Improved Spatio-temporal Salient Feature Detection for Action Recognition

Amir H. Shabani[1,2]
hshabani@uwaterloo.ca
David A. Clausi[1]
dclausi@uwaterloo.ca
John S. Zelek[2]
jzelek@uwaterloo.ca

[1] Vision and Image Processing Lab.
[2] Intelligent Systems Lab.
University of Waterloo, Waterloo, ON, Canada, N2L 3G1

Spatio-temporal salient features localize the local motion events and are used to represent video sequences for many computer vision tasks such as action recognition. The robust detection of these features under geometric variations such as affine transformation and view/scale changes is however an open problem. Existing methods use the same filter for both time and space and hence, perform an isotropic temporal filtering.

The contribution of this paper is on improving the temporal filtering for robust salient feature detection (Fig. 1) and consequently, improving the action recognition accuracy. We thus develop a novel biologically-consistent multi-resolution filtering which is asymmetric and causal in the temporal direction. This development is based on an anisotropic diffusion to the past frames. Given the functional nature of this filter, we refer to it as the "**asymmetric sinc**". We also introduce the use of **Poisson** and **truncated exponential** filters. We then compare the complex form of these three asymmetric filters with the standard symmetric **Gabor** filtering (Fig. 2). The performance of these filters are evaluated by answering three research questions: (1) whether asymmetric temporal filtering is better than symmetric Gabor filtering for salient motion feature detection, (2) which filter is the best, and (3) the features detected by which filter provide higher action classification accuracy.

We use a common framework (Fig. 1) for spatio-temporal salient feature detection to investigate the role of differen temporal multi-resolution filters. To be consistent with the biological vision, we compute a **phase and contrast polarity insensitive** [2] motion map. We preformed three experimental tests of precision, reproducibility, and action classification to evaluate the salient features. The precision score determines the ratio of the true positive features over all detected features. The reproducibility score determines the percentage of re-detected salient features under different geometric deformations such as scale/view change or affine transformation.

To provide quantitative results for the precision and reproducibility tests, we used the Weizmann classification and robustness datasets for which the ground truth (i.e., foreground masks) is provided. For the action classification test, we used the KTH and the UCF sports datasets as well. A discriminative bottom-up approach based on bag-of-words (BOW) representation of the video content is used for the action classification.

Our observation (Fig. 3) is that the asymmetric temporal filters perform better than the symmetric Gabor filtering in detecting precise, robust, and informative salient features. Among all, our asymmetric sinc filtering performs the best. Without video segmentation, our 95.5% accuracy on the Weizmann dataset and 93.3% accuracy on the KTH dataset competes with the state-of-the-art methods which use similar discriminative BOW setting. Moreover, our 91.5% accuracy on the UCF sports dataset is much better than the dense trajectory [4] with 88.2%, the unsupervised feature learning [1] with 86.5%, and the dense sampling [3] with 85.6% accuracy. This is in contrast to the previous observations about better performance of dense sampling/trajectories, showing the importance of proper temporal filtering for robust and informative salient feature detection.

[1] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *CVPR*, 2011.

[2] A. H. Shabani, J.S. Zelek, and D.A. Clausi. Human action recognition using salient opponent-based motion features. *CRV*, 2010.

[3] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *BMVC*, 2009.

[4] H. Wang, A. Klaser, C. Schmid, and C.L. Liu. Action recognition by dense trajectories. *CVPR*, 2011.

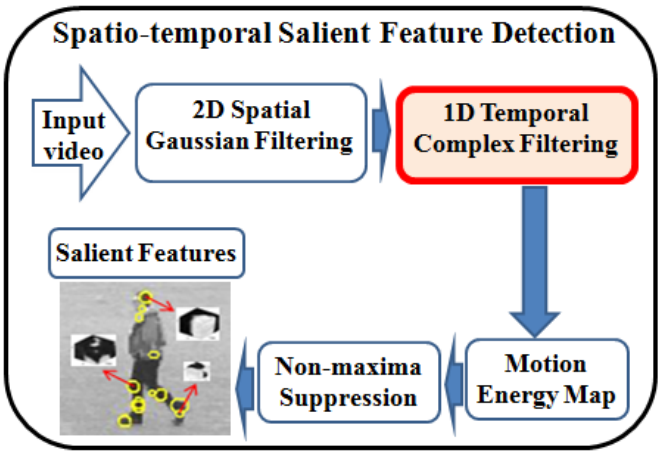

Figure 1: The focus of this paper is on the temporal complex filtering for better salient feature detection and action classification.

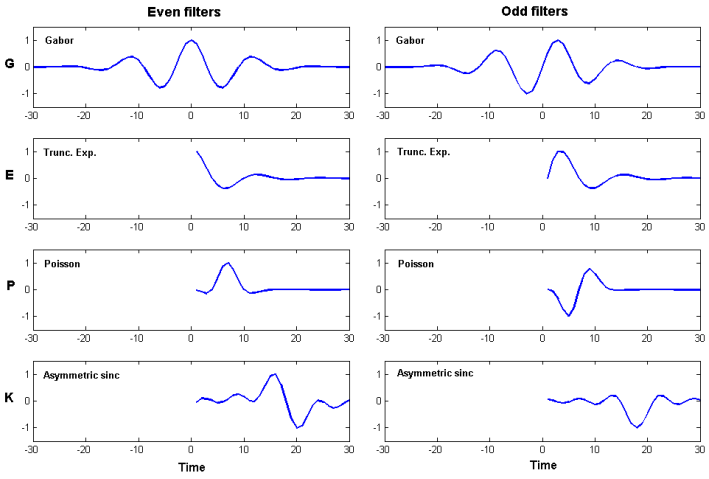

Figure 2: Even and odd components of four different complex temporal filters for multi-resolution video filtering.

		Temporal filters			
Test	Type	Gabor [5]	**Trunc. exp.**	**Poisson**	**Asym. sinc**
Precision	classification data	64.8 %	78 %	74.7 %	**79.6** %
	robustness data	91.9 %	94.7 %	93.6 %	**96.1** %
Reproducibility	view change	55.5 %	71.3 %	73%	**73.4** %
	rotation change	86.6 %	88 %	87.8 %	**93.3** %
	shearing change	84.1 %	88.4 %	86.8 %	**91.8** %
	scale change	**50.7** %	47.5 %	47.1 %	45.2 %
Classification	Weizmann	91.7 %	93.1%	94.6 %	**95.5** %
	KTH	89.5%	90.8%	92.4%	**93.3** %
	UCF sports	73.3%	76%	82.5%	**91.5** %

Figure 3: Asymmetric temporal filtering provide more precise and more robust features by which higher classification accuracy is obtained.