Efficient Unsupervised Abnormal Crowd Activity Detection Based on a Spatiotemporal Saliency Detector

Yilin Wang
Arizona State University
Tempe, Arizona
ywang370@asu.edu

Qiang Zhang
Samsung
Pasadena, California
zhangtemplar@gmail.com

Baoxin Li
Arizona State University
Tempe, Arizona
baoxin.li@asu.edu

Abstract

Approaches to abnormality detection in crowded scene largely rely on supervised methods using discriminative models. In this paper, we presents a novel and efficient unsupervised learning method for video analysis. We start from visual saliency, which has been used in several vision tasks, e.g., image classification, object detection, and foreground segmentation. To detect saliency regions in video sequences, we propose a new approach for detecting spatiotemporal visual saliency based on the phase spectrum of the videos, which is easy to implement and computationally efficient. With the proposed algorithm, we also study how the spatiotemporal saliency can be used in two important vision tasks, saliency prediction and abnormality detection. The proposed algorithm is evaluated on several benchmark datasets with comparison to the state-of-the-art methods from the literature. The experiments demonstrate the effectiveness of the proposed approach to spatiotemporal visual saliency detection and its application to the above vision tasks.

1. Introduction

Automatic abnormality detection for online multimedia content has been an active area in recent years due to it potential applications for crowded surveillance[30], social media behavior monitoring[35, 37, 34]and event retrieval[7]. Early approaches [30, 8, 24] focus on either generating discriminative model for semantic indexing the video or decompose it into semantic parts. These approaches, which rely on frame-based video labels, have been shown effective on certain datasets. Unfortunately, frame-based labels are in general hard to obtain. Especially, for massive YouTube videos, it is too labor-and time-intensive to obtain labeled sets large enough for robust training. Thus, the unsupervised approach would be more desirable. This paper studies unsupervised video abnormality detection.

Typically, video features such as optical flow, motion trajectory, and spatialtemporal interest points, lack of semantic meanings required by the abnormality detection. In the supervised case, label information could be directly utilized to build the connection between video features and video labels. Thus, unsupervised video abnormality detection is inherently more challenging than its supervised counterpart. In this paper, we start from visual saliency, which has attracted a lot of interests in the vision community in recent years. One early work that is widely known is the approach by Itti *et al.* [19]. Since then, a lot of different models have been proposed for computing visual saliency. Moreover, visual saliency often depends on not only a static scene but also the changes in the scene. To this end, spatiotemporal saliency has been proposed, which tries to capture regions attracting visual attention in the spatiotemporal domain. Spatiotemporal saliency has been applied to vision tasks such as video summarization, human-computer interaction [18], and video compression. However, these approaches only focus on the video objects or foreground, but ignore irregular motion pattern changes, which is an essential part in abnormality event detection. On the other hand, the saliency information can be regarded as an abstract of the video frame (image) [32], which may be exploited to enable unsupervised abnormality detection. How to achieve this is the objective of our approach.

In this paper, we study unsupervised video abnormality detection based on a spatiotemporal saliency detector by investigating two related challenges: (1) how to model the interaction between video content and spatiotemporal saliency systematically so as to augment video analysis using the information from saliency detection, and (2) how to use spatiotemporal saliency information to enable unsupervised video analysis. In addressing these two challenges, we propose a novel spatiotemporal visual saliency detector for video content analysis, based on *the phase information of the video*. With the saliency map computed using the proposed method, we analyze how it can be used for two fundamental vision tasks, namely saliency detec-

tion and abnormal event detection. We evaluate the performance of the proposed algorithm using several widely used datasets, with the comparison to the state-of-the-art in the literature. Our main contribution can be summarized as following: (1) A parameter free approach to enabling unsupervised video event detection. Neither normal examples nor abnormal examples are required for abnormality detection; (2) A novel and efficient framework for spatiotemporal saliency detection, which captures the global motion information and can be used to model complex activities. We demonstrate the complexity of the proposed algorithm is only $O(N \log N)$, where N is the size of the input; and (3) Comprehensive comparisons and evaluations using several benchmark datasets on saliency detection and abnormality detection are used to demonstrate that the effectiveness of the proposed approach, suggesting its potential application for future video analysis tasks.

