Spatiotemporal Salience via Centre-Surround Comparison of Visual Spacetime Orientations

Andrei Zaharescu and Richard Wildes

York University, Toronto, Canada

Abstract. Early delineation of the most salient portions of a temporal image stream (e.g., a video) could serve to guide subsequent processing to the most important portions of the data at hand. Toward such ends, the present paper documents an algorithm for spatiotemporal salience detection. The algorithm is based on a definition of salient regions as those that differ from their surrounding regions, with the individual regions characterized in terms of 3D, (x, y, t), measurements of visual spacetime orientation. The algorithm has been implemented in software and evaluated empirically on a publically available database for visual salience detection. The results show that the algorithm outperforms a variety of alternative algorithms and even approaches human performance.

1 Introduction

1.1 Motivation

Temporal image streams (e.g., videos) are notorious for the vast amounts of data they comprise. Correspondingly, the efficient processing of such data would benefit greatly from early delineation of the most salient portions of the data, so that subsequent operations can focus on those components. Moreover, automatic detection of salient patterns in video will provide an indication of where humans will likely focus their attention when viewing similar data streams and thereby help in filtering vast video databases to provide information of most interest to humans. As examples, an algorithm for spatiotemporal salience detection could detect an object moving against the dominant direction of motion (e.g., wrong way motion detection), as well as a coherent motion against a more complicated background (e.g., a person moving amidst water waves or camouflaging wind-blown vegetation). Detection of such salient patterns in visual spacetime will be able to cue a wide range of subsequent visual processes (e.g., target tracking, action recognition, video indexing and browsing), without relying on extensive a priori information (e.g., knowledge of targets of interest or background models).

In response to the observations above, this paper presents a spatiotemporal salience detection algorithm. In particular, the algorithm accepts as input a temporal stream of images (e.g., a video) and outputs a corresponding saliency map (e.g., where regions in the video with greater salience are indicated by larger numerical values in the derived map). The approach defines salient regions as those that differ from their surrounding neighborhood in terms of the dynamics of their particular image streams. Significantly, such centre-surround antagonism is a pervasive principle in the organization of biological sensory systems and it is believed to serve the purpose of accentuating salient regions for subsequent processing [1]. To realize this principle in computational terms, two major issues must be considered. First, a base representation must be specified over which the centre-surround comparison is performed. Second, a comparison metric for the centre and surround must be selected. In the present work, the base representation is comprised of pixelwise distributions of visual spacetime, (x, y, t), orientation measurements that serve to capture the local spatiotemporal structure. Such measurements provide an integrated approach to characterizing both the spatial texture and dynamics of a region [2]; correspondingly, centre-surround differences of such distributions can provide the basis for salience detection. For the comparison operation, local measurements are aggregated separately in centre and surround regions and a standard approach to quantifying the difference between two distributions, the Kullback-Leibler divergence [3], is employed.

1.2 Related research

Owing to its potential to guide and optimize subsequent processing, a variety of computational approaches have been developed for spatiotemporal salience detection. Perhaps the most widespread approach is use of learned background models, with salience defined in terms of differences from the acquired model [4]. To date, acquisition and maintenance of reliable background models remains a challenging task and fundamentally entails access to reference training imagery, which is not always available. Other research is limited in applicability by assuming a static camera [5–7], that foreground appearance change is slower than that of background [8] or that background motion compensation is adequate to remove undesired background motion [9, 10]. Still, other approaches rely on accumulation of extended foreground tracks [11, 12]. Formulations of salience detection also have been developed in terms of "Bayesian surprise" [13] and information maximization [14]. Previous work also has made use of centre-surround comparisons for spatiotemporal salience detection, with the comparisons being variously defined over a combination of color, intensity, orientation, flicker, and motion features [15] or an autoregressive, linear dynamical systems (AR-LDS) model of dynamic texture [16, 17]. It also is of interest to note that biological systems appear to be tuned to detecting salient dynamic patterns against their background [18].

In the present work, spatiotemporal oriented energy filters serve in defining the representation of observed dynamics over which centre-surround processing is defined. Previous research has used similar oriented energy filtering for image sequence analysis toward various ends, including optical flow estimation[19–21], analysis of actions/behaviours [22–25], dynamic texture recognition [2] and dynamic scene recognition [27]. Also, alternative approaches to capturing spatiotemporal orientation (e.g., HOG3D [28]) or spatial orientation combined with optical flow might be considered (e.g., HOG/HOF [29]); however, in application to alternative tasks (e.g., action recognition [30] and dynamic scene recognition

[27]) spatiotemporal oriented energy has been shown to outperform the alternatives. In any case, it appears that the present work is the first to use spacetime orientation as the computational basis for spatiotemporal salience detection.

