problems in early vision besides correspondence, such as image restoration. The second is to choose more intelligently among the plausible component hypotheses, for

example by incorporating global contextual constraints.

5.1 A variable neighborhood approach to early vision

While our method was originally designed for computing correspondence, it is applicable to a range of early vision problems. Many of these problems involve reconstructing a piecewise constant function from noisy data. Existing methods make widespread use of rectangular windows of fixed size, primarily for efficiency. However, fixed rectangular windows poorly model the boundaries of real world objects. Our work provides a new approach to these problems.

For example, consider the problem of image restoration, in which the intensities of a piecewise constant image are corrupted by noise, and the goal is to reconstruct the original image. We approach this problem by considering hypotheses of the form P^i, which states that the pixel P has the intensity i in the original (uncorrupted) image. A hypothesis P^i is plausible if i is more likely to be the original intensity of P than not for our observed data. Plausibility testing for this problem is very similar to section 2.3, and the absence of occlusions simplifies the equations. We then compute the plausible components S^i by forming connected components from the pixels P with plausible hypotheses P^i. Finally, we assign each pixel P the intensity i such that P^i is the largest plausible S^i containing P.

5.2 Ranking the plausible hypotheses

In the second step of our method we choose among the plausible component hypotheses. Our work to date has used a simple, local criterion: every pixel chooses the largest plausible component hypothesis containing it. However, it is possible to make this choice in a more interesting manner. Suppose, for example, that a pixel participates in two component hypothesis S and S' (we omit disparity superscripts for legibility). S might be smaller than S', yet still be a better hypothesis. For example, S' could be highly irregular in shape and riddled with holes, while S could be a compact, simple region. It is straightforward to extend our basic method to rank component hypotheses based on criterion other than size.

We are particularly interested in the use of global contextual constraints to rank component hypotheses. In a particular task, some hypotheses may be more reasonable than others. For example, suppose the camera is pointed at a ground plane, as in figure 2. The components at the bottom of such images tend to be horizontally elongated.

Another extension would also consider the measurement errors that result from a given S. If all the pixels in S have disparity d, then the measurement errors is $I(P) - I'(P+d)$ for each P in S. Note that due to the way we construct S there will be no pixel P with a measurement error greater than $\epsilon(P)$ (otherwise, P^d is not plausible, and P would not be included in S). We can, for example, require that the measurement errors from S be unbiased. Other methods for evaluating the measurement errors could also be used, such as minimizing the squared error.

A final extension concerns the manner in which we assign disparities to pixels, once we have ranked the plausible component hypotheses. Our current method is greedy, and purely local. As a consequence, it is possible that P selects disparity d based on the component S, but that the other pixels in S select some different disparity. We are designing other methods which enforce a kind of consistency in the selection process.

References

[1] Ingemar Cox, Sunita Hingorani, Satish Rao, and Bruce Maggs. A maximum likelihood stereo algorithm. *Computer Vision, Graphics and Image Processing*, 63(3):542–567, 1996.

[2] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[3] Berthold Horn and Brian Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[4] David Jones and Jitendra Malik. A computational framework for determining stereo correspondence from a set of linear spatial filters. In *2nd European Conference on Computer Vision*, pages 395–410, 1992.

[5] Takeo Kanade and Masatoshi Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, September 1994.

[6] Tomaso Poggio, Vincent Torre, and Christof Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[7] Anselm Spoerri and Shimon Ullman. The early detection of motion boundaries. In *International Conference on Computer Vision*, pages 209–218, 1987.

[8] Charles Stewart, Robin Flatland, and Kishore Bubna. Geometric constraints and stereo disparity computation. *International Journal of Computer Vision*, 20(3):143–168, December 1996.

Figure 2. Left tree image.

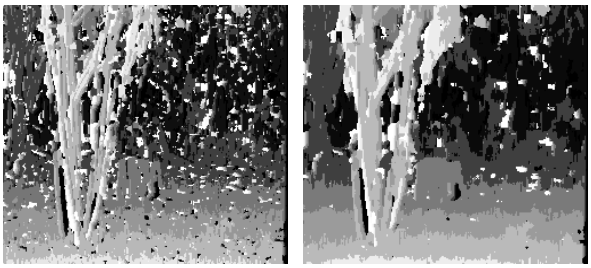

Figure 3. Disparity maps from normalized correlation, 5×5 and 8×8 windows.

Figure 4. Disparity map from our algorithm.

Figure 5. Left meter image.

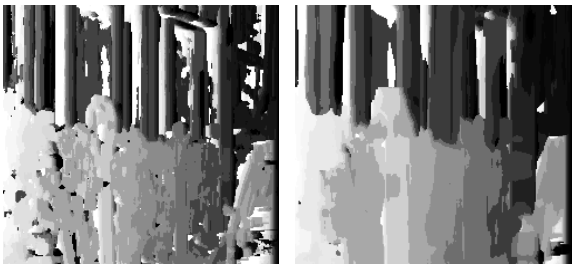

Figure 6. Disparity maps from normalized correlation, 9×9 and 20×20 windows.

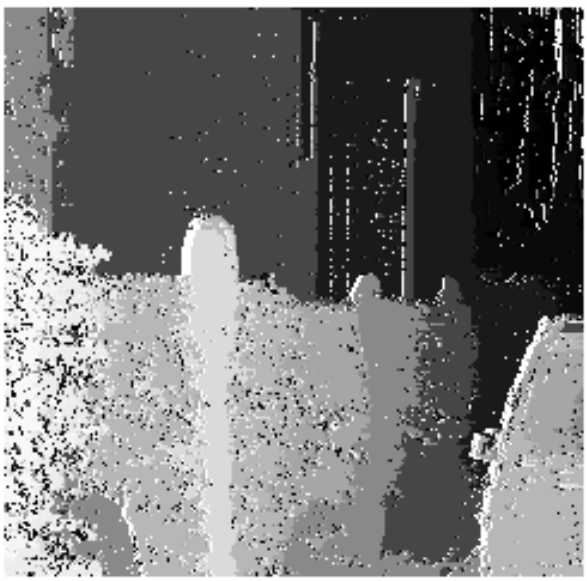

Figure 7. Disparity map from our algorithm.