Matching Images by Comparing their Gradient Fields

Daniel Scharstein*

Department of Computer Science
Cornell University, Ithaca, NY 14853, USA
schar@cs.cornell.edu

Abstract

We present a simple yet powerful method to perform point-to-point matching between two images. The method uses an evidence measure*, whose value for a given displacement reflects both the similarity between two locations and the confidence in a correct match. The measure is based on the gradient fields of the images, and can be computed quickly and in parallel. Accumulating the evidence measure for different displacements allows (1) stable computation of correspondences without smoothing across motion boundaries, and (2) detection of dominant motions. The method works well both on highly textured images and on images containing regions of uniform intensities, and can be used for a variety of applications, including stereo, motion, and object tracking.*

1 Introduction and related work

A fundamental problem in computer vision is the so-called *correspondence problem*, that is, to establish point-to-point correspondences across a pair of images. Most algorithms for computing correspondence have a *point-oriented* control strategy: For each location in one image, find the displacement that aligns it with the best matching location in the other image. The method presented here uses a *displacement-oriented* control strategy: Given a certain displacement, find all the locations that match well. Under the assumption that the motion between two images can be locally approximated by pure translation, near points corresponding to the same object have similar displacements, which can be detected by accumulating evidence for matches over a larger area.

To compare locations in two images, most existing methods rely on a *similarity* criterion reflecting how well two locations in the two images resemble each other, and also sometimes on a *confidence* criterion reflecting the likelihood that a match is correct [1]. While these two criteria are usually treated separately, our method uses a single measure based on the *gradient fields* of the images, which—given a certain displacement—gives a (strong) positive response where points match with (high) confidence, a negative response where there is a clear mismatch, and zero response in regions where there is neither evidence for a match nor evidence against. This approach has the following advantages:

- The evidence measure is local and can be computed quickly and in parallel.
- For a given displacement, the measure can be accumulated by averaging over a larger area. The average value represents evidence for or against a match, enabling the use of a displacement-oriented control strategy.
- Finding maxima in the accumulated measure is a stable way of computing correspondences without smoothing across motion boundaries.

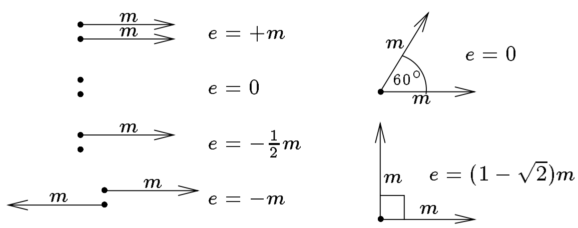
For reviews of correspondence methods see [1, 2]. Seitz [8] uses local gradients for object recognition. Prazdny [6] describes a stereo algorithm that collects support for different disparity hypotheses similarly to our method, but requires an initial set of possible disparity hypotheses collected by explicit feature matching. Coombs and Brown [3] describe an active stereo vision system that finds points at the depth of fixation by means of a feature-based *zero-disparity filter*. Olson and Lockwood [5] describe a way of disparity filtering using a multi-scale correlation method. Both approaches differ from ours in that they do not return a measure that reflects the evidence for a match at a certain position. Discussion of other related work can be found in the full paper [7].

2 Measuring evidence for matches

As mentioned above, our method combines the notions of similarity and confidence (or distinctiveness)

*This work was supported in part by NSF PYI grant IRI-9057928 and matching funds from Xerox Corp.

into a single measure of *evidence* for or against a match based on the two gradients at a certain location. In particular, if $\mathbf{g}_L$, $\mathbf{g}_R$ are the two gradient vectors to be compared, we use their average magnitude $\overline{m} = (|\mathbf{g}_L| + |\mathbf{g}_R|)/2$ to represent confidence, and the negated magnitude of their difference $-d = -|\mathbf{g}_L - \mathbf{g}_R|$ to represent similarity. We define the *evidence* for a match to be the sum of these two terms: $e = \overline{m} - d$. For illustration, we compare vectors of length m and 0 in the following examples:



If both gradients are zero, one can't tell whether or not they match, and consequently $e = 0$. (The measure ignores the original intensities, although one can argue that they provide additional information. However, comparing absolute intensities has proven to be not very stable in practice.) Note that e can also be zero for two non-zero gradient vectors, for example, in the case of two vectors of equal length defining an angle of 60°. Intuitively, this reflects the situation where the directions of gradients are too different to be considered a match, but not different enough to be counted as a mismatch. Figure 1 is a contour plot of e for comparing any vector (x, y) to the unit vector $(1, 0)$. The unit vector of angle 60° is shown as an example; note that its endpoint lies on the $e = 0$ curve.

