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

Abstract

Spatio-temporal salient features can 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. A novel anisotropic temporal filter for better spatio-temporal feature detection is developed. The effect of symmetry and causality of the video filtering is investigated. Based on the positive results of precision and reproducibility tests, we propose the use of temporally asymmetric filtering for robust motion feature detection and action recognition.

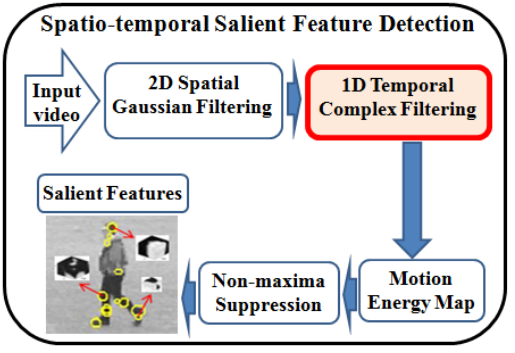
1 Introduction

Motion event detection is important in video analytic applications such as human action recognition, video summarization, and video retrieval. In action recognition, for example, the statistics of the motion events in video can be utilized to discriminate different motion patterns. The global motion maps such as motion history image/volumes [4, 36] or histogram of optical flow [3] are sensitive to occlusion or clutter. Local motions are more robust to these situations and their detection does not necessarily require a video segmentation [16].

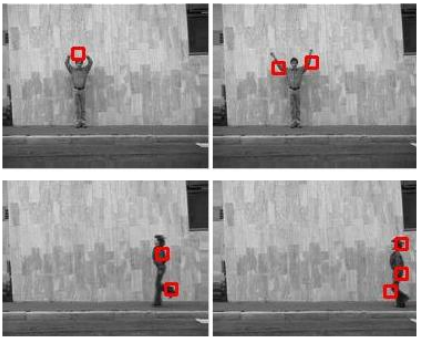
Spatio-temporal salient features can localize the local motion events in space and time. A salient feature is defined by its center point referred to as spatio-temporal interest/key point and is described using the salient appearance and motion features in its extension. Spatial and temporal scale events require the detection of salient features at multiple scales for which a multi-resolution spatio-temporal representation of the video is required. The convolution of a video with linearly separable spatial and temporal kernels provides such a representation [5, 16]. The local maxima in a saliency map of this representation is the location where the local motion events occur (Fig. 1(a)).

Both Gaussian and Gabor filters are, when implemented spatially, multi-resolution filters that are consistent with the functionality of the receptive fields in the human's vision system [18, 34]. For salient feature detection, the existing methods [5, 16, 37] treat the temporal

BMVC 2011 http://dx.doi.org/10.5244/C.25.100



(a) Detection of salient features in video



(b) 2D projection of volumetric salient features

Figure 1: (a) The process of detecting salient features to localize the local motion events is depicted. The focus of this paper is on the temporal complex filtering for better salient feature detection. (b) 2D projection of the volume of salient features at scale $\sigma = 2$ and $\tau = 4$ detected using our novel asymmetric filtering. First and second row show the relevant (i.e., true positive) salient features for two-hand waving and running action, respectively.

domain as a third dimension of space and apply these symmetric non-causal spatial filters in the temporal direction. The V1 cells physiology [1, 31, 34], however, motivates the asymmetry and time causality. In scale-space theory, Lindeberg [18] considered the truncated exponential as the candidate kernel for convolution with a time-causal signal. By cascading a series of truncated exponential kernels, they then derived the Poisson kernel as the canonical time-causal kernel for a discrete signal. In practice, as the delay of using non-causal filtering is negligible, the causal and asymmetric filters are neglected. However, the choice of the preferred temporal filter for improved salient feature detection is unknown.

The contribution of this paper is on improving the temporal filtering for robust salient feature detection 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 these three asymmetric filters with the standard symmetric Gabor filtering to validate the performance of these filters by answering three research questions: (1) whether asymmetric temporal filtering is better than symmetric Gabor filtering for salient feature detection, (2) which filter is the best, and (3) the features detected by which filter provide higher action classification accuracy.

