Motion Saliency Map Generations for Video Data Analysis: Spatio-temporal Signatures in the Array Operations

Jun Hu, Nikos Pitsianis and Xiaobai Sun
Department of Computer Science, Duke University
Durham NC 27708, USA
{junhu, nikos, xiaobai}@cs.duke.edu

Introduction

Motion saliency map is a critical feature in rapid analysis of video data, especially in coping with visual information overload and cluttered background [1]. Its potential applications include automatic action recognition, recognition of moving objects, and target tracking [2, 3, 4]. Similar to the other feature saliency maps, a motion saliency map is a topographic representation of the motion salience at every location in the visual scene. The generation of a saliency map involves typically an intermediate representation of a hierarchical structure with multiple spatial scales, which is computationally intensive. Array data structures are a natural choice in the early processing stage at the fine scale levels, and processors supporting fast array operations in parallel are preferable. The generation of a motion saliency map for video data analysis is more computationally costly in memory space and data access or movement, in comparison to intensity, color, (edge) orientation saliency maps. The latter are generated from individual image frames while a motion saliency map generation involves multiple frames. Moreover, there are different approaches emerged in recent years for generating motion saliency maps, based on physical sensor models, neuron-physiological vision models [5, 6] or information-theoretic arguments [7, 5]. We describe three of the motion saliency map generation methods that are rich in array operations, which can be carried out rapidly on many-core processors such as the graphics processing units (GPUs). These methods differ in their spatial-temporal characteristics or signatures of the array operations. We show a wide range in the data locality and concurrency in each array operation and the data relay between array operations. In the following description and discussion we assume the simplified case that an image frame at time t in a video sequence is a square array $I(x, y; t)$ of N pixels of grayscale values with two dimensional Cartesian coordinates (x, y).

Optical-flow based motion saliency

This approach uses directly the conventional optical-flow model often seen in computer vision applications [8, 9, 10]. The model assumes, among others,

$$\begin{aligned} & I(x+u, y+v, t+1) - I(x, y, t) = 0, \\ \text{or,} \quad & I_x u + I_y v + I_t = 0, \end{aligned} \tag{1}$$

where (u, v), to be determined at each and every pixel (x, y), is the displacement from the frame at t to that at $t+1$, (I_x, I_y) is the spatial gradient, and I_t is determined by the two adjacent frames. The first equation is the grey-value constancy, the second one is its linearization referred to as the optical flow constraint. The flow field in terms of (u, v) over the image domain is then described by a variational model, in which the condition (1) is coupled with some regularization term(s) on smoothness by an energy function f, for example,

$$\min_{(u,v)} f(u, v) = \phi(dI(u, v)) + \lambda\, \phi(\|(\nabla(u, v)\|^2) \tag{2}$$

where $dI(u, v) = I(x+u, y+v, t) - I(x, y)$, ϕ is the 2-norm or a non-negative, differentiable, convex function of the flow field, and ∇ denotes $(\partial_x, \partial_y, \partial_t)$. The temporal gradient component $\partial_t(u, v)$ is absent when (u, v) at $t-1$ is not available. Once the flow field at time t is obtained, one may extract a motion saliency map by the magnitude $|u(x, y)| + |v(x, y)|$. The solution to (2) between every two frames is of iterative nature in general and intensive in array operations. A solution method via the Euler-Lagrange equation, for example, needs at least 8 additional arrays for holding the partial derivatives, I_p, I_{pq}, with $p \in \{x, y, t\}$ and $q \in \{x, y\}$. At every iteration step, there are element-wise addition and multiplication operations between array operands. There may also be certain reduction operations, such as the summation across an entire array for the 2-norm, depending on the detailed structure of the energy function. There are up and down sampling operations when a multi-grid technique is used. The number of iteration steps to reach the equilibrium (or the fixed point), however, depends on many parameters, the condition number of the linearized equation, the initial guess of (u, v) and the termination criteria.

Motion contrast between orientation feature pyramids

In a relatively new approach, one extracts motion feature and motion saliency from other feature representations, in particular, the orientation feature pyramids between two consecutive image frames [1]. The orientation feature pyramid associated with an image frame I contains the feature components $O(\sigma, \theta)$ obtained from I by a bank of Gabor filters, for instance, at multiple spatial scales $\sigma = 0, 1, \cdots, 8$ and multiple local orientations $\theta = 0^o, 45^o, 90^o, 135^o$. Based on a spatio-temporal energy model of motion under suprathreshold conditions, the motion contrast at level σ and in orientation θ is extracted as follows,

$$R(\sigma, \theta) = |O_t(\sigma, \theta) * S_{t-1}(\sigma, \theta) - O_{t-1}(\sigma, \theta) * S_t(\sigma, \theta)|, \tag{3}$$

where $S_t(\sigma, \theta)$ is $O_t(\sigma, \theta)$ shifted by one-pixel along the direction orthogonal to the feature one. The extraction treats the two adjacent frames symmetrically. The motion contrast pyramid is then transformed by the following *center-surround* operations, based on a model for visual receptive fields,