We now extend the measure to entire images. Let I_L, I_R be two images, and let $\mathbf{G}_L$, $\mathbf{G}_R$ be their gradient vector fields. For a given displacement $\delta = (\delta_x, \delta_y)$, the *evidence* E_δ for a match at (x, y) is

$$\begin{aligned} E_\delta(x,y) = \; & \tfrac{1}{2}\left(|\mathbf{G}_L(x,y)| + |\mathbf{G}_R(x+\delta_x, y+\delta_y)|\right) \\ & -|\mathbf{G}_L(x,y) - \mathbf{G}_R(x+\delta_x, y+\delta_y)|. \end{aligned}$$

In order to apply the method to discrete images, we approximate the gradients by differences after initial smoothing to compensate for quantization error and noise. In the experiments reported here, we used a Gaussian filter with $\sigma = 0.5$. It should be noted that E_δ can be computed very fast, since only a few floating point operations and a single square root is needed at each pixel (the two magnitudes of gradients $|\mathbf{G}_L|$ and $|\mathbf{G}_R|$ only need to be computed once). The local nature of the computations makes the method ideally suited for a parallel implementation. A sequential implementation on a Sparc workstation takes less than one second to compute E_δ for a 256×256 pixel image.

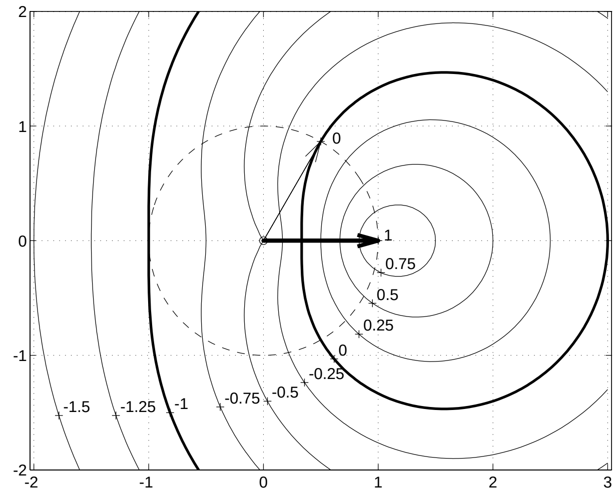


Figure 1: Contour lines of e for the vector $(1, 0)$

3 Accumulating results

To find the best match for an isolated point, all we can do is to maximize E_δ at this point for all δ under consideration. Doing so independently for every point is not very stable and might produce a noisy and inconsistent displacement field. To deal with this problem, motion computation methods usually make the assumption that nearby points have similar displacements, based on the observation that motion in real scenes varies smoothly almost everywhere. Furthermore, it is often assumed that motion can be described locally by pure translation, i.e., rotational components and effects of perspective foreshortening are small enough. Many point-oriented methods utilize the assumption of a smooth motion field *after* computing initial matches by smoothing the displacement field, often employing some confidence measure associated with each match to constrain the smoothing process [4, 1]. The problem is that this tends to smooth over motion discontinuities, which contain important information about the scene geometry.

In contrast, our displacement-oriented method uses the assumption of a smooth motion field *while* finding the matches. The idea is that if a certain displacement δ aligns two matching objects, E_δ will have a strong positive response at the location of the match. By accumulating E_δ over a certain area (i.e., computing the average or smoothing with a Gaussian filter), dominant motions can be detected. That is, only the correct displacement E_δ will yield support for a match over a larger area, thereby creating a maximum among all δ under consideration. Note that our method does not smooth over motion boundaries, since it is not assumed that *all* close pixels have the same disparity. A point on a motion boundary will give rise to a positive response for two different displacements, corresponding to the two different motions. If necessary, the local response at that point can help to break the tie.

E_δ can also be accumulated over very large areas, such as a quarter of the image or even the entire image, to find an initial set of interesting displacements. Since most displacements will only align a small subset of features, only the displacements that align larger parts of the image will yield an above-average response, which can serve to select an initial set of displacements, for which the matching with smaller windows is undertaken. A scale-space approach could be used to speed up the initial selection of interesting displacements. Peaks in the accumulated E_δ as a function of δ can also serve as attention cues for active vision systems.