Detected using a common framework (Fig. 1(a)), the features of asymmetric filters are more precise and robust under geometric deformations than the features of the symmetric Gabor filter. Moreover, these features provide higher classification accuracy than the features detected using the symmetric Gabor filter, in a common discriminative bag-of-words (BOW) framework [5, 32]. Overall, our novel asymmetric sinc provides robust features with the highest classification performance.

The rest of this paper is organized as follows. Section 2 reviews the existing methods for salient feature detection for localizing the motion events in video. Section 3 describes the development of the novel asymmetric sinc filter. Section 4 explains the salient feature

detection from the energy map of the filters. Section 5 presents the evaluation tests and the experimental results. Finally, Section 6 summarizes the results.

2 Literature review

Detection of spatio-temporal salient features at multiple scales consists of three main steps: (1) scale-space representation of the video, (2) saliency map construction, and (3) non-maxima suppression to localize the salient features. To this end, existing methods treat the sequence of images as a 3D volume. A Gaussian filtering is thus performed spatially, but different filters have been used for the time direction. For example, Laptev et al. [16] used Gaussian filtering and extended the Harris corners saliency map to detect start and stop of video events. Willems et al. [37] used an approximation of the Gaussian filter with the trace of a Hessian to detect ellipsoids. Dollar et al. [5] used temporal Gabor filtering to detect the "cuboids" from the energy of the motion map [32].

Treating a 2D+t video signal as a 3D spatial volume for motion analysis raises a fundamental question about its validity as the time and space are different in nature. 3D spatial filtering is performed by three orthogonal filters. For motion analysis in a video, however, the orthogonality of the temporal filter with the spatial filters is not required [30], but rather a quadrature pair of temporal filters is required for phase insensitivity [1, 34]. Moreover, it is well-known that the Gaussian diffusion filtering results in blurring and dislocation of semantically meaningful structures such as edges and corners in a still image [35]. In the temporal domain, Gaussian filtering dislocates the salient motions and smoothes both the motion boundaries and the regions with fast motions [28]. Consequently, we argue that the temporal dislocation of salient motions might result in false positive detection.

According to scale-space requirements for real-time multi-resolution video representation [6, 7, 8, 15, 18, 31], the diffusion in the time direction should be limited to just the past frames to respect the time causality observed in the biological vision. For a discrete signal, Lindeberg et al. [18] obtained a Poisson kernel as a canonical time-causal scale-space kernel by cascading a set of truncated exponential kernels. In biological vision, Fourtes [9] modeled the functionality of an axon in the Limulus visual system using a Poisson kernel. This kernel has been utilized to fit with the data representing the temporal contrast sensitivity of the human visual system [1, 34]. This filtering is utilized for a biologically-inspired filtering by Jhuang et al. [14]. Shabani et al. [27] showed the promising performance of using this filtering in extraction of opponent-based motion features and robust local motion event detection in [29]

We model the scale-space filtering of a 2D+t video signal using circuit theory [10, 12, 23], but considering the time causality. For our development, a graphical network of connected pixels in video is constructed. The diffusion in any direction can then be steered by the weight on the edge which is the intensity similarity (i.e., conductivity) between the pixels. We obtain an isotropic spatial Gaussian and an anisotropic temporal asymmetric sinc as the spatio-temporal kernels for a linear multi-resolution motion map of a video signal. We then show that the truncated exponential kernel of the Lindeberg [18] is actually an approximation to our temporal kernel. We therefore evaluate all of these kernels to find out the best choice for localizing the motion event by detection of salient features. Note that the anisotropic temporal filtering has shown promising performance for video segmentation [33]. We thus expect to see better performance of the causal and asymmetric filters for motion detection. Compared to the Poisson, truncated exponential, and Gabor filtering,

our asymmetric sinc filter provides better feature detection and consequently, better action recognition (Section 5).

3 Time-causal multi-resolution video filtering

In this section, we introduce a circuit representation to derive a time-causal multi-resolution representation of a video signal which is then utilized for salient feature detection in the next section. The circuit representation is used in modeling the neurobiological systems [9, 10] and modeling the random walk process [12].