$$\Re(c, s, \theta) = |R(c, \theta) \ominus R(s, \theta)|, \tag{4}$$

where $c = 2, 3, 4$ and $s = c + 3, 4$ indicate the center and surround scales, respectively, and $\ominus$ denotes the cross-scale difference associated with an interpolation scheme. This transformed representation is sensitive to local motion contrast instead of raw feature magnitude, unlike the optical-flow based extraction. Finally, each and every component array $\Re(c, s, \theta)$ is treated further by a pixel-wise normalization procedure, and the motion saliency map is obtained by combining/summing up the normalized components across all dimensions $c = 2, 3, 4$, $s = c+3, 4$ and $\theta = 0^o, 45^o, 90^o, 135^o$. The two processing stages have different spatio-temporal signatures. The Gabor filtering process to obtain the orientation feature pyramid for each frame is a data driven, feed-forward process. The multiple filtering processes can be carried independently, the spatial support of each filter affects the data locality. The stage for extracting the motion contrast from the pyramids of two adjacent frames involves iterative processes in normalization of scale-wise contrasts $\Re(c, s, \theta)$. These iterative processes are independent of each other and can be carried out concurrently.

Spatiotemporal filtering

Spatiotemporal filtering methods treat a sequence of image frames as 3D data, more than two frames are used to extract motion feature, but they differ in the way the spatial and temporal dimensions are coupled in the filters. As a direct extension from the spatial filtering for static images, one may use 3D Gabor filters [11]. We describe a different method. In [12], 2D Gabor filters with two local motion directions, right and left $(45^o, 135^o)$, are used to the (x, t) slices/planes, rendering energy maps $E_{y,R}(x, t)$ and $E_{y,L}(x, t)$ by the center-surround operations and cross-scale combinations. The horizontal motion energy map is then described as follows,

$$E_h(x, y) = \bigcup_{k=1}^{K} \frac{|E_{y,R}(x, t_k) - E_{y,L}(x, t_k)|}{|E_{y,R}(x, t_k) + E_{y,L}(x, t_k)| + \varepsilon} \qquad (5)$$

where ε is a positive constant. Similarly, 2D Gabor filters are used to the (y, t) slices/planes and render the vertical motion energy map $E_v(x, y)$. Finally, a motion saliency map is extracted from (E_h, E_v), for example, by the magnitude of the vector field. Although it seems a straightforward extension of the spatial filtering methods, this method takes many more image frames and hence incurs a delay in the generation of the motion saliency map, in addition to the increase in memory buffers and the scope of memory accesses. The design of 2D or 3D spatial-temporal filers with multiple scales in space and time is in active research.

Discussion

The processing latency is an important factor in evaluation and deployment of a saliency analysis model and an algorithmic interpretation of the model. Along with the spatiotemporal features of the three methods described above, we demonstrate also the resulting saliency maps generated by these methods for the same video segment as the critical evidence and reference. There still remains a gap in processing latency between what is achievable on GPUs with the above methods and what is desirable. There are many active efforts in developing saliency models and algorithms, in two basic approaches or their combinations. One is to extend existing motion detection models [13, 14, 15] , beside the optical flow model, for motion saliency analysis. The other is to exploit the new knowledge of the vision system in the primate brain [6].

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, Nov 1998.

[2] L. Itti, "Computational cognitive neuroscience and its applications," *National Academy of Engineering, 2007 U.S. Frontiers of Engineering Symposium, Redmond, WA*, Sep 2007.

[3] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, (San Siego, CA), pp. 631–637, June 2005.

[4] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology* (B. Bosacchi, D. B. Fogel, and J. C. Bezdek, eds.), vol. 5200, (Bellingham, WA), pp. 64–78, SPIE Press, Aug 2003.

[5] N. Bruce and J. Tsotsos, "Spatiotemporal saliency: Towards a hierarchical representation of visual saliency," in *International Workshop on Attention in Cognitive Systems*, (Berlin, Heidelberg), pp. 98–111, Springer-Verlag, 2008.

[6] J. Tsotsos, Y. Liu, J. Martinez-Trujillo, M. Pomplun, E. Simine, and K. Zhou, "Attending to visual motion," *Comput. Vis. Image Underst.*, vol. 100, no. 1-2, pp. 3–40, 2005.

[7] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18* (Y. Weiss, B. Schölkopf, and J. Platt, eds.), pp. 155–162, Cambridge, MA: MIT Press, 2006.

[8] B. Horn, *Robot Vision*, pp. 279–298. The MIT Press, 1986.

[9] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conference on Computer Vision (ECCV)* (T. Pajdla and J. Matas, eds.), vol. 3024 of *LNCS*, pp. 25–36, Springer, May 2004.

[10] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proceedings of the Eleventh IEEE International Conference on Computer Vision*, pp. 1–8, 2007.

[11] D. Gabor, "Theory of communication," *IEE*, vol. 93, pp. 429–457, 1946.

[12] A. Belardinelli, F. Pirri, and A. Carbone, "Motion saliency maps from spatiotemporal filtering," *International Workshop on Attention in Cognitive Systems*, pp. 112–123, 2009.

[13] J. Rehg and A. Witkin, "Visual tracking with deformation models," in *IEEE International Conference on Robotics and Automation*, pp. 844–850, April 1991.

[14] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *7th International Joint Conference on Artifical Intelligence*, 1981.

[15] J. Shi and C. Tomasi, "Good features to track," in *1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pp. 593 – 600, 1994.