2. The Proposed Method

2.1. Spectrum Analysis for Saliency Detection

There has been several explanations for why spectral domain based approach is able to detect saliency region from the image. For example, In [3], it has been shown that human visual system will select a subset of objects to focus. In other words, an attention competition exists among objects in the image. Only a small portion of objects, which are more distinctive, will be popped out, and rest of objects, which are usually in a uniform or common patterns, are suppressed. The spectral magnitude measures the total response of cells tuned to the specific frequency and orientation. According to lateral surround inhibition, similarly tuned cells will be suppressed depending on their total response, which can be modeled by dividing its spectral by the spectral magnitude [36]. [13] provided another explanation from sparse representation, which states that, if the foreground is sparse in spatial domain and background is sparse in DCT domain (e.g., periodic textures), the spectral domain based approach will highlight the foreground region in the saliency map. In a word, given an image (or 2D signal), $f(x, y)$, the saliency map can be calculated as:

$$S(x, y) = \mathcal{F}^{-1}[h(m, n) * \mathcal{A}(m, n) \cdot e^{-i\mathcal{P}(m,n)}] \quad (1)$$

where h is a high pass filter and $\mathcal{A}, \mathcal{P}$ represent amplitude and phase of Fourier transform $\mathcal{F}$.

2.2. Spectrum Analysis for Normal Videos

Spectrum analysis in spatiotemporal data, e.g., videos, is still in its infancy. In [16], the author studied how motion patterns contribute to saliency. It demonstrate that by setting the target object having different flicker rate, moving direction or motion velocity from other objects, the target object can be easily identified by human subjects. In this paper, we model the video abnormality detection as a spatiotemporal saliency detection problem, where normal video frame is regarded as non-salient and abnormality is perceptually salient.

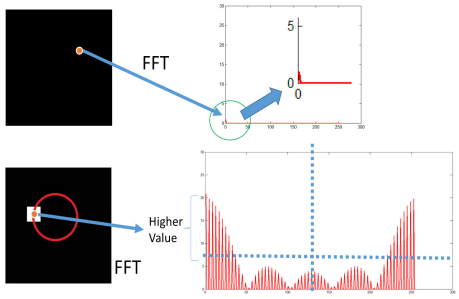

Figure 1. The spectrum analysis for normal videos. The first column is sample frame, the second column shows sampled video points in frequency domain. Please see the figures in color print.

In [23, 13], it has been shown that, for natural image, the amplitude spectrum of background (non-salient region) has higher value at lower frequency. Essentially, our observation demonstrates that, in the temporal domains, the normal video frames, where object can be modeled in a uniform motion pattern , will have higher amplitude in lower frequencies than higher frequencies. Thus, the phase information of temporal domain can be used for abnormality modeling. We show that one can exploit such informative observation through spectrum analysis.

In order to demonstrate the property of spectrum of normal videos, we generate two synthetic videos. The first video contains a uniform background (black) and second video with a moving object (33 by 33) which has its value uniformly distributed (white). Moreover, the motion trajectory of the object is followed as the red circle with same speed (we call it regular motion) in Figure 1. Specifically, two points, which are sampled from background and motion trajectory respectively, are further plotted in the frequency domain through the time period.

From Figure 1, we can interpret the result as following: 1) if no global motion in the video, the background (even with dynamic scenes, we show later) has higher response in the lower frequency domain. 2)Since the result of the spectrum obeys the symmetry [2], the amplitude from the points within regular motion object also trend to be higher in lower frequencies.