A digital image can be viewed as a graph network in which each pixel x is a node connected to the neighbor pixel by a resistance R. The brightness u at each pixel is the charge on a capacitor C. The flux of the brightness between each two adjacent pixels depends on their relative brightness. We extend this model to a 2D+t video signal by considering the effect of the potential of the corresponding pixel from the immediate past frame on the potential evolution at the current frame. This consideration satisfies the time causality and is modeled as an external conductance R_{ext}^{-1}. For simplicity, we show this modeling for a $1D+t$ signal in Fig. 2(a) and derive the equations for this circuit. From Kirchhoff's Laws (1-4), we can derive the diffusion equation as the change of the brightness at a given (time-like) diffusion scale s. The potential at this pixel is denoted by $u(x,t;s)$ and the current by $i(x,t;s)$.

$$
\begin{aligned}
i(x,t;s) &= (u(x,t;s)-u(x-dx,t;s))/R_1 && (1)\\
i(x+dx,t;s) &= (u(x,t;s)-u(x+dx,t;s))/R_2 && (2)\\
i_{ext} &= (u(x,t;s)-u(x,t-dt;s))/R_{ext} && (3)\\
-C\frac{\partial u(x,t;s)}{\partial s} &= i(x,t;s)+i(x+dx,t;s)+I_{ext} && (4)
\end{aligned}
$$

Consider $R_1 = R_2 = R$ and the per unit quantities, $R = rdx$, $R_{ext} = r_{ext}dt$, $C = cdx$, and $I_{ext}(x,t;s) = i_{ext}(x,t;s)dx$. Substitute equations 1, 2 and 3 in the KCL equation 4. At the limit $dt \longrightarrow 0$ and $dx \longrightarrow 0$, the KCL equation results in the spatio-temporal diffusion equation 5.

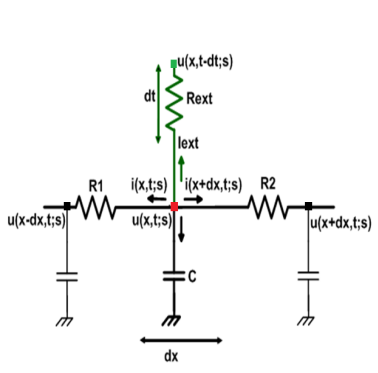
$$\frac{\partial u(x,t;s)}{\partial s} = \frac{1}{rc}\frac{\partial^2 u(x,t;s)}{\partial x^2} - \frac{1}{r_{ext}c}\frac{\partial u(x,t;s)}{\partial t} \tag{5}$$

This equation is consistent with the observation of Lindeberg [13] in that the variation of signal with respect to scale ($\frac{\partial u}{\partial s}$) should be proportional to its second-order derivative in space ($\frac{\partial^2 u}{\partial x^2}$), but first-order derivative in time ($\frac{\partial u}{\partial t}$). For a 2D spatial network of pixels, (x,y), the second-order spatial derivative $\frac{\partial^2}{\partial x^2}$ is replaced with a 2D Laplacian $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$.

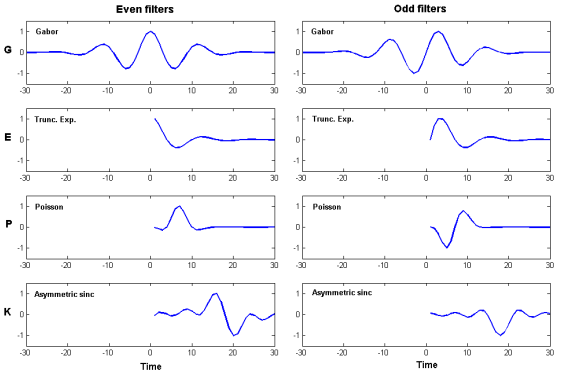
The solution of the diffusion equation 5 is obtained using the Fourier transform. Consider $\widetilde{U}(w_x,w_y,w_t;s)$ as the (discrete-time) Fourier transform of $u(x,y,t;s)$ with respect to pixel coordinates $X = (x,y,t)$, in which (w_x,w_y,w_t) are the corresponding spatial frequencies and temporal frequency. $\alpha = 1/rc$ and $\beta = 1/r_{ext}c$ are related to the image resolution and frame rate. We consider constant value α and β to obtain a close form solution.

