VISUAL MOTION CORRESPONDENCE BY REGION–BASED APPROACHES

Chiou-Shann Fuh
Comp. Sci. & Info. Eng. Dept.
National Taiwan University
Taipei, Taiwan 106

Petros Maragos
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Luc Vincent
Xerox Imaging Systems
9 Centennial Drive
Peabody, MA 01960

Abstract

This paper presents a correspondence method to determining motion displacement fields in sequences of intensity images where the motion tokens to be matched between consecutive image frames are 2-D regions. These regions contain perceptually important image features. The computation of the 2-D image velocity field is done in three stages: region extraction, region matching, and velocity smoothing. Overall, the proposed region-based methods for computing image velocities are simple, efficient, less computationally complex than intensity correlation methods, and (as our experiments on real images indicate) more robust than iterative gradient methods especially for medium or long-range motion.

1. Introduction

This paper presents a correspondence approach to measuring the 2-D image velocity field by using *regions* as the simple tokens to extract from each image frame and track over time. In recent research considerable attention has been given to edges (e.g., zero-crossings of the Laplacian of a Gaussian) as being perhaps the most desirable features to match in binocular stereopsis or motion analysis. However, without doubting the general usefulness of edges as important image features, we view the region matching as more robust than edge matching, because noise perturbs the coherence of a region less than its boundaries (edges). This was demonstrated by Nishihara [4] who solved the correspondence problem for binocular stereo by cross-correlating the binary regions (sign areas) bounded from the zero-crossing contours of the band-pass filtered images $\nabla^2 G * I$. (∇^2 is the operator $\partial^2/\partial x^2 + \partial^2/\partial y^2$, $G(x,y) = \exp[-(x^2+y^2)/2\sigma^2]/2\pi\sigma^2$ is a Gaussian function with standard deviation (scale parameter) σ, and $*$ denotes 2-D convolution.) Mayhew and Frisby [2] have also found that intensity edges cannot by themselves disambiguate some correspondences in binocular stereopsis unless they are supplemented by region features such as intensity peaks and valleys. Additional strong evidence for the possible effectiveness of blob-like regions is provided by the psychophysical experiments of Ramachandran and Anstis [5], which demonstrated that the human visual system during its first short-term phase of perceiving apparent motion is more likely to detect correspondences between regions of similar brightness or texture before it detects sharp outlines or edges.

2. Region–Based Approaches

Motivated by all the above evidences, in this paper we study and compare several region-based approaches [1] to motion correspondence. The common procedure in all our approaches consists of three stages: (I) *Region Extraction:* This part carries the main emphasis of our paper, deals with pre-cleaning the image, extracting the regions, and cleaning these regions. We study four different approaches to extracting regions: sign representation of the image convolution with $\nabla^2 G$, morphological peak/valley detectors [3], morphological image segmentation by watersheds [7], and watershed segmentation of distance functions of binarized regions resulting from segmentation of graylevel images. Although the sign of the convolution with $\nabla^2 G$ offers reasonably effective regions, the operators and algorithms of mathematical morphology for feature extraction and segmentation have the advantage of providing multiscale region features without blurring their boundaries. Thus, our preferance for using morphological feature extraction and segmentation approaches to extract regions is based on the inherent ability of morphological operators to easily relate to shape and hence to provide regions that may correspond to more easily identifiable subparts of the moving object. II) *Region Matching* where Ullman's general correspondence theory [6] is applied to region tokens by using several similarity criteria for matching. These criteria are based on a more extended set of region features than the affinity measure used by Ullman. After the region matching, ve-

locity estimates are then identified as the spatial displacements among centroids of corresponding regions. III) *Velocity Smoothing* where the 2-D velocity data are smoothed with a spatio-temporal vector median filter.