2.3. Modeling Video Abnormality via Saliency Detection

Based on the spectrum analysis of normal video, we have observed some potential properties. Then two research questions arises: 1) How to model the video abnormality only using the information from amplitude spectrum? 2)

How to automatically find the abnormality in a video? Answering these questions leads us to further analyze the amplitude spectrum with phase information. Given a signal $f(x, y, t)$ it is first transformed to the frequency domain $f(x, y, t) \xrightarrow{\mathcal{F}} \mathcal{F}(m, n, k)$, with the amplitude $\mathcal{A}(i, j, k) = |\mathcal{F}(m, n, k)|$ and phase $\mathcal{P}(m, n, k) = angle(\mathcal{F}(m, n, k))$. Based on the Fourier Transform and inverse Fourier Transform, we have:

$$\begin{aligned} f(x, y, t) &= \mathcal{F}^{-1}\left[\mathcal{F}(m, n, k)\right] \\ &= \mathcal{F}^{-1}\left[\mathcal{A}(m, n, k) \cdot e^{i \cdot \mathcal{P}(m,n,k)}\right] \end{aligned} \tag{2}$$

In order to extract the video abnormality, and inspired by the saliency detection, we perform a high pass filtering on the frequency domain in temporal dimension, which will suppress the signals from normal videos. Then we model the abnormality in a saliency fashion:

$$S(x, y, t) = g * \mathcal{F}^{-1}[h(k) * \mathcal{A}(m, n, k) \cdot e^{i \cdot \mathcal{P}(m,n,k)}] \tag{3}$$

where $h(k)$ is the high-pass filter along the temporal direction in the frequency domain, and g is a low pass filter in spatiotemporal domain, e.g., 3D Gaussian filter, which smooths the result. However, Eq 3 only considers the temporal information for video abnormality detection, which may involve the noise from background if the video contains global motion. To alleviate this issue, we further incorporate the spatial saliency information to refine the detection results. The improved model is described as below:

$$S(x, y, t) = g * \mathcal{F}^{-1}[h(k) * l(m, n) * (A(m, n, k) \cdot e^{i \cdot \mathcal{P}(m,n,k)})] \tag{4}$$

where $l(m, n)$ is the high pass filter along the spatial direction in the frequency domain. In frequency domain, setting the spectrum magnitude to uniform can achieve similar effect of high pass filter. In order to reduce the computation cost, we further relax Eq 4 (further analysis shown in Sec 2.4):

$$\begin{aligned} S(x, y, t) &= g * \mathcal{F}^{-1}[e^{i \cdot \mathcal{P}(m,n,t)}] \\ &= g * \mathcal{F}^{-1}(\frac{\mathcal{F}(m, n, k)}{|\mathcal{F}(m, n, k)|}) \end{aligned} \tag{5}$$

Eq 5 actually adopts the phase information of a video for saliency detection, it can be easily paralleled. The Fourier transform for multiple dimensional data can be computed as a sequence of 1D Fourier transform on each coordinate of the data. Thus the computation cost of the proposed spatiotemporal saliency detector is $O(MNTlog(MNT))$ when the input data size is $X \in \mathbb{R}^{M \times N \times \times T}$. If the data has P feature channel, then the computational cost is $O(PMNTlog(MNT))$.

2.4. Analysis

In this section, we provide evidence that, for a foreground object with irregular motion pattern, the proposed spatiotemporal saliency detector can approximately obtain its location in a video based on **Parseval's theorem** [2].

Parseval's theorem: *The energy in $u(t)$ equals the energy in $U(f)$, where $u(t) = \int_{f=-\infty}^{+\infty} U(f) \cdot e^{i2\pi ft} df$*

Now, given a 3D signal and reconstruct it with only phase information:

$$\begin{aligned} S(x, y, t) &= g * \mathcal{F}^{-1}[\frac{\mathcal{F}(m, n, k)}{|\mathcal{F}(m, n, k)|}] \\ &= g * \mathcal{F}^{-1}[1(m, n, k) \cdot e^{i \cdot \mathcal{P}(m,n,k)}] \end{aligned} \tag{6}$$

Based on **Parseval's theorem**, we know the summation across all the dimensions of $f(x, y, t)$ is equal to the summation across all the frequency component in the frequency domain. From the Eq 6, we can see that when only using the phase information it is equal to replace the amplitude spectrum $\mathcal{A}(t)$ to a cube. In other words, all of the elements which have non-zero value in magnitude spectrum are set to one. The region with repeat (regular) motion pattern creates a high peak in the magnitude spectrum (Figure 1) is suppressed; while the region with salient (irregular motion pattern instead corresponds to the spread-out magnitude spectrum will pop-out. Additionally, based on the proposition in [13], we can easily extend the sparse condition for saliency detection in the spatial domain to the spatiotemporal domain, which means the proposed method is also bounded with the ratio of salient region and non salient region. Due to the space limits, we omit the proof.