1. Apply a Fourier transform to (5) and solve the resulting ordinary differential equation.

$$\frac{\partial \widetilde{U}}{\partial s} = [-\alpha(w_x^2+w_y^2) - \beta(jw_t)]\widetilde{U} \Longrightarrow \widetilde{U} = [e^{-\alpha[w_x^2+w_y^2]s}e^{-j\beta w_t s}]\widetilde{U}_0 \tag{6}$$



(a) Modeling time-causal 1D+t filtering



(b) Even and odd components of the temporal filters

Figure 2: (a) The brightness $u(x,t;s)$ is the charge on the capacitor C at node x at the current time t and current diffusion scale s. A pixel is connected to the neighbor spatial pixels by a resistor R and to temporal past neighbor pixel by R_{ext}. (b) Even and odd components of four different complex temporal filters introduced in Section 3.

2. The inverse Fourier transform of (6) provides the convolution ($\star$) of the original image sequence u_0 with the (discrete analogue of) spatial Gaussian kernel G with standard deviation $\sigma = \sqrt{2\alpha s}$ and a time-causal asymmetric sinc $k(t;\tau)$ with temporal scale $\tau = \beta s$. The $S(t)$ denotes the Heaviside step function.

$$u(x,y,t;\sigma,\tau) = G \star k \star u_0 \ , \quad G(x,y;\sigma) = \frac{e^{-\frac{x^2+y^2}{2\sigma^2}}}{2\pi\sigma^2} \ , \quad k(t;\tau) = sinc(t-\tau)\, S(t) \qquad (7)$$

Note that by releasing the causality constraint, the diffusion spreads to the future frames as well. The resulting temporal kernel will be a Gaussian due to appearance of the second-order temporal derivative in (5).

By considering the stability criteria of $|w_t \tau| < 1$, the Taylor approximation of the temporal term in (6) equals to $e^{-jw_t\tau} \sim 1/(1+jw_t\tau)$. The temporal kernel can thus be approximated with a truncated exponential function ($E(t,\tau) = e^{-t/\tau} S(t)/\tau$) for a continuous signal. The stability constraint shows that for small time scales, higher temporal frequencies are captured and vice-versa. As explained in Section 2, the truncated exponential is the base for the derivation of the Poisson kernel $P(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$ (n denotes the (positive) discrete time and $\lambda = \tau$). For the sake of completeness, we also consider both the truncated exponential and Poisson for the evaluations. This consideration helps us to better understand the effect of both symmetry and causality in the salient feature detection.

4 Salient feature detection

Local motion events occur at different spatial and temporal scales and hence, the salient features should be detected at multiple spatio-temporal scales. These features are detected from a motion saliency map constructed from the multi-resolution representation of the video signal. From biological vision [1, 27, 34], a motion map should be phase insensitive to provide a response independent of the phase of the motion stimulus and be constant for a

periodic motion. Moreover, the motion response should be the same for two stimuli with opposite contrast but similar motion. The latter requirement ensures a similar response for a white foreground against dark background and vice versa.

To obtain a **phase insensitive** motion map, we need two orthogonal kernels to capture all phase information. We therefore construct a quadrature pair $k_h(t;\tau) = k(t;\tau) \star h(t)$ of the asymmetric sinc using the Hilbert transform $h(t) = \frac{1}{\pi t}$ [34].

To obtain a **contrast polarity insensitive** response, the energy of the filters $R = (G \star k \star u_0)^2 + (G \star k_h \star u_0)^2$ is utilized. The local maxima in the motion energy map localizes the salient features [5, 27] (Fig. 1(a)).