2.1. Region Extraction

2.1.1. Sign Representation of $\nabla^2 G * I$

Regions are the complementary representation of edges. Hence, they can be obtained from edge operators. Specifically, the edge detection operation $\nabla^2 G * I$ is applied to each image frame I. Binary edge information is obtained by the zero-crossing contours of the operator's output. For each image frame, the regions are identified as the connected subsets of the image plane whose boundaries are these edge contours. Thus, the set of image pixels at which this edge signal has a positive sign identifies the collection of *positive regions*, and its set complement yields the *negative regions*. There is a trade-off in selecting a value for the scale parameter σ. For large σ, the regions are large, and their number per frame is small. To achieve dense velocity estimates, small values of σ are preferred. On the other hand, to achieve a matching that is more robust and less susceptible to noise, a larger σ is preferred. In our experiments we implemented the $\nabla^2 G$ as the difference of two Gaussians, one (the excitatory) with $\sigma = 2.25$ and another (the inhibitory) with $\sigma = 0.75$; the size of the convolution kernel was 9×9 pixels.

2.1.2. Binarized Peak/Valley Detection Transformations

If I is the intensity image at some time frame, two morphological operators that can extract its intensity peaks and valleys, respectively, are the opening and closing residuals (known as "top-hat" transformations and due to Meyer):

$$Peak(I) = I - (I \circ B) \geq 0 \quad (1)$$

$$Valley(I) = (I \bullet B) - I \geq 0 \quad (2)$$

where B is a flat convex structuring element. The opening $I \circ B$ smooths I by cutting down its peaks; hence the residual signal $Peak(I)$ contains only the peaks of I. The shape and size of B control the shape and maximum size of of the binary regions of support of these peaks. Similarly for the valleys. In our experiments, we use as structuring element an octagon $S = \{(x, y) : x^2 + y^2 \leq 5\}$ of size 2; i.e., $B = S \oplus S$. Note that the resulting element B has the same size as the truncated impulse response for the $\nabla^2 G * I$ operation used in Section 2.1.1 so that both the linear smoothing via the Gaussian convolution and the morphological smoothing via the opening or closing refer to the same scale.

The value of $Peak(I)$ at a certain pixel location determines the contrast (or the "strength") of the peak at that location. We produce binary *peak regions* by thresholding at level T, i.e., by setting all pixels (x, y) at which $[\text{Peak}(I)](x, y) \geq T$ equal to 1 and 0 elsewhere. It is not a simple task to find an optimum T for general-purpose detection. In the approach we used, all the nonzero values of the peak signal $Peak(I)$ were sorted for each frame and T was selected as the 70% percentile value. (The value 70 was experimentally found to give reasonable results.) Similarly, the binary *valley regions* result from thresholding the valley signal Valley(I) at T. Figures with examples from the above peak/valley region extraction will be shown later.

2.1.3. Watershed Segmentation

Here, we shall make use of one of the most powerful tools provided by mathematical morphology, namely the watershed transformation [7]. It is defined for grayscale images via the notion of a *catchment basin:* let us regard the image under study as a topographic relief and assume it is raining on it. A drop of water falling at a point p flows down along a steepest slope path until it is trapped in a minimum m of the relief. The set $C(m)$ of the pixels such that a drop falling on them eventually reaches m is called catchment basin associated with the minimum m. The set of the boundaries of the different catchment basins of an image constitute its *watersheds.*

In other words, the watershed edges or lines are located on the crest-lines of the image which actually separate two different minima (the watershed elements are always closed edges). The basic idea of watershed segmentation consists therefore in applying this tool to the gradient of the image I to be segmented. Note that by *gradient,* we mean here a morphological gradient of I, i.e., an image where the gray-level of each pixel is indicative of the slope in the original image. One of the most popular gradients, often referred to in literature as Beucher's gradient, is obtained by taking the algebraic difference between an elementary dilation and an elementary erosion of I:

$$grad_B(I) = (I \oplus B) - (I \ominus B), \quad (3)$$

(with B being an elementary square or disc).

The direct application of the watershed transformation to a gradient image usually leads to poor results.

Indeed, even after dramatic filtering of the original image or of its gradient, the latter often exhibits far too many minima, and thus far too many catchment basins. Hence, straightforward watershed segmentation of the gradient mostly leads to oversegmented images.