To verify the correctness of the proposed model, we generate one synthetic video to test the abnormality detection. The video contains a dynamic background with two moving squares with same texture. One of squares follows the red circle and moves steadily (we call normal object), another moving square moving randomly (we call abnormal object). The motion trajectory of these square is defined as following:

$$\begin{aligned} \Gamma_1(t) &= \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 128 + 64cos(\frac{\pi t}{32}) \\ 128 + 64sin(\frac{\pi t}{32}) \end{bmatrix} \\ \Gamma_2(t) &= \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 64 + 32cos(\frac{\pi t}{32}) + \epsilon \\ 64 + 32sin(\frac{\pi t}{32}) + \epsilon \end{bmatrix} \end{aligned}$$

where ϵ is a random variable uniformly draw from $[0, 128]$. We view the object with trajectory Γ_1 moving regularly. For the Gauss filter used to smooth the saliency map, we set the standard deviation as $0.006\sqrt{N^2 + M^2}$ and te filter size as $1 + 6\sigma$, where $N \times M$ is the size of each frame.

In Fig. 2, we show some sample frames of the video (top), the results from the proposed method (middle) with the comparison to the results of the method proposed in [11] (bottom), where the frame differences of two adjacent frames are used as temporal information. From the figure, we can find the proposed method highlight the moving objects over the dynamic background; in addition, the object moving "irregularly" (i.e., with trajectory Γ_2) gets higher values in the saliency map than the other object (the one with trajectory Γ_1) does. In contrast, the method proposed in [11] not only has problem in segmenting the moving objects from changing background, but also can't discriminate the one moving "irregularly" from the one moving regularly. One explanation could be that simply the frame differences of two adjacent frames can't distinguish the object moving irregularly from the object moving reguarly. This reveals the potential of the proposed method to detect irregular events from the video (or abnormal events), as detailed in the next section.

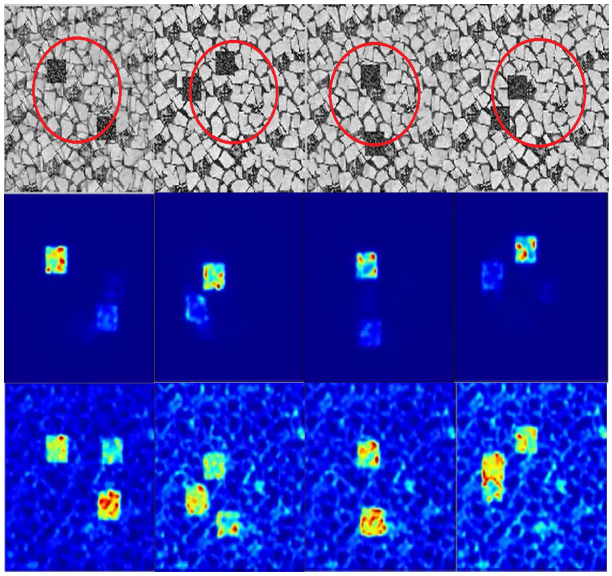

Figure 2. Top: some sample frames from the input video, where the red circle indicates the trajectory Γ_1; middle: the corresponding saliency maps; and bottom: the saliency map computed with method described in [11]. For the saliency map, the warm color indicates high value and cold color for low value. The video can be found in the supplementary material. Please see the figures in color print.

3. Experiment

Since the proposed method is based on saliency detection, to verify the correctness in saliency detection, we first evaluate the proposed method on both syntheic data (Sec. 3.1) and two real-world video datasets (Sec. 3.2), CRCNS-ORIG and DIEM, for saliency detection. Then we evaluate the proposed method on several benchmark datasets on abnormality detection. The performance of the proposed methods are compared with the existing methods, some of which are state-of-the-art methods.