Note that the Gabor filter is complex due to multiplying the complex exponential term $e^{-jw_0 t}$ with the Gaussian kernel. We can therefore use the energy of the filter outputs for the motion detection which provides us the "cuboids" [5]. However, to address the phase and contrast polarity insensitivity of the motion map for the truncated exponential and Poisson kernel, we make them complex by multiplying the same complex exponential term as the Gabor (Fig. 2(b)). Multiplying this complex term, in fact, shifts the magnitude of the Fourier transform of the kernel to the specific frequency of interest $w_0 = 1/2\tau$.

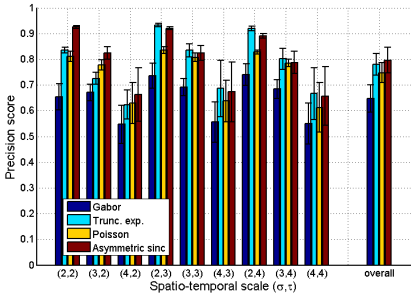
5 Experimental tests

This section explains the experiments utilized to compare the performance of a Gabor filter with three asymmetric filters (Poisson, truncated exponential, and asymmetric sinc) in a common framework (Fig. 1(a)) for the task of salient feature detection. The features are detected at nine different spatio-temporal scales ($\sigma_i = \tau_i = 2\,(\sqrt{2})^i, i = \{0,1,2\}$) in which σ is in pixels and τ is in frames. Fig. 1(b) shows sample features detected using asymmetric sinc at spatio-temporal scale of $\sigma = 2$ and $\tau = 4$. Three different scenarios are evaluated: precision, reproducibility/repeatibility, and action classification.

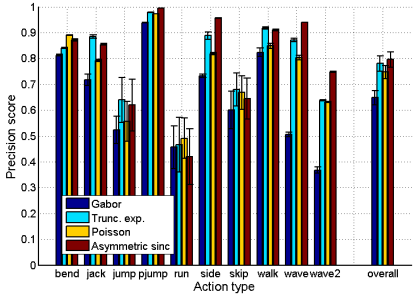
To provide quantitative results for the precision and reproducibility tests, we need video samples with ground truth. More specifically, we need segmented videos of the same action/subject taken under different views and clutter/occlusion. For these tests, we used the robustness and classification datasets of the Weizmann dataset [2] for which the ground truth has been provided. For the action classification test, we used the the KTH [25] and the UCF sports [24] datasets as well.

The **Weizmann robustness data set** includes the "deformations" and "view-change" videos. This set has video samples of "walk" captured under different views (0^o-81^o with 9^o intervals) and under motion pattern deformations, occlusions, and clutter.

The **Weizmann classification data set** consists of ten different actions (run, jack, jump, jump-in-place, wave-two-hands, wave-one-hand, side, bend, skip, and walk) each performed by at least nine different people in front of a fixed camera (with a total of 93 videos). The **KTH data set** [25] consists of six actions (running, boxing, walking, jogging, hand waving, and hand clapping) with 600 video samples. The **UCF Sports dataset** [24] includes actions such as diving, golf swing, kicking, lifting, riding horse, run, skate boarding, swing baseball, and walk with 150 video samples collected from Youtube website. This dataset is more challenging due to diverse ranges of views and scene changes with moving camera, clutter, and partial occlusion.



(a) Scores averaged over 93 video samples.



(b) Scores averaged over all scales and all action samples.

Figure 3: Precision scores for the classification video samples. The Gabor filter [5] has the worst overall performance. (a) As the temporal scale increases the precision decreases. (b) There exist more false positives in the videos such as "run" or "jump" with fast motions. Asymmetric sinc is very good at detecting localized motions such as "wave" and "wave2".

5.1 Precision test

We performed the precision test on video samples of both the classification and robustness "deformations" data sets to compare the performance of four temporal filters. The precision score is the ratio of the number of true positive features to all detected features. A feature is true positive if it has sufficient volumetric intersection with the foreground mask.