1.3 Contributions

In the light of previous research, the main contributions of the present work are twofold. First, a novel approach for spatiotemporal salience detection is proposed. The approach makes use of centre-surround differences in measurements of visual spacetime oriented structure as the basis for salience detection. While previous research has made use of centre-surround processing for salience detection (see above), it appears that the present work is the first to apply it in conjunction with visual spacetime orientation analysis for spatiotemporal salience detection. Second, the proposed approach is empirically evaluated on a publically available dataset, including quantitative comparison to seven alternative approaches. The results show that the proposed approach yields the best overall performance relative to the alternatives considered.

2 Technical approach

2.1 Overview

The basic principle to be explored for spatiotemporal salience detection is that the salience of a region is a function of how dissimilar the region's spacetime structure is in comparison to its surrounding area: The more dissimilar a region is from its surround, the higher its salience will be rated. Computational realization of this principle entails specification of two matters. First, a base representation for characterizing visual spacetime structure must be defined. Second, an algorithmic approach to quantifying the difference in represented structure of a region and its surround must be given. The next two subsections of this paper describe each of these components in detail.

2.2 Base representation

In the developed approach to salience detection, visual spacetime structure is represented in terms of local distributions of multiscale 3D, (x, y, t), spacetime orientation measurements. These measurements are extracted from input imagery (e.g., a video) via application of a spatiotemporal orientation tuned filter bank. This representation is advantageous as it provides a uniform way to capture both spatial and dynamic properties of imagery: Orientations that lie along (x, y)-planes capture spatial pattern; orientations that extend into the temporal dimension, t, capture dynamic properties. While the basic filtering mechanisms employed here have mostly been documented previously (e.g., [31, 20, 26]) they are reviewed in the remainder of this section for the sake of keeping the paper self-contained. Also, the particular approach to dealing with normalization embodied in (2), which encompasses a noise floor even while yielding properly normalized measurements, appears to be novel.

To extract the orientation measurements, oriented energy filtering is realized in terms of third derivative of 3D Gaussian filters, $G_3(\mathbf{x};\theta,\sigma)$, where θ represents the direction of the filter's axis of symmetry, σ scale and $\mathbf{x}=(x,y,t)$ spacetime coordinates. These particular filters are selected due to their (moderately) broad tuning, which allows for a wide range of orientations to be captured with a relatively small number of filters. Additionally, these filters admit a steerable and separable formulation [31], which leads to efficient computations. The filter responses are rectified (squared) and aggregated over a local support region, to yield the following local oriented energy measure,

$$E(\mathbf{x};\theta,\sigma)=\sum_{\mathbf{x}\in\Omega} \mid G_3(\theta,\sigma)*I(\mathbf{x})\mid^2, \tag{1}$$

where Ω is an aggregation region and care is taken to normalize the filters to ensure that their energy across scale is constant [32]. Notice that while the filtering, (1), is not performed in quadrature, a reasonable measure of energy is achieved owing to the summation over a support region, Ω: Parseval's theorem states that the squared modulus of a signal aggregated over a region is proportional to the squared modulus of its spectrum [33]; thereby, in the present case, (1) provides a measure of image energy along direction θ at scale σ [26].

The resulting oriented energies are confounded with local image intensity contrast that is not indicative of spacetime orientation. This state of affairs makes it impossible to determine whether a high response from a particular filter is indicative of a close match with the underlying orientation structure or is instead a low match that yields a high response due to significant contrast in the overall signal intensity. To arrive at a purer measure of oriented spacetime structure, the energy measures are divided by the sum of the oriented responses at each point,

$$\bar{E}(\mathbf{x};\theta,\sigma)=\frac{E(\mathbf{x};\theta,\sigma)+\epsilon}{\sum\limits_{\tilde{\theta},\tilde{\sigma}\in\mathcal{M}}(E(\mathbf{x};\tilde{\theta},\tilde{\sigma})+\epsilon)}, \tag{2}$$

where $\mathcal{M}=\sigma\times\theta$ denotes the set of multiscale oriented energies, (1), $\tilde{}$ denotes variables of summation and ϵ a constant, set to 1% of the maximum filter response, introduced as both a noise floor and to avoid instabilities at points where the overall energy is small. Notice that by adding ϵ in both the numerator and denominator of the normalization formula, (2), the final result is a set of pointwise defined normalized energies, $\bar{E}(\mathbf{x};\theta,\sigma)$; whereas, adding ϵ in only the denominator does not yield proper normalization, c.f., [26].

In the currently implemented representation, 10 different directions, θ, are made explicit, as they span the space of 3D orientations for the G_3 filters that were used [31]. The particular orientations selected were the normals to the faces of an icosahedron, as they evenly sample the sphere [34], and antipodal directions are identified. Filters of size 9 are used. Three scales, σ, are considered corresponding to factor of $\sqrt{2}$ subsampling between a multiscale pyramid representation of the input images [35]. By construction, these measures are normalized, which allows for a degree of robustness to unimportant variability in observations and makes them amenable to subsequent comparison metrics that are defined over normalized distributions. Further, the representation is simply realized by an alternating series of linear (i.e., separable convolution and pointwise addition) and pointwise non-linear operations (i.e., squaring and division); thus, efficient computations are realized, including real-time implementations [24]. Overall, pointwise distributions of normalized oriented energy measurements, $\hat{E}(\mathbf{x};\theta,\sigma)$, are made available, with the distributions maintained as 10 (orientations) × 3 (scales) = 30 dimensional, pointwise histograms.