3.1. Simulation Experiment

In this section, we evaluate the proposed method on synthetic data. In [16], how three properties of motion, namely flicker, direction and velocity, contribute to the saliency was studied. In this section, we generate the synthetic data according to the their protocol. The input data is a short clip where the resolution is 174×174 with 400 frames at the frame rate of 60 frames per second. We put 36 objects of size 5×13 in a 6×6 grid and a target object is randomly selected out of those 36 objects. All the objects are allowed to move within a 29×29 region centered at their initial position (and warped back, if they move out of this region). The video is black-and-white. We design the following three experiments:

1. **Flicker**: we set the objects on-off at a specified rate and the target object at a different rate from the other 35 objects;
2. **Direction**: we set the objects moving in a specified direction and the target object in a different direction. The velocity of all the objects are the same;
3. **Velocity**: we set the objects moving in a specified velocity and the target object moves in a different velocity. The moving direction of the all the objects are the same.

All the other parameters are the same as used in [16]. According to [16], the target object could be easily identified by human subjects, when its motion property (e.g., flicker rate, moving direction, velocity) is different from the other objects. We also include some "blind" trials, where the target object has the same motion property as the other 35 objects. In this case, the target object can't be identified by the human subjects, i.e., there is no salient region.

We apply the proposed method to the input data. For comparison, we also evaluate the method proposed in [4] and [13]. We use the area under receiver operating characteristic curve as the performance metric. The ground truth mask is generated according to the location of the target object. The experiment result is shown in Figure 3.

From the experiment results, we can find that the proposed method detects the salient region much more accurately than [4] and [13] in all except the "blind" trials, which should be as lower as possible in terms of abnormality. However, [4] and [13] don't survive in those "blind" trials. Surprisingly, [4] and [13] achieves quite similar performances, though [4] was supposed to achieve better result as it include the differences of two adjacent frames as motion (temporal) information.

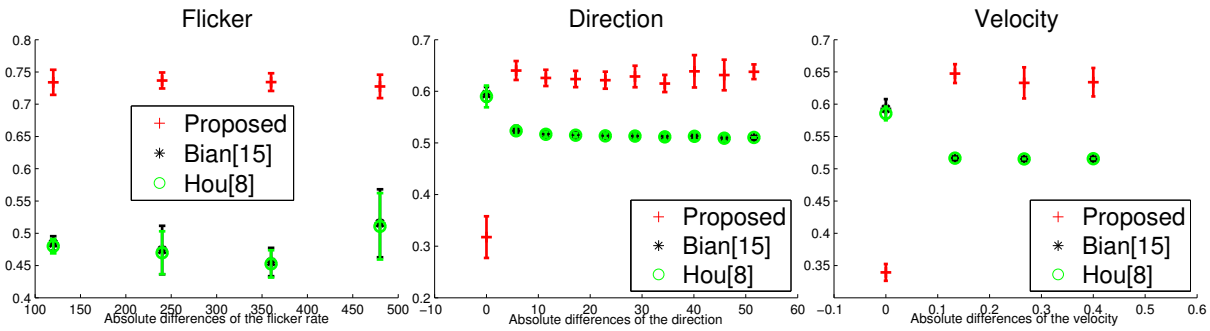

Figure 3. The AUC on the synthetic data for the proposed method and two existing methods. For "Direction" and "Velocity", we also include some "blind" trials (X-axis has value 0), where the target object has exactly the same motion property as the other 35 objects. In those trials, the target object can't be identified by human subjects, i.e., there is no salient object [16].