To get rid of this problem, one of the best solutions consists in making use of markers of the regions to be extracted. By marker of a given region, we mean a connected component of pixels located inside this region. The assumption used here is that it is easier to design robust methods to extract markers than to directly extract the precise contours of the desired regions. The method we used is sometimes called generalized maxima/minima extraction, or dome/basin extraction [7]. For the domes, e.g., the principle is to subtract an arbitrary constant h from the original image I and to perform a grayscale geodesic reconstruction of I from $I - h$ The grayscale reconstruction process can itself be viewed as an iteration of elementary dilations of $I - h$ with the constraint that at each step, the resulting image must be smaller than I for each pixel. The reconstructed image is then subtracted from the original one, thus yielding a grayscale image J of *all* the domes and crest-lines of I. From J, it is then easy to extract a binary picture of the most important domes: it suffices to keep each dome which has at least one pixel with value greater than a given constant h'. Usually, one takes $h' = h/2$. This last operation is realized via binary reconstruction of J thresholded at level 1 from J thresholded at value h'. The dual process can be used to extract the basins and valleys of I. The combined dome and basin form the marker image to modify the image gradient to impose on it as new minima.

2.1.4. Watershed Segmentation Followed by Binary Region Segmentation

One can notice that the obtained regions, though very accurate, sometimes exhibit very strange shapes. As we shall see later, this characteristics may have a bad effect on the matching algorithms, which usually work by using the centroid of the extracted regions. Furthermore, when the matching is done with few regions, the results are not as dense as one would expect, especially in comparison with the results provided by block matching techniques. Therefore, it is interesting at this point to cut the regions obtained after the above watershed segmentation into smaller pieces.

Although several approaches may be considered to achieve this goal, watershed-based methods seem once again to provide the most appropriate answer. We used here a technique which is commonly used for binary segmentation tasks, i.e., to separate binary shapes into their perceptually relevant *components*. Its first step consists in determining the *distance function* of the binary image under study: each pixel belonging to the previously extracted regions is assigned a gray-level corresponding to its distance to the outer boundary of this region. Then, the maxima of such a distance function image are called *ultimate erosion* and mark the centroid of the different components in which the regions will be decomposed. In actuality, in order to avoid getting too many markers, constrained maxima are used again here. Finally, the components are obtained by computing the catchment basins of the negation of the distance function.

Examples of regions resulting from the algorithms described in the previous subsections (including the pre- and post-smoothing) are presented in Figure 1. As Figure 1.b shows (with white areas representing the positive sign regions), the regions from the edge operator convey similar information as the edges. In contrast, the peak/valley regions in Figures 1.c and 1.d (where the peaks and valleys are represented by the white areas) correspond to intensity bright or dark blobs. Figure 1.e shows that watershed segmentation of the original gray-level image based on dome/basin markers yields binary regions that are generally consistent with the concept of image segmentation. Finally, as shown in Figure 1.f, watershed segmentation of distance functions of binary regions resulting from graylevel watershed segmentation yields the densest region fields. (In Figures 1.e,f the region boundaries are overlapped to the original image.)

2.2. Region Matching

Our region matching algorithm is guided by Ullman's general correspondence principles [6], but it also has two differences. First, the tokens Ullman used were 1-D line segments, whereas we use 2-D regions. Second, Ullman used affinity measures for matching, whereas we select the best region matching pair by comparing the similarities of regions based on an extended set of region features. Specifically, let R_i and R_j be two regions with areas $A(R_i)$ and $A(R_j)$, respectively, extracted from two consecutive image frames (at times $t = t_k, t_{k+1}$), and let $\vec{c_i}, \vec{c_j}$ denote their centroids. Then, fixing R_i, a region R_j from the frame at $t = t_{k+1}$ is a possible candidate to match with R_i if it successfully passes the following matching criteria:

1. *Centroid Distance:* For two centroids to match,

their distance should not exceed an upper bound; i.e., both the x- and y-components of the displacement vector $\vec{c_i} - \vec{c_j}$ should not exceed L pixels.