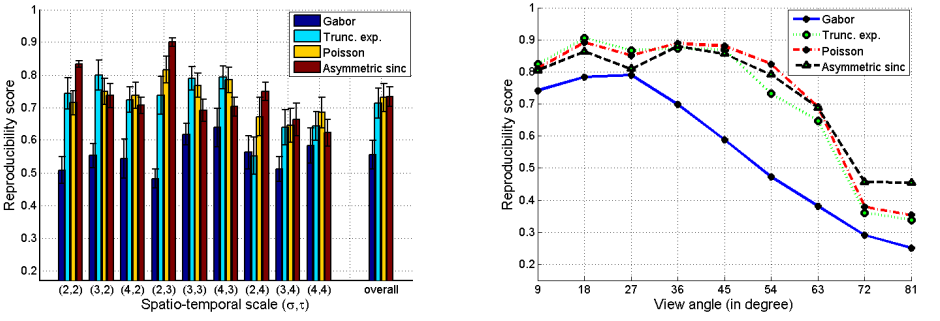
Fig. 3 shows the precision score for different scales and different actions. We report the results for the threshold of at least 25% volumetric intersection; however, the internal testing shows that the results are relatively consistent across different thresholds. Results show that causal filters provide salient features that are more precise than those detected using temporal Gabor filter. Due to linear diffusion, there are two situations in which the uncertainty in temporal correlation will increase: (a) at high temporal scales (e.g.,$\tau = 4$) (Fig. 3(a)) and (b) when we have fast motions (e.g., run) (Fig. 3(b)). Higher temporal uncertainty results in more temporal dislocation and hence, a decrease in the precision.

5.2 Reproducibility tests

In many tracking or recognition applications, the same features should be re-detected if the geometric deformations such as view or scale change occur [21, 37]. The reproducibility score is defined as the ratio of the number of matched features between two videos divided by the total number of detected features. This score is a measure of feature's robustness.

For **view change**, we used the videos of "walking" recorded under different views from the Weizmann robustness data set. Due to the temporal asynchrony of these videos, we used the feature descriptor (3DSIFT [26]) for matching. For in-plane **rotation** test, we created eight videos out of each video from 93 video samples of the Weizmann classification dataset, rotated from -60^o to $+60^o$ with an interval of 15^o. To test the reproducibility under in-plane **shearing**, we used the vertical (horizontal) shearing factor of ($\{0, 0.2, 0.4, 0.6\}$) as part of the shearing transformation. We therefore have 15 shearing transformations for each of 93 video samples (excluding the $(0,0)$). For the spatial **scale change**, we generated three ($\{1/2, 1/3, 1/4\}$) sub-sampled videos from each of the 93 classification videos.

As Fig. 4 shows, the asymmetric filters significantly perform better than the Gabor under



(a) Scores averaged over 93 video samples. (b) Scores averaged over nine scales and ten samples.

Figure 4: Reproducibility score under view change. Note that the Gabor has the worst overall performance. (a) Asymmetric sinc and Poisson are more robust under view change at all spatio-temporal scales. (b) The reproducibility of the Gabor filter drastically decreases after 36^o in view change, while asymmetric filters are more robust.

view change. Due to space limitation to show the plots, we summarize the results of other reproducibility tests in Table 1. Under rotation and shearing, the asymmetric filters perform better than the symmetric Gabor filter. The Gabor filter performs slightly better under spatial scale change because the symmetric structures are better preserved under spatial subsampling. Except from the scale change test, in all other tests, the asymmetric sinc performs the best.

The results of both the precision and reproducibility tests favor asymmetric filters for preferred salient feature detection. This observation is in support of the fact that consistency with the human visual system [34] provide better motion detection. Among introduced filters, our asymmetric sinc filter performs generally better than the Poisson and the truncated exponential. With respect to their peaks, the proposed asymmetric sinc and Poisson are asymmetric and hence, they incorporate a few future frames, but much fewer than past frames. The exponential kernel is the extreme case and considers just the past frames. The Gabor kernel is symmetric and incorporates the same number of frames from past and future. In essence, the asymmetric filters are aware of time, but Gabor is not.

A possible explanation for better performance of asymmetric sinc and Poisson is that they capture wider set of features from asymmetric to symmetric due to the shape change of these kernels with scale (i.e., they are scale-covariant). The truncated exponential favors more the asymmetric features and the Gabor is more responsive to the symmetric features. A higher performance in asymmetric filters for the feature detection promises better action classification.