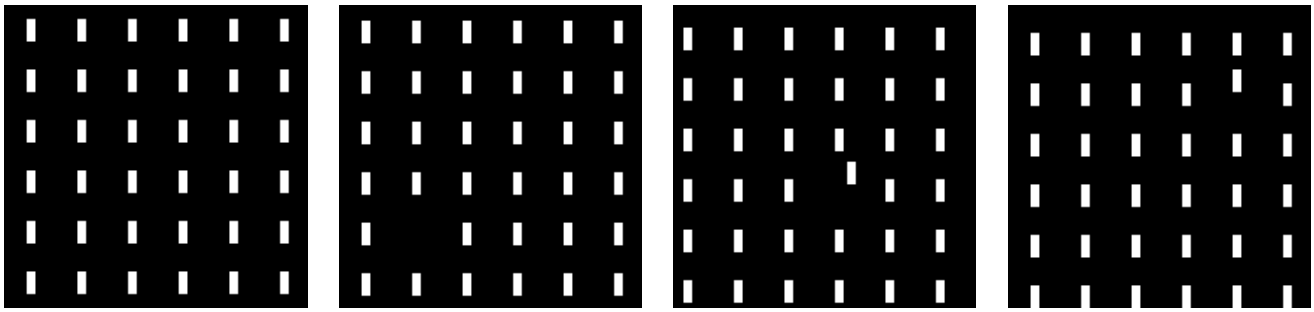

Figure 4. Some visual sample of the synthetic data for different experiments.

3.2. Spatiotemporal Saliency Detection

In previous section, we test the proposed spatiotemporal saliency detector on synthetic videos, with the comparison to two other saliency detectors, where the proposed detector shows better performances in capturing the temporal information. In this section, we evaluate the proposed spatiotemporal saliency detector on two challenging video datasets for saliency evaluation, CRCNS-ORIG [20] and DIEM [29]. For this experiment, we first convert each frame into the LAB color space, then compute the spatiotemporal saliency in each channel independently and the final spatiotemporal saliency is the summation of the saliency maps of all three channels.

CRCNS-ORIG includes 50 video clips from different genres, including TV programs, outdoor scenes and video games. Each clip is 6-second to 90-second long at 30 frames per second. The eye fixation data is captured from eight subjects with normal or correct-normal vision. In our experiment, we downsample the video from 640×480 to 160×120 and keep the frame rate untouched, then apply the our spatiotemporal saliency detector. To measure the performance, we compute the area under curve (AUC) and F-measure (harmonic mean of true positive rate and false positive rate). The experiment result is shown in Fig. 5, where the area under curve (AUC) is 0.6639 and F-measure is 0.1926. Tab. 1 compares the result of the proposed method with some state-of-art methods on CRCNS-ORIG, which indicates that our method outperforms them by at least 0.06 regarding AUC.

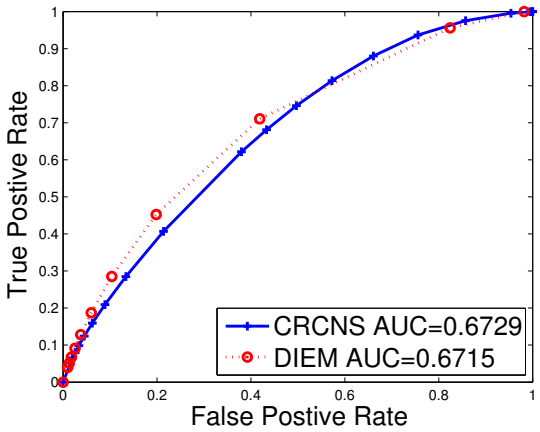

Figure 5. The receiver operating characteristic curve of the propose method in CRCNS-ORIG dataset and DIEM dataset. The area under the curve is 0.6639 and 0.6896 accordingly.

DIEM dataset collects data of where people look during dynamic scene viewing such as film trailers, music videos, or advertisements. It currently consists of data from over

Method	AUC	Method	AUC
AWS [10]	0.6000	AWS [10]	0.5770
HouNIPS [15]	0.5967	Bian [4]	0.5730
Bian [4]	0.5950	Marat [27]	0.5730
IO	0.5950	Judd [21]	0.5700
SR [14]	0.5867	AIM [6]	0.5680
Torralba [33]	0.5833	HouNIPS [15]	0.5630
Judd [21]	0.5833	Torralba [33]	0.5840
Marat [27]	0.5833	GBVS [12]	0.5620
Rarity-G [26]	0.5767	SR [14]	0.5610
CIOFM [17]	0.5767	CIO [17]	0.5560
Proposed	0.6639	Proposed	0.6896

Table 1. The result the proposed method compared with the results of the top ten existing methods on CRCNS dataset (left) and DIEM dataset (right) according to [5]. From this table, we can find that the propose method gets obvious better performances than the state-of-arts on both two datasets.