2.3 Centre-surround comparison

Given the defined measurements of spacetime structure, (2), it is necessary to aggregate the locally defined $\hat{E}(\mathbf{x};\theta,\sigma)$ over centre and surround regions to allow for subsequent comparisons. The notation used in defining the initial oriented energy filtering, (1), allows for this consideration via appropriate definition of the aggregation region, Ω. At any given point, $\mathbf{x}$, a central spacetime support region, $\mathcal{C}$, is defined, using a radius $r_{\mathcal{C}}$. Similarly, at each point a surround support region, $\mathcal{S}$, is defined using a radius $r_{\mathcal{S}}$, which extends beyond, but excludes, $\mathcal{C}$. To calculate the centre distibution, $\hat{E}_{\mathcal{C}}(\mathbf{x};\theta,\sigma)$, let $\Omega = \mathcal{C}$ in (1). To calculate the surround distibution, $\hat{E}_{\mathcal{S}}(\mathbf{x};\theta,\sigma)$, let $\Omega = \mathcal{S}$ in (1). Importantly, the operation (2) ensures that the final centre, $\hat{E}_{\mathcal{C}}(\mathbf{x};\theta,\sigma)$, and surround, $\hat{E}_{\mathcal{S}}(\mathbf{x};\theta,\sigma)$, measurements are properly normalized distributions, albeit defined over different support regions. An efficient implementation is realized by using integral images [36].

Finally, to compare the centre and surround measurements, a dissimilarity metric is required so that larger values in the resulting computation imply greater salience. A variety of metrics might be considered, e.g., Kullback-Leibler divergence, χ^2 and earth mover's distance as well as L^1 and L^2 norms [37]. In the present approach, Kullback-Leibler divergence is used as the measure of dissimilarity between two distributions as it provides a principled approach based on relative entropy [3]. Further, in preliminary empirical investigation the Kullback-Leibler divergence showed overall most reasonable performance in comparison to a sampling of alternatives. Correspondingly, salience, $\rho(\mathbf{x})$, is define in terms the Kullback-Leibler divergence between centre, $\hat{E}_{\mathcal{C}}(\mathbf{x};\theta,\sigma)$, and surround, $\hat{E}_{\mathcal{S}}(\mathbf{x};\theta,\sigma)$, spacetime structure representations as

$$\rho(\mathbf{x}) = \sum_{\tilde{\theta},\tilde{\sigma}\in\mathcal{M}} \hat{E}_{\mathcal{C}}(\mathbf{x};\theta,\sigma) \log \frac{\hat{E}_{\mathcal{C}}(\mathbf{x};\theta,\sigma)}{\hat{E}_{\mathcal{S}}(\mathbf{x};\theta,\sigma)} \tag{3}$$

Thus, larger values of ρ are taken as indicative of greater spatiotemporal salience.

3 Empirical evaluation

3.1 Dataset and experimental protocol

The performance of the presented approach for spatiotemporal saliency detection has been evaluated on the USC publically available dataset [13]. The dataset consists of 50 video clips, recorded at 640 × 480 spatial resolution and 30 fps temporal resolution, for a total of over 25 minutes of playtime. The dataset includes a wide variety of very challenging scenarios including home videos, television broadcasts of news, sports, talk shows, commercials and video games. Example frames for the 50 clips are shown in Figures 1, 2 and 3.

The dataset was groundtruthed for salience with respect to human fixation patterns. In particular, human eye tracking data was recorded from 8 subjects. Each subject watched a subset of the collection of video clips, so that eye movement traces for 4 distinct subjects were obtained for each clip. Overall, a total of 200 eye movement traces containing 10, 192 saccades were gathered for the 50 video clips. In essence, locations to which subjects preferentially fixate via saccade eye movements in comparison to randomly selected locations in the videos are taken as indicative of salience. A valuable aspect of this groundtruthing methodology is that it avoids the highly subjective nature of approaches that make use of humans manually labeling imagery for salience.

Given an algorithm for salience detection, the experimental protocol for evaluation with respect to the human eye track-based groundtruth is to compare algorithm recovered salience at human fixated points vs. randomly sampled points. In order to ensure proper random sampling, 100 randomized runs are performed for each sequence. To account for image processing border effects, recovered salience near image borders is down weighted (proportionally to the amount of underlying filter support in the salience computation that does not lie within the images). Letting R and S correspond to distributions of recovered salience (maintained as 10 bin histograms) at randomly sampled and human fixated points (resp.), the difference is quantified in terms of the Kullback-Leibler divergence

$$D_{KL}(R, S) = \sum R_i \log \frac{R_i}{S_i}, \tag{4}$$

with subscripts i indexing individual histogram bins. (N.b, while this same measure is used to compare centre-surround regions in the salience detection algorithm, (3), no bias is thereby introduced, as different distributions are being compared.) Thus, larger scores for $D_{KL}(R, S)$ indicate that performance of the salience algorithm is closer to the human-based groundtruth.