4 Experiments

A striking experiment is to just observe E_δ for different displacements δ. As test data we use a stereo pair from the street image sequence[1], depicting a woman crossing a street (Figure 2, top row). This image pair is an interesting example in that it contains large regions with little texture. Also, the absolute intensities are quite different between the two images. To illustrate the power of using maxima in the accumulated measure E_δ as attention cues, we have selected the displacements that yield the strongest response (maximal $\sum E_\delta$) in each quadrant of the image. Rows 2 and 3 in Figure 2 show E_δ for the resulting four displacements δ. Gray corresponds to a value of 0, light to positive values, and dark to negative values. Note that these displacements align the dominant features in each quadrant. One can also see that the measure is not sensitive to the brightness difference between the original images.

We implemented a simple stereo matcher that uses the evidence measure to select matches. For a range of different δ, We accumulate E_δ by smoothing with a Gaussian filter with $\sigma = 2$. The disparity at each point is taken to be the displacement that maximizes the accumulated measure at this point. We ran the matcher on two highly textured images from the Stanford tree sequence[2]. The considered range of disparities is $\delta_x = 0 \dots 12$. Simply picking maxima in the accumulated measure already gives surprisingly good results. Figure 3 shows one of the original images and the computed disparities, which are displayed with different gray levels: lighter corresponds to closer, darker to farther away.

[1] The street images were provided by Wilfried Enkelmann.

[2] The tree images were provided by SRI.

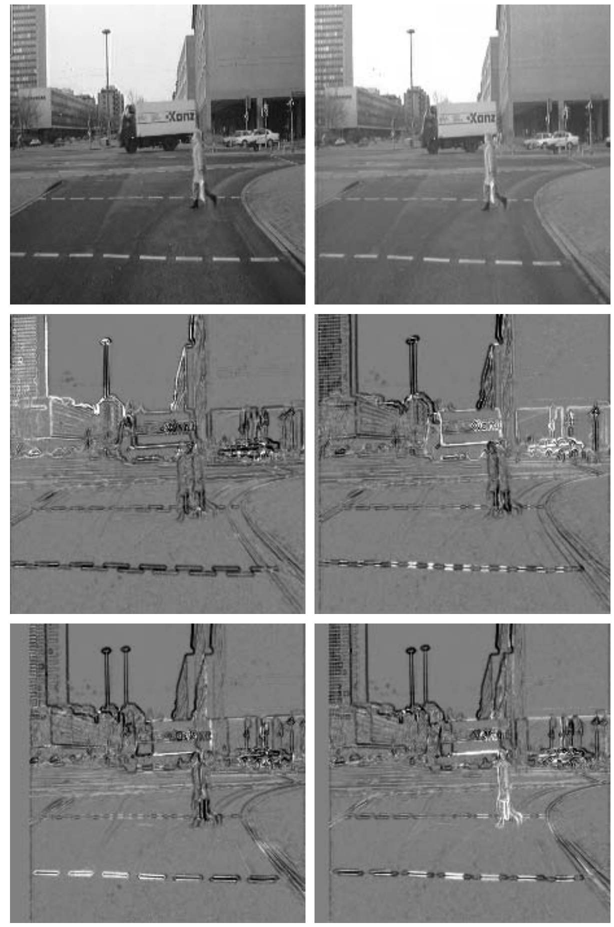


Figure 2: The street image pair and plots of E_δ for the maximizing displacements δ for each quadrant.

The next experiment shows how confidence can be incorporated in the matcher to be able to deal with images with less texture, where it is harder to find clear maxima in the evidence measure. An advantage of the measure we use is that the *value* of the achieved maximum is related to the gradient magnitude at that point, and thus represents the confidence for the match being correct. To demonstrate this, we will use the street image pair described above. Unreliable matches can be suppressed by setting a threshold for the actual achieved maximum at each point. Figure 4 shows the disparities in different gray levels; in the right image all unreliable matches are displayed in black. The considered range of disparities is $\delta_x = -3 \dots 21$. Note that while feature-based matchers try to decide beforehand which locations to match, our method allows the selection of reliable points after the matching process.