5.3 Action classification test

For action classification, we incorporated the same bag-of-words (BOW) setting utilized in [5, 27, 32]. Salient features are clustered based on their 3DSIFT descriptors [26] into visual words using K-means algorithm with random seed initialization. We experimentally set the number of clusters to 200 for the Weizmann dataset and 1000 for the KTH and UCF sports datasets, consistent with the experimental setting in [32]. We formed the action signature for a video as the frequency of the occurrences of these clusters in the video. The

Table 1: Summary of the evaluation tests of different temporal filters for salient feature detection and action classification. Bold numbers represent the highest performance. Note that overall our proposed asymmetric sinc filter has the best performance.

		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** %

χ^2 distance metric is then utilized for the matching of the action signatures. For the Weizmann dataset, a nearest neighbor classifier is used with leave-one-out setting. For the KTH and UCF sports datasets, the SVM classifier and the training-testing setting similar to [32] is used. Table 1 shows that the salient features detected using asymmetric and causal filters provide higher average action classification compared to non-causal symmetric Gabor filter. As we expected, our proposed asymmetric sinc filtering performs the best, the Poisson is second best, and the truncated exponential is third best. Note that there are many action recognition schemes that we could have used, some superior to the BOW that we used. However, the focus of this paper is on the quality of the features and the choosing an appropriate action recognition method is left to future work.

5.4 Summary of the results

We introduced three asymmetric filters which provide more precise and more robust salient features than the widely used symmetric Gabor filter. Moreover, we used these features with a standard discriminative BOW framework for action classification. Our novel complex temporal filter of asymmetric sinc has the best overall performance. Even though the features detected using the Gabor filter are more robust under the spatial scale change, the action classification test on the KTH and UCF sports datasets in which there exists camera zooming shows that the features of asymmetric filters perform better than those of Gabor due to better overall precision and robustness.

Without video segmentation and with a similar BOW setting, our 95.5% accuracy on the Weizmann dataset is comparable with the method in [11] with 83.7% and in [26] with 82.6%. Similarly, we obtained 93.3% accuracy on the KTH dataset which is comparable with 84.3% using 3D Hessian features [37], 88.3% accuracy using dense features [23], and 92.1% accuracy using 3D Harris features [32]. Moreover, our 91.5% accuracy on the UCF sports dataset is much better than the accuracy obtained using the dense trajectory [13] with 88.2%, the unsupervised feature learning [17] with 86.5%, and the dense sampling [32] with 85.6%. 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.

6 Conclusion

We developed a novel asymmetric complex video filtering using isotropic diffusion in space and anisotropic causal diffusion in time. We then evaluated the role of causality and symmetry of the video filtering for the spatio-temporal salient feature detection. We observed that asymmetric filters perform better than symmetric Gabor filtering. *As the asymmetric filters are aware of time, we conclude that a taste of the future is good, but the motion history is more important*. In a standard discriminative BOW framework, we also obtained better action classification accuracy using asymmetric sinc filter. We are currently experimenting on more real-world datasets such as HOHA [20] and UCF Youtube actions [19] dataset.

Acknowledgment

GEOIDE (Geomatics for Informed Decision), a Network for Centers of Excellence supported by the NSERC Canada, is thanked for the financial support of this project.

References

[1] E. H. Adelson and J.R. Bergen. Spatio-temporal energy models for the perception of motion. *Optical Society of America*, pages 284–299, 1985.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE International Conference on Computer Vision, Beijing, China*, pages 1395–1402, October 2005.

[3] R. Chaudry, A. Ravichandran, G. Hager, and R. Vidal. Histogram of oriented optical flow and binet-cauchy kernel on nonlinear dynamic systems for the recognition of human actions. *IEEE International Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA*, pages 1932–1939, June 2009.

[4] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. *Computer Vision and Pattern Recognition, San Juan, Puerto Rico*, pages 928–934, June 1997.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal filters. *IEEE International Workshop VS-PETS, Beijing, China*, pages 65–72, August 2005.