A measure of human salience detection performance also has been derived relative to the groundtruthed dataset, which reflects interobserver consistency [13]. In essence, given the 4 observers of any particular video, the fixation pattern of 1 is compared to the remaining 3; the resulting KL divergence performance measure for human observers across the entire dataset is 0.679 ± 0.011, with 0.011 the standard deviation.

Additional information about the USC dataset and experimental protocol is available elsewhere [13].

3.2 Results

Image-based results of the proposed salience detection algorithm are shown in Figures 1, 2, 3 and 4, where video clip names, example input frames and derived salience maps are depicted in correspondence. Each depicted salience map, $\rho(\mathbf{x})$, is derived according to the comparison measure, (3), with larger values depicted as brighter image intensities. The illustrated results were recovered using centre and surround radii of 2 and 48 pixels (resp.), relative to 320×240 spatial image resolution.

Quantitative results of the KL divergence performance metric, (4), are shown in Table 1. The right side of the table shows histogram representations of the salience at random (green) and human fixated (blue) positions (resp.) as recovered by the proposed algorithm. The left side of the table shows the KL divergence and standard deviation for the proposed approach, denoted spatiotemporal oriented energies (**SOE**), as well as seven alternative approaches. The alternative algorithms include the best performing single image measure reported elsewhere, local *Entropy* [13], centre-surround comparisons of temporal *Flicker* [13], *Motion* [13] and a combination of features (colour, 2D spatial orientation, motion and flicker) centre-surround *Saliency* [15], *Outliers* and Bayesian *Surprise* in the same feature combination [13] and Attention by Information Maximization (*AIM*) [14]. It is seen that the proposed approach (SOE) outperforms all the alternative approaches by a significant margin (i.e., beyond the standard deviation separations).

Name	KL Score
Entropy	0.151 ± 0.005
Flicker	0.179 ± 0.005
Motion	0.180 ± 0.005
Outliers	0.204 ± 0.006
Saliency	0.205 ± 0.006
Surprise	0.241 ± 0.006
AIM	0.328 ± 0.009
SOE	$\mathbf{0.624 \pm 0.012}$

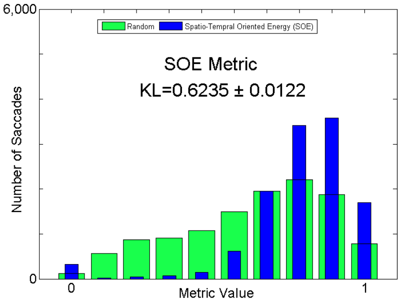

Table 1. *Left*: Comparison of the proposed approach (SOE) with a variety of alternatives. *Right:* Results for the proposed method (SOE): histogram representation comparing saliency values at eye-fixations locations (blue) versus random (green) locations. The KL-divergence score for the proposed algorithm is 0.624, a significant improvement over the previous top performer (AIM, KL=0.328) and reaching close to human performance (KL=0.679).

The key free parameters of the proposed approach are the radii for the centre, $r_{\mathcal{C}}$, and surround, $r_{\mathcal{S}}$, aggregation regions in the salience measure, (3). To illustrate the variability of the approach with respect to these parameters, Table 2 shows results for $r_{\mathcal{C}} \in \{2, 4, 8\}$ and $r_{\mathcal{S}} \in \{12, 24, 48\}$. Here, it is seen that the best performance, 0.624±0.012, with 0.012 the standard deviation, resulted from $r_{\mathcal{C}} = 2$ and $r_{\mathcal{S}} = 48$, resp, which are those presented elsewhere in this paper.

Finally, the current implementation of the proposed salience detection algorithm runs at an average rate of 473 ms/frame as implemented in unoptimized C++ and running under a Windows 7 OS on an Intel i7 2.4GHz processor (single core, no threading). The bulk of the processing time is taken by the computation of the spatiotemporal oriented energies at 3 scales. Here, it is interesting to note that a GPU implementation of the spatiotemporal oriented energy filtering can reduce its run time to 30 ms/frame [24].

Centre $r_{\mathcal{C}}$	2	2	**2**	4	4	4	8	8	8
Surround $r_{\mathcal{S}}$	12	24	**48**	12	24	48	12	24	48
KL (mean)	0.4931	0.5380	**0.6235**	0.1292	0.3414	0.4805	0.0000	0.2154	0.2408
KL (std.dev.)	0.0127	0.0117	**0.0122**	0.0051	0.0092	0.0103	0.0002	0.0065	0.0061

Table 2. Results of the proposed methods using different centre and surround radius values.

