Perceptual Grouping from Motion Cues Using Tensor Voting

1. Research Team

Project Leader: Prof. Gérard Medioni, *Computer Science*

Graduate Students: Mircea Nicolescu

2. Statement of Project Goals

The project addresses the problem of visual motion analysis by establishing a voting-based computational framework. We present a novel approach for matching and motion capture that recovers the dense velocity field, motion boundaries and regions from a sequence of images, based on a 4-D layered representation of data, and a voting scheme for token affinity communication.

3. Project Role in Support of IMSC Strategic Plan

Our ability to extract velocity information, as well as motion regions and boundaries from image sequences can be very useful to other research projects at IMSC. It can be used by other projects that involve motion capture, avatars, face and body animation, and video compression.

4. Discussion of Methodology Used

We propose a novel approach based on a *layered 4-D representation* of data, and a *voting scheme* for token communication. Our methodology is formulated as a 4-D Tensor Voting computational framework. The position (x,y) and potential velocity (v_x,v_y) of each token are encoded as a 4-D tuple. By propagating this information through voting, distinct moving regions emerge as *smooth surface layers* in the 4-D space of image coordinates and pixel velocities.

4.1. Generating candidate matches

We take as input two image frames that involve general motion – an input image is shown in Figure 1(a). For every pixel in the first image, the goal at this stage is to produce candidate matches in the second image. We use a normalized cross-correlation procedure, where all peaks of correlation are retained as candidates. Each candidate match is represented as a $(x\ y\ v_x\ v_y)$ point in the 4-D space of image coordinates and pixel velocities, with respect to the first image. Since we want to increase the likelihood of including the correct match among the candidates, we repeat this process at multiple scales, by using different correlation window sizes.

The resulting candidates appear as a cloud of $(x\ y\ v_x\ v_y)$ points in the 4-D space. Figure 1(b) shows a 3-D view of the candidate matches - the 3 dimensions shown are x and y (in the horizontal plane), and v_x (the height). The motion layers can be already perceived as their tokens are grouped in smooth surfaces surrounded by noisy matches.

4.2. Extraction of motion layers

Within our 4-D representation, the smoothness constraint is embedded in the concept of *surface saliency* exhibited by the data. By letting the tokens communicate their mutual affinity through voting, noisy matches are eliminated as they receive little support, and distinct regions are extracted as smooth, salient surface layers.

Selection. Since no information is initially known, each potential match is encoded into a 4-D ball tensor. Then each token casts votes by using the corresponding ball voting field. During

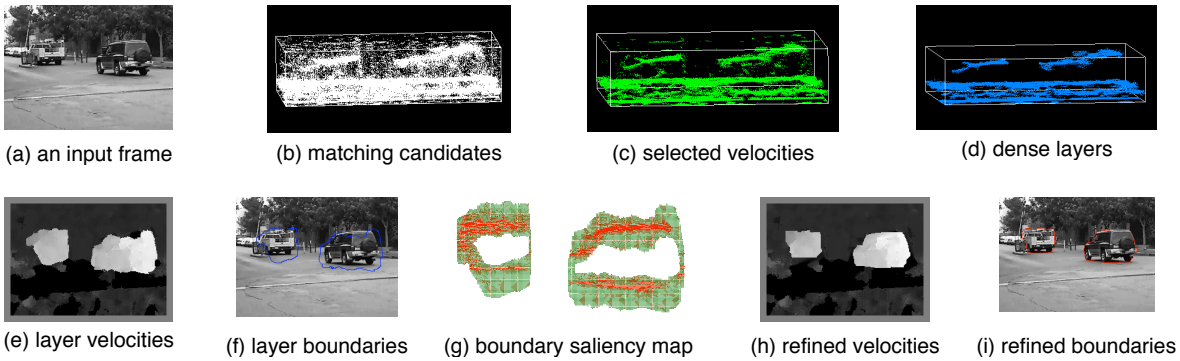

Figure 1. Barrier sequence

voting there is strong support between tokens that lie on a smooth surface (layer), while isolated tokens receive little or no support. For each pixel we retain the candidate match with the highest surface saliency, and we reject the others as outliers. Selected matches are shown in Figure 1(c).

Orientation refinement. In order to obtain an estimation of the layer orientations as accurate as possible, we perform an orientation refinement through another voting process, but now with the selected matches only. After voting, the eigenvectors give the local layer orientations at each token. The remaining outliers are also rejected at this step, based on their low surface saliency.