[6] R. Duits and B. Burgeth. Scale spaces on lie groups. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 300–312, 2007.

[7] D. Fagerstrom. Temporal scale spaces. *International Journal of Computer Vision*, 64 (2-3):97–106, 2005.

[8] D. Fagerstrom. Spatio-temporal scale-spaces. In *International Conference on Scale Space and Variational Methods in Computer Vision, Ischia, Italy*, pages 326–337, June 2007.

[9] M. G. F. Fourtes and A. L. Hodgkin. Changes in the time scale and sensitivity in the omatidia of limulus. *Journal of Physiology*, 172:239–263, 1964.

[10] W. Gerstner and W.M. Kistler. *Spiking Neuron Models. Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

[11] T. Goodhart, P. Yan, and M. Shah. Action recognition using spatio-temporal regularity based features. *IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegad, USA*, pages 745–748, March 2008.

[12] L. Grady. Random walks for image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1–17, 2006.

[13] C. Schmid C.L. Liu H. Wang, A. Klaser. Action recognition by dense trajectories. *IEEE Conference on Computer Vision and Pattern Recognition, Colorado Spring*, pages 3169–3176, June 2011.

[14] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[15] J. J. Koenderink. Scale-time. In *Biological Cybernetics*, pages 162–169, 1988.

[16] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, pages 207–229, 2007.

[17] 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. *IEEE Conference on Computer Vision and Pattern Recognition, Colorado Spring*, pages 3361–3368, June 2011.

[18] T. Lindeberg and D. Fagerstrom. Scale-space with causal time direction. In *European Conference on Computer Vision, Cambridge, UK*, pages 229–240, April 1996.

[19] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". *IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL*, pages 1–8, June 2009.

[20] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA*, pages 2929–2936, June 2009.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, pages 43–72, 2006.

[22] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 629–639, 1990.

[23] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1454–1461, 2009.

[24] M.D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, Alaska, USA*, pages 1–8, June 2008.

[25] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. *IEEE International Conference on Pattern Recognition, Cambridge, UK*, 3: 32–36, August 2004.

[26] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional *SIFT* descriptor and its application to action recognition. In *ACM Multimedia, Augsburg, Germany*, pages 357–360, September 2007.

[27] A. H. Shabani, J.S. Zelek, and D.A. Clausi. Human action recognition using salient opponent-based motion features. *IEEE Canadian Conference on Computer and Robot Vision, Ottawa, Canada*, pages 362 – 369, May 2010.

[28] A.H. Shabani, D.A. Clausi, and J.S. Zelek. Towards a robust spatio-temporal interest point detection for human action recognition. *IEEE Canadian Conference on Computer and Robot Vision*, pages 237–243, 2009.

[29] A.H. Shabani, J.S. Zelek, and D.A. Clausi. Robust local video event detection for action recognition. *Advances in Neural Information Processing Systems (NIPS), Machine Learning for Assistive Technology Workshop, Whistler, Canada*, December 2010.

[30] E. P. Simoncelli and D. J. Heeger. A model of neuronal responses in visual area mt. *Vision Research*, 38(5):743–761, 1997.

[31] B.M. ter Haar Romeny, L.M.J. Florack, and M. Nielsen. Scale-time kernels and models. In *Scale-Space and Morphology in Computer Vision, Vancouver, Canada*, pages 255–263, July 2001.

[32] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Conference, London, UK*, pages 1–11, September 2009.

[33] J. Wang, B. Thiesson, Y. Xu, and M. Cohen. Image and video segmentation by anisotropic kernel mean shift. *European Conference on Computer Vision, Prague, Czech Republic*, pages 238–249, May 2004.

[34] A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *Journal of Optical Society of America*, 2(2):322–342, 1985.

[35] J. Weickert. A review of nonlinear diffusion filtering. In *International Conference on Scale-space theory in computer vision, Utrecht, Netherlands*, pages 3–28, July 1997.

[36] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 103(2):249–257, 2006.

[37] G. Willems, T. Tuytelaars, , and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *European Conference on Computer Vision, Marseille, France*, pages 650–663, October 2008.