3.3 Discussion

Qualitatively, the image-based results illustrated in Figures 1, 2 and 3 show that the proposed approach typically provides its largest salience responses in intuitively reasonable locations. Upon visual inspection, it can be observed that the approach preferentially highlights dominant moving objects in home videos (e.g. *beverly-01*, *monica-05* and *beverly-07*) and television sports (e.g. *tv-sports01*, *tv-sports02* and *tv-sports05*), heads and especially mouths in talk shows (e.g. *tv-talk01*, *tv-talk03* and *tv-news06*) and video game targets (e.g. *gamecube04*, *gamecube13* and *gamecube18*). While the results suggest that response peaks can be fairly broad, their centroids are generally centred on targets of interest.

Qualitative comparisons to human performance are offered in Figure 4. Here, it becomes evident that humans are more conservative in their salience detection than is the proposed algorithm. In particular, while the proposed algorithm provides salience responses directly in terms of centre-surround contrast in spatiotemporal orientation, humans appear to focus their fixations further based on higher level information and preferences. For example, humans ignore cast shadows in favour of moving people and targets in beverly03 and gamecube18, while the proposed algorithm maps both people/target and shadow contours to high salience values. Similarly, humans selectively fixate on heads and faces in tv-music01 and tv-news04, while the algorithm also finds other contours as salient. Still, it is interesting to note that humans do not always restrict themselves to single salience peaks for a given video, as shown in beverly03, tv-ads05 and tv-music01. Overall, it appears that the proposed salience algorithm largely succeeds in detecting those points selected by humans as salient, but also marks additional points as salient.

The qualitative observations are strongly supported by the quantitative results. In particular, the proposed approach outperforms all the considered alternatives to set a new state-of-the-art. For example, its performance value, $D_{K,L} = 0.624$, is almost twice that of the second best performer (AIM), which achieves $D_{K,L} = 0.328$. Consequently, it is of interest to compare the AIM vs. the proposed SOE approach in terms of how they operate. The AIM approach makes use of a procedure to learn spacetime oriented Gabor-like filters from a large corpus of video data, which are not guaranteed to span the underlying visual spacetime structure; whereas, the proposed approach makes use of a predefined set of spacetime Gaussian derivative filters that are defined to provide a spanning basis set for 3D orientation structure (e.g., of visual spacetime). Further, while the AIM approach in essence considers global image support in making local salience determinations, the proposed approach employs more local comparisons, i.e., restricted to the employed centre-surround support regions. Significantly, these design choices allow the proposed approach to not only best a wide range of alternative algorithms, but to even approach human performance, $D_{K,L} = 0.679$.

The algorithm also appears to be well behaved with respect to variation in the supports used in centre-surround aggregations (Table 2). Interestingly, it is seen that maintaining relatively small centre and large surround yields best overall performance, which suggests that salience is reasonably defined (at least in the considered dataset) in terms of relatively small support regions in comparison to their surrounds.

4 Conclusions

This paper has presented a novel approach to spatiotemporal salience detection. The approach is based on two key ideas. First, visual spacetime is usefully characterized in terms of local distributions of 3D, (x, y, t), orientation measurements. Second, salience is defined in terms of the discrepancy between centre and surround aggregation regions of the local orientation measurements. While the current approach makes use of multiscale orientation measurements, it operates at a single pair of aggregation scales in its centre-surround comparisons. A potential direction for future research would be to make use of multiple paired aggregation regions. In this way both inner scale (local scale of oriented filtering) and outer scale (centre and surround aggregation regions) could be more fully exploited [38] in salience detection.

The entire approach has been implemented in software and evaluated on a publicly available dataset. The empirical results show that the approach outperforms a variety of alternative state-of-the-art algorithms for spatiotemporal