Densification. Since the previous step created holes (i.e., pixels where no velocity is available), we infer this information from the neighbors by using a smoothness constraint. This is performed through an additional dense voting step, by generating discrete velocity candidates, collecting votes at each such location, and retaining the candidate with maximal surface saliency. By following this procedure at every image location we generate a dense velocity field. A 3-D view of the dense layers (the height represents v_x) is shown in Figure 1(d).

4.3. Boundary inference

After grouping the tokens into regions, based on the smoothness of both velocities and layer orientations, it becomes apparent that the extracted layers may still be over or under-extended along the true object boundaries. Figure 1(e) illustrates the recovered v_x velocities within layers (dark corresponds to low velocity), and Figure 1(f) shows the layer boundaries superimposed on the first input image.

This situation typically occurs in areas subject to occlusion, where the initial correlation procedure may generate wrong matches that are consistent with the correct ones, and therefore could not be rejected as outlier noise. However, the key observation is that one should not only rely on *motion cues* in order to perform motion segmentation. Examining the original images reveals a multitude of *monocular cues*, such as intensity edges, that can aid in identifying the true object boundaries.

The boundaries of the extracted layers give us a good estimate for the position and overall orientation of the true boundaries. We combine this knowledge with monocular cues (intensity edges) from the original images in order to build a boundary saliency map within the uncertainty zone along the layers margins. The smoothness and continuity of the boundary is then enforced through a 2-D voting process, and the true boundary is extracted as the most salient curve within the saliency map. Finally, pixels from the uncertainty zone are reassigned to regions according to the new boundaries, and their velocities are recomputed. Figure 1(g) shows the boundary saliency map, while Figure 1(h) shows the refined velocities within layers, and Figure 1(i) shows the refined motion boundaries, that indeed correspond to the actual objects.

4.4. Results

The examples in Figure 2 and Figure 3 illustrate the performance of our approach for boundary inference in a cluttered environment, when texture edges strongly compete with the true object edges. Through voting, the saliency of the spurious texture edges has been diminished by the overall dominance of saliency and orientation of the correct object edges.

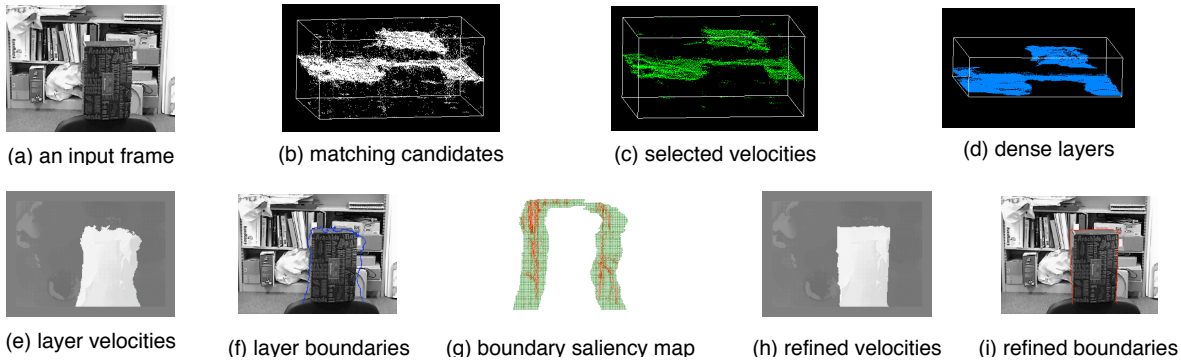

Figure 2. Candy box sequence

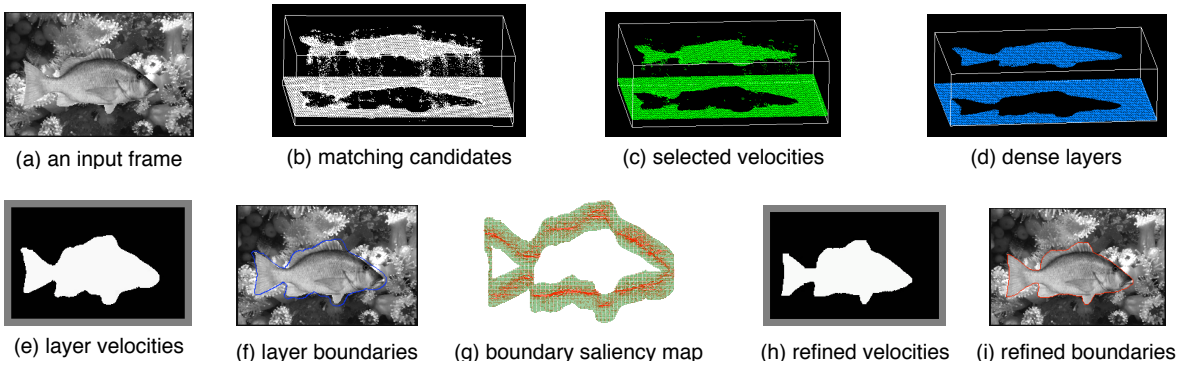

Figure 3. Fish sequence

5. Short Description of Achievements in Previous Years

During the initial phase of this research, the core of the voting-based computational framework had been developed and implemented. This includes finding a suitable representation for the motion layers, as smooth surfaces in 4-D, and establishing an efficient mechanism for token communication.