2. *Region Identity*: The sign (positive or negative) of two regions if they resulted from the $\nabla^2 G * I$ approach, or their peak vs. valley (respectively, dome vs. basin) identities if they resulted from the morphological peak/valley (respectively, watershed segmentation) approach must be identical for allowing them to match.

3. *Area Difference*: The area of matching regions should not vary too much; i.e., $|A(R_i) - A(R_j)| < P \cdot A(R_i)$ where $0 < P < 1$.

4. *Intensity Difference:* The average intensities of the two regions should not vary too much; i.e.,

$$\left| \sum_{(x,y)\in R_i} \frac{I(x,y,t_k)}{A(R_i)} - \sum_{(x,y)\in R_j} \frac{I(x,y,t_{k+1})}{A(R_j)} \right| < ID_{max}.$$

Clearly, the fixed numbers L, P, and ID_{max} are control parameters for the correspondence process. Specifically, L controls the range of correspondence. Region R_i may be matched with R_j only if the centroid of R_j lies inside a square window of $(2L+1) \times (2L+1)$ pixels centered at the centroid of R_i, and their $\nabla^2 G * I$ signs or their peak/valley or dome/basin identities are the same. P and ID_{max} determine, respectively, the maximum percentage of area difference and the maximum average intensity difference between two regions above which a match is impossible. The parameters we used for these screening criteria in our experiments are $L = 25$ pixels, $P = 0.3$, and $ID_{max} = 20$.

If there are no regions R_j in the frame at $t = t_{k+1}$ satisfying the matching criteria, then there is no match for the particular region R_i. If more than one candidate regions R_j pass the matching criteria, the one having the smallest mean absolute intensity difference

$$\sum_{(x,y)\in R_i} |I(x,y,t_k) - I(x+d_x, y+d_y, t_{k+1})|$$

is selected, where $(d_x, d_y) = \vec{c_i} - \vec{c_j}$ is the centroid displacement vector.

3. Experiments and Discussion

Figure 2 (a) and (b) show two "toy truck" images with no rotation and the objects have an equal amount of translation to the left, downward, and toward camera. The lower left truck is the closest (170mm away), the lower right truck is at middle (220mm away), and the upper tractor truck is the farthest (360mm away). Figure 2 (c), (d), (e), and (f) show the vector median smoothed velocity field generated by matching the regions extracted by the four algorithms. The smoothed velocity fields clearly show that the closest object (the lower left truck) has the largest displacement and the farthest object (the upper tractor truck) has the least displacement.

Overall, the proposed region-based methods for computing image velocities are simple, efficient, less computationally complex than intensity correlation methods, and (as our experiments on real images indicate) more robust than iterative gradient methods especially for medium or long-range motion.

Finally, in addition to their usefulness for motion tracking, the developed morphological region extraction methods can also serve as efficient systems for robust 2-D feature extraction in a variety of computer vision tasks.

References

[1] C.S. Fuh, P. Maragos, and L. Vincent, "Region-Based Approaches to Visual Motion Correspondence," *Technical Report 91-18, Harvard Robotics Lab.*, Nov. 1991.

[2] J. E.W. Mayhew and J. P. Frisby, "Psychophysical and Computational Studies towards a Theory of Human Stereopsis", *Artificial Intelligence*, 17, pp.349-385, 1981.

[3] F. Meyer, "Contrast Feature Extraction", in *Special Issues of Practical Metallography*, Stuttgart, Germany: Riederer Verlag, GmbH, 1978.

[4] H. K. Nishihara, "Practical real-time imaging stereo matcher", *Optic. Enginr.*, 23(5), pp.536-545, 1984.

[5] V. S. Ramachandran and S. M. Anstis, "The Perception of Apparent Motion", *Scientific American*, pp. 102-109, June 1986.

[6] S. Ullman, *The Interpretation of Visual Motion*, MIT press, 1979.

[7] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13 (6), pp. 583–598, 1991.