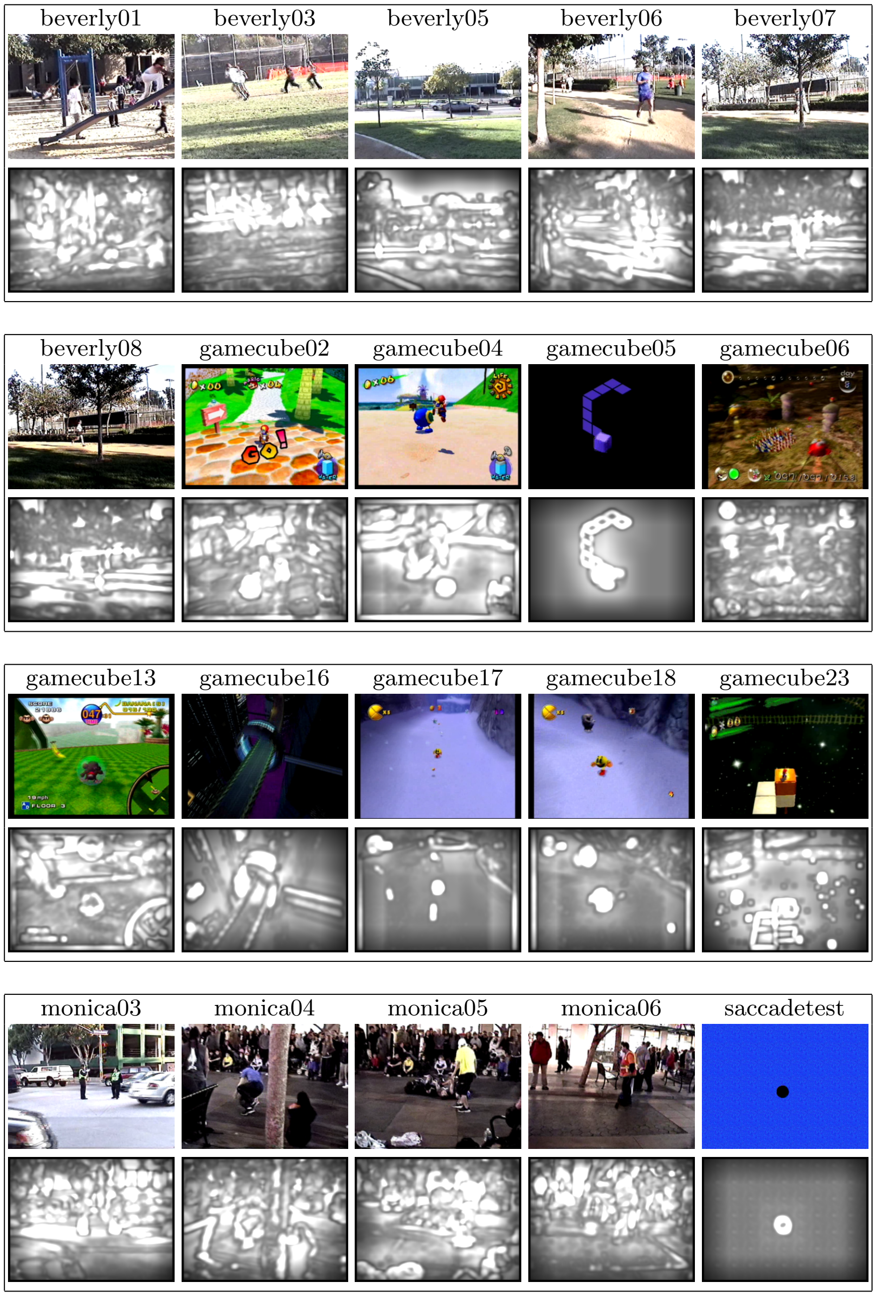

Fig. 1. Sample results from the USC dataset (one image per video) - videos 1-20: Upper row shows input image frames; lower row shows the saliency map produced by the proposed approach.