5a. Detail of Accomplishments During the Past Year

During the past year we developed a consistent methodology for integrating both monocular and motion cues within a tensor voting computational framework. Combining motion information with intensity cues, such as texture edges, is crucial for the accurate segmentation of real image sequences.

Based on this research, we have published a journal paper [1] and two conference papers [2,3] on perceptual grouping from motion cues. One of our papers [2] received the "Best Student Paper Award" at the 2002 International Conference on Pattern Recognition in Quebec City, Canada.

6. Other Relevant Work Being Conducted and How this Project is Different

Optical flow techniques [4] rely on local, raw estimates of the flow field to produce a partition of the image. However, the flow estimates are very poor at motion boundaries and cannot be obtained in uniform areas. Past approaches have also used Markov Random Fields [5] or regularization techniques to handle discontinuities [6].

Significant improvements have been achieved by using layered representations [7]. The difficulties range from a severe restriction in motion representation (as rigid or planar), to over-fitting and instability due to high-order parameterizations.

A computational framework that successfully enforces the smoothness constraint in a unified manner, while preserving smoothness discontinuities is Tensor Voting [8]. The first to propose using Tensor Voting for motion analysis were Gaucher and Medioni [9]. They employ successive steps of voting, first to determine the boundary points as tokens with maximal motion uncertainty, then to locally refine velocities on each side of the boundary. However, their voting communication is essentially a 2-D process that does not inhibit neighboring elements with different velocities from influencing each other.

Our approach proposes a 4-D space for a layered representation, and a voting-based scheme for token communication. This allows for a spatial separation of the points according to both velocities and image coordinates. Consequently, the proposed framework allows tokens from the same layer to strongly support each other, while inhibiting influence from other layers or from isolated tokens. Despite the high dimensionality, our voting scheme is both time and space efficient. It is non-iterative and the only free parameter is scale, which is an inherent characteristic of human vision.

7. Plan for the Next Year

We plan to extend our approach by using polarity voting for boundary extraction, extending the framework for multiple scale analysis, studying the occlusion relationships, and incorporating information from multiple frames.

8. Expected Milestones and Deliverables

We have recently submitted a paper at the IEEE Conference on Computer Vision and Pattern Recognition, and we anticipate more publications in the future, near and long-term, as we are addressing a fundamental problem in Computer Vision.

9. Member Company Benefits

N/A

10. References

[1] M. Nicolescu and G. Medioni, "Layered 4-D Representation and Voting for Grouping from Motion", to appear in the *IEEE Trans. on PAMI - Special Issue on Perceptual Organization in Computer Vision*, 2003.

[2] M. Nicolescu and G. Medioni, "4-D Voting for Matching, Densification and Segmentation into Motion Layers", *ICPR*, 2002.

[3] M. Nicolescu and G. Medioni, "Perceptual Grouping from Motion Cues Using Tensor Voting in 4-D", *ECCV*, vol. III, pp. 423-437, 2002.

[4] J. Barron, D. Fleet, S. Beauchemin, "Performance of Optical Flow Techniques", *IJCV*, 12:1, pp. 43-77, 1994.

[5] F. Heitz, P. Bouthemy, "Multimodal Estimation of Discontinuous Optical Flow Using Markov Random Fields", *PAMI*, 15: 12, pp. 1217-1232, December 1993.

[6] S. Ghosal, "A Fast Scalable Algorithm for Discontinuous Optical Flow Estimation", *PAMI*, 18:2, pp. 181-194, 1996.

[7] S. Hsu, P. Anandan, S. Peleg, "Accurate Computation of Optical Flow by Using Layered Motion Representations", *ICPR*, pp. 743-746, 1994.

[8] G. Medioni, Mi-Suen Lee, Chi-Keung Tang, "A Computational Framework for Segmentation and Grouping", *Elsevier Science*, 2000.

[9] L. Gaucher, G. Medioni, "Accurate Motion Flow Estimation with Discontinuities", *ICCV*, pp. 695-702, 1999.