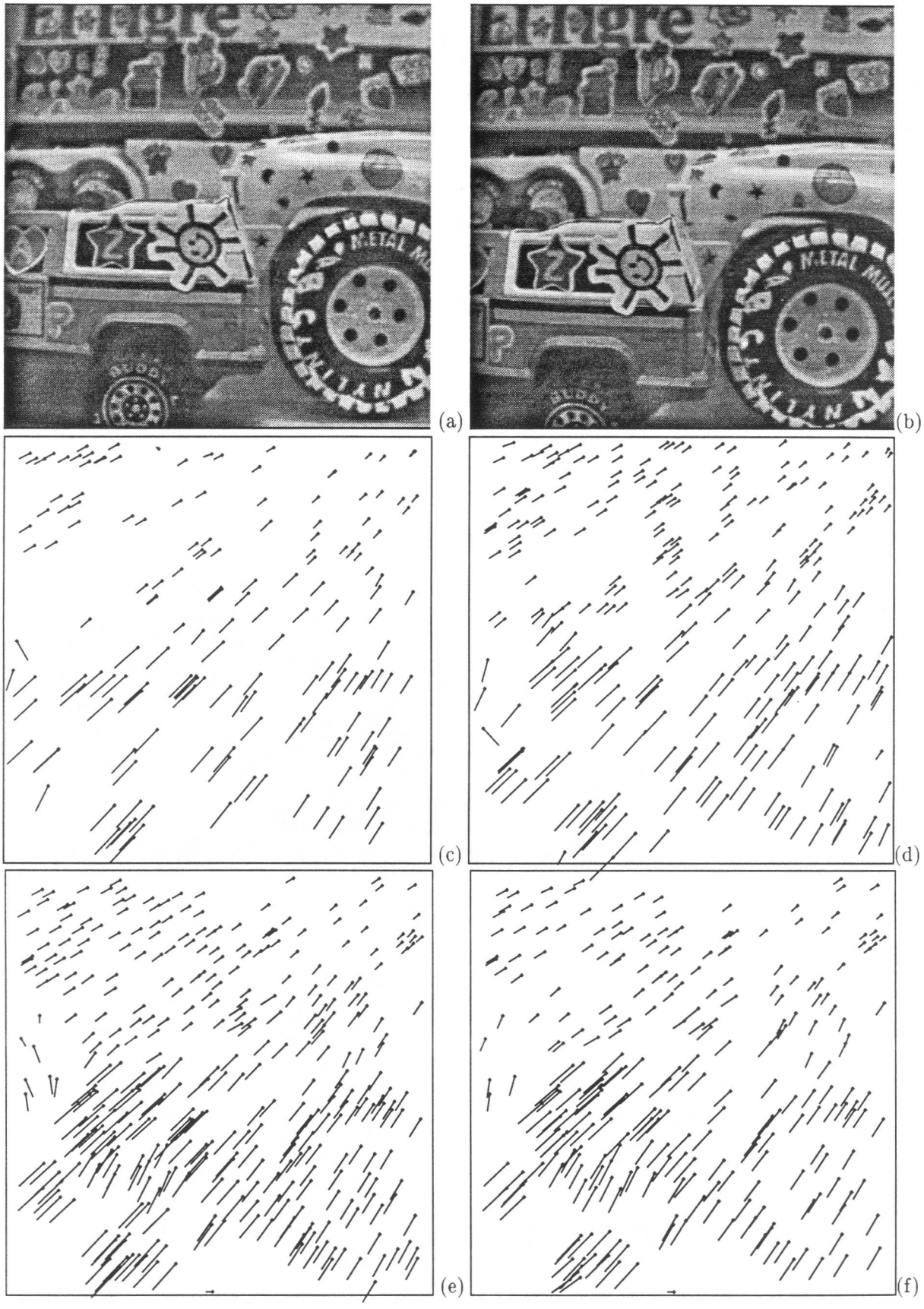

Figure 2: *Toy truck image sequence, (a) Frame 1 (386 × 386 pixels, 8 bit/pixel) (b) Frame 2 of the image sequence. (c), (d), (e), (f) Vector median smoothed velocity fields generated by matching the regions extracted by the four algorithms.*