To test the method on general motion, we used two images from the cat sequence[3]. This sequence depicts

[3] The cat images were provided by John Woodfill.

Figure 3: Disparities for the tree images.

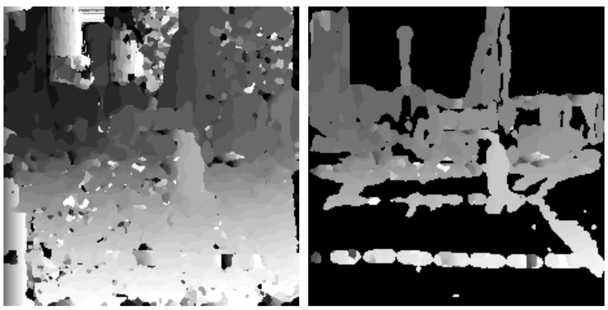


Figure 4: Disparities for the street images. In the right image, uncertain matches are displayed in black.

a cat walking on a lawn in front of some bushes. The camera follows the cat, so that the visual motion of the cat is almost only caused by its (non-rigid) change of shape, whereas the background moves by more than 10 pixels to the left. Like the tree images, the cat images are well textured, so we don't use the confidence information here. Figure 5 shows one of the original images and the x-components of the displacements that maximize the accumulated measure. The considered ranges are $\delta_x = -15 \ldots 4$, $\delta_y = -2 \ldots 1$; accumulation is done with a Gaussian filter with $\sigma = 2$.

5 Conclusion and future work

We have presented a simple yet powerful method to perform point-to-point matching between two images. The method uses an *evidence measure* that is based on the gradient fields of the images and that combines the notions of *similarity* between two locations, and *confidence* for a correct match. The computation of the measure is simple and highly parallelizable. For a given displacement, the measure can be accumulated over a larger area, to collect evidence for or against a match at this location. Using a *displacement-oriented* control strategy that accumulates evidence for a range of different displacements, dominant motions can be detected, which can serve as attention cues in an active vision system. Finding maxima in the accumulated measure is a stable way of computing correspondences without smoothing across motion boundaries.

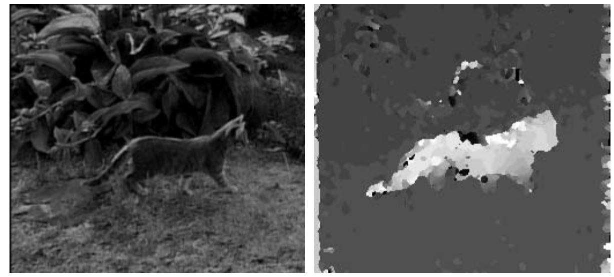


Figure 5: Disparities for the cat images.

The method works well both on highly textured images and on images containing regions of uniform intensities, and can be used for a variety of applications, including stereo vision, motion segmentation, object tracking, and active vision.

A problem with the measure discussed here is that partially aligned intensity edges yield a positive response, which can make it hard to find the component of the displacement that is parallel to this edge. For example, in Figure 4 one can observe errors in the computed disparities of the street marks in the foreground of the scene. This is due to the so-called *aperture problem*, which states that, locally, only the component of displacement in the direction of the intensity gradient can be recovered. A possible way to deal with this problem is to use a measure that is only sensitive to corners. We are currently investigating functions that combine gradients and traditional "cornerness" measures, as well as measures that combine original intensities, first and second order derivatives. See the full paper [7] for other possible extensions.

References

[1] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *IJCV*, 2(3):283–310, 1989.

[2] S. Barnard and M. Fischler. Computational stereo. *ACM Comp. Surveys*, 14(4):553–572, 1982.

[3] D. Coombs and C. Brown. Real-time binocular smooth pursuit. *IJCV*, 11(2):147–164, 1993.

[4] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[5] T. Olson and R. Lockwood. Fixation-based filtering. In *Proc. SPIE*, volume 1825, pages 685–695, 1992.

[6] K. Prazdny. Detection of binocular disparities. *Biological Cybernetics*, 52(2):93–99, 1985.

[7] D. Scharstein. A gradient-based evidence measure for image matching. Cornell CS TR 94-1439, 1994.

[8] P. Seitz. Using local orientation information as image primitive for robust object recognition. In *Proc. SPIE*, volume 1199, pages 1630–1639, 1989.