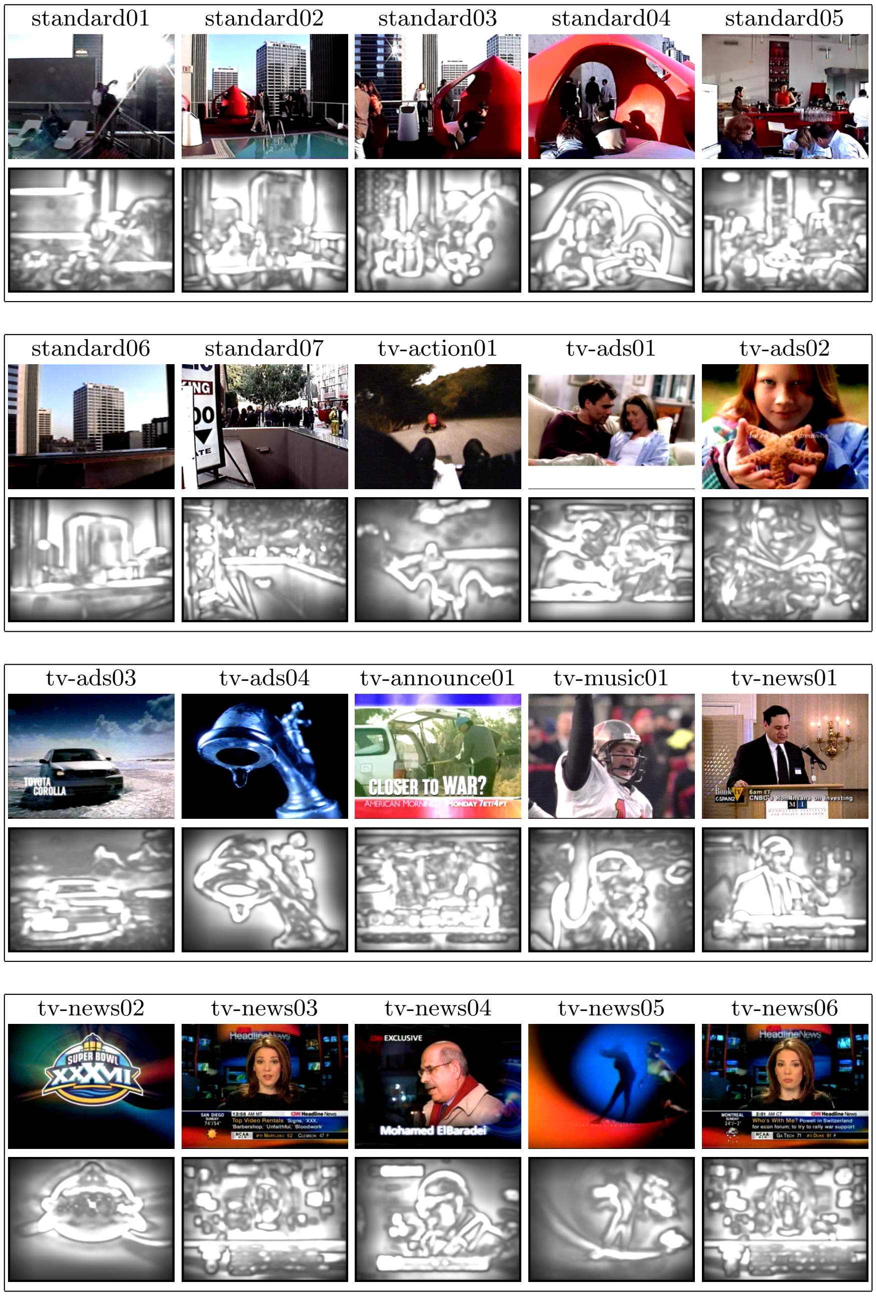

Fig. 2. Sample results from the USC dataset (one image per video) - videos 21-40: Upper row shows input image frames; lower row shows the saliency map produced by the proposed approach.

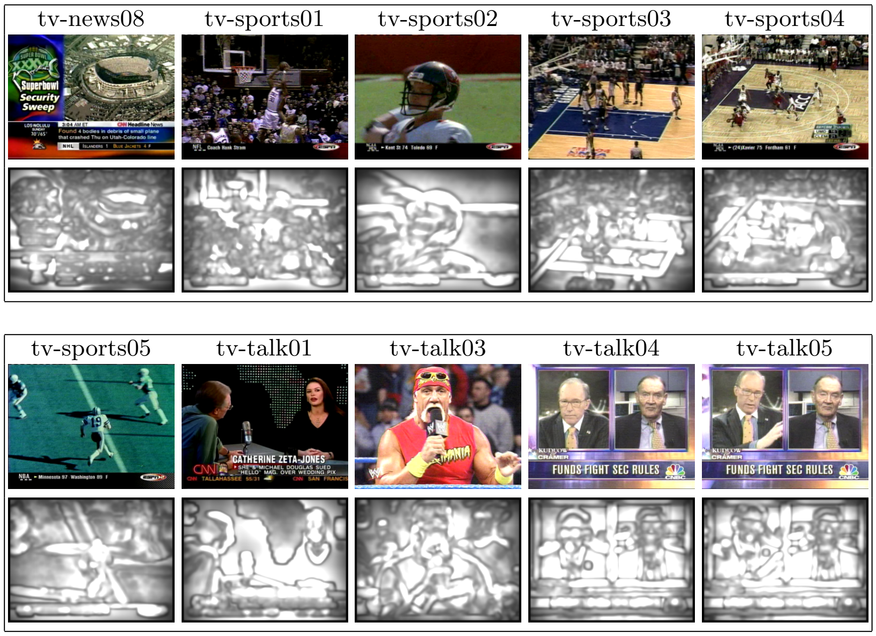

Fig. 3. Sample results from the USC dataset (one image per video) - videos 41-50: Upper row shows input image frames; lower row shows the saliency map produced by the proposed approach.

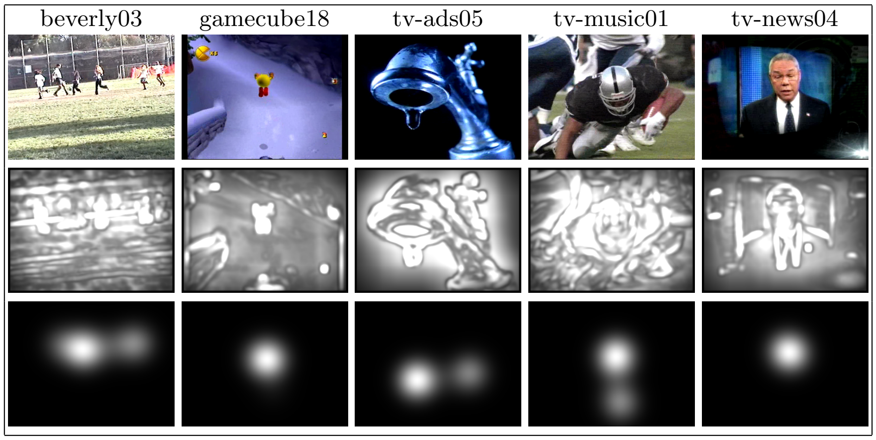

Fig. 4. Sample comparative results of the proposed approach vs. human fixation maps: The top row shows input image frames; the middle row shows saliency maps produced by the proposed approach; the bottom row shows the human derived fixation maps.

salience detection. Moreover, the results show that the evaluated implementation approaches human performance in selection of spatiotemporal salient points in video imagery.

References

1. Kandel, E., Schwartz, J.: Principles of Neuroscience. Elsevier, NY, NY (1996)
2. Derpanis, K., Wildes, R.: Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. PAMI **34** (2012) 1193–1205
3. Kullback, S.: Information Theory and Statistics. Dover, Mineola, NY (1968)
4. Picard, M.: Background subtraction techniques: A review. In: SMC. (2004) 3099–3104
5. Stauffer, C., Grimson, E.: Adaptive background mixture models for real-time tracking. In: CVPR. (1999) 2246–2252
6. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.: Background and foreground modeling using nonparametric kernel density for visual surveillance. Proc. IEEE **90** (2002) 1151–1163
7. Monnet, A., Mittal, A., Paragios, N., Ramesh, V.: Background modeling and subtraction for a moving observer. In: ICCV. (2003) 1305–1312
8. Sheikh, Y., Shah, M.: Bayesian modeling of dynamic scenes for object detection. PAMI **27** (2005) 1778–1792
9. Hayman, E., Eklundh, J.: Statistical background subtraction for a moving observer. In: ICCV. (2003)
10. Ren, Y., Chua, C., Ho, Y.: Motion detection with nonstationary background. MVA **13** (2003) 332–343
11. Wixson, L.: Detecting salient motion by accumulating directionally-consistent flow. PAMI **22** (2000) 774–780
12. Bugeau, A., Perez, P.: Detection and segmentation of moving objects in highly dynamic scenes. In: CVPR. (2007)
13. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vis. Res. **49** (2009) 1295–1306
14. Bruce, N., Tsotsos, J.: Towards a hierarchical representation of visual saliency. In: WAPCV. (2009) 98–111
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. PAMI **20** (1998) 1254–1259
16. Doretto, K., Chiuso, A., Wu, Y., Soatto, S.: Dynamic textures. IJCV **51** (2003) 91–109
17. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. PAMI **32** (2010) 171–177
18. Bence, M., Olveczky, P., Baccus, S.: Segregation of object and background motion in the retina. Nature (2003)
19. Heeger, D.: Model for the extraction of image flow. JOSA-A **2** (1987) 1455–1471
20. Granlund, G., Knuttson, H.: Signal Processing for Computer Vision. Kluwer, Norwell,MA (1995)
21. Simoncelli, E., Heeger, D.: A model of neuronal responses in visual area MT. Vis. Res. **38** (1996)
22. Chomat, O., Crowley, J.: Probabilistic recognition of activity using local appearance. In: CVPR. (1999) 104–109

23. Derpanis, K., Sizintsev, M., Cannons, K., Wildes, R.: Efficient action spotting based on a spacetime orientation structure representation. In: CVPR. (2010)
24. Zaharescu, A., Wildes, R.: Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event driven processing. In: ECCV. (2010)
25. Sadanand, S., Corso, J.: Action bank: A high-level representation of activity in video. In: CVPR. (2012)
26. Derpanis, K., Wildes, R.: Dynamic texture recognition based on distributions of spacetime oriented structure. In: CVPR. (2010)
27. Derpanis, K., Lecce, M., Daniilidis, K., Wildes, R.: Dynamic scene understanding: The role of orientation features in space and time in scene classification. In: CVPR. (2012)
28. Klaser, A., Marszalek, M., Schmid, C.: A spatiotemporal descriptor based on 3d-gradients. In: BMVC. (2008)
29. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
30. Derpanis, K., Sizintsev, M., Cannons, K., Wildes, R.: Action spotting and recognition based on a spatiotemporal orientation analysis. PAMI (to appear)
31. Freeman, W., Adelson, E.: The design and use of steerable filters. PAMI **13** (1991) 891–906
32. Lindeburg, T.: Scale-Space Theory in Computer Vision. Kluwer, Norwell, MA (1993)
33. Bracewell, R.: The Fourier Transform and its Applications. McGraw-Hill, NY, NY (2000)
34. Pearce, P., Pearce, S.: The Polyhedra Primer. Van Nostrand Reinhold, NY, NY (1978)
35. Jahne, B.: Digital Image Processing. Springer, Berlin (2005)
36. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV **57** (2004) 137–154
37. Duda, R., Hart, P., Stork, D.: Pattern Classification, Second Edition. Wiley, NY, NY (2000)
38. Koenderink, J.: The structure of images. Biological Cybernetics **50** (1984) 363–370