250 participants watching 85 different videos. Each video in DIEM dataset includes 1000 to 6000 frames at 30 frames per second. Similarly as CRCNS, we downsample the video to $1/4$ (e.g., from 1280×720 to 320×180) while maintaining the aspect ratio and frame rate. We observe that each video in DIEM dataset is consisted of a sequence of short clips, where each clip has 30 to 100 frames. To properly detect the saliency from those videos, we apply the window function to our spatiotemporal saliency detector, where the size of the window (along temporal direction) is 60-frame. The experiment result is shown in Fig. 5 and Tab. 1, where the AUC is 0.6896 and F-measure is 0.35. From the table, we can find that the proposed method outperforms the state-of-arts by over 10%.

To compare the performances of combining four visual cues via QFT and performances via summation of saliency maps of each visual cues, we design the following experiment. We run 1000 simulations and in each simulation we generate a $r \times c \times 4$ array, where r and c is a random number between $[1, 1000]$ and 4 is the number of feature channels. We compute the saliency map with different methods then measures their similarities via cross-correlation, where 0.91 is reported for QFT and FFT. After smoothing the saliency map with a Gaussian kernel, the correlation is over 0.998. For natural image, we could expect an even higher correlation.

This suggests that, we can compute the saliency map for each visual cue independently and then add them together, which will yield quite similar result by using quaternion Fourier transform. In addition, the proposed method other than QFT provides more flexibility, e.g., we can assign different weights to the visual cues as [21].

We also include the AUC of the proposed method for each video from the CRCNS-ORIG (Figure 6) and DIEM dataset (Figure 7).

Method	AUC
Optical flow [28]	0.84
Social force [28]	0.96
NN [9]	0.93
Sparse reconstruction [9]	0.978
Proposed	0.9378

Table 2. The result on UMN dataset. Note, we have cropped out the region which contains the text "abnormal", and results in frame resolution 214×320. *Please note that, most of those methods, except the proposed one, need a training stage.*

3.3. Abnormality Detection

In this section, we show how can we utilize the proposed spatiotemporal saliency detector to detect abnormality from the video.

Method	Ped1	Ped2	Overall
Social force [28]	31%	42%	37%
MPPCA [22]	40%	30%	35%
MDT [25]	25%	25%	25%
Adam [1]	38%	42%	40%
Reddy [31]	22.5%	20%	21.25%
Sparse [9]	19%	$N.A.$	$N.A.$
Proposed	27%	19%	23%

Table 3. The frame level EER (the lower the better) for UCSD dataset. Please note that, most of those methods, except the proposed one, need a training stage. From the result, we can found that the proposed method, even without traing stage or training data, can still outperform social force, MPPCA.

For abnormality detection, we start with computing the saliency map for the input video as described above. The regions containing abnormalities can be detected by founding the region where the saliency value is above a threshold. Then the saliency score of a frame is computed as the average of saliency value of the pixels in that frame, i.e.,

$$\mathbf{s}(t) = \frac{1}{NM} \sum_i \sum_j \mathbf{X}(i, j, t) \tag{7}$$

where $\mathbf{s}(t)$ is the saliency score of t_{th} frame, $N \times M$ is the size of one frame, i, j, t are row, column and frame index of the 3D saliency map accordingly. The frame with high saliency score would contain abnormality. To show the proposed method is not sensitive to the value of threshold, we choose the average value of the saliency in the video as threshold.

We evaluate the proposed method for abnormality detection in videos from two datasets: UMN abnormal dataset[1] and UCSD dataset [25]. Abnormal detection has attracted

[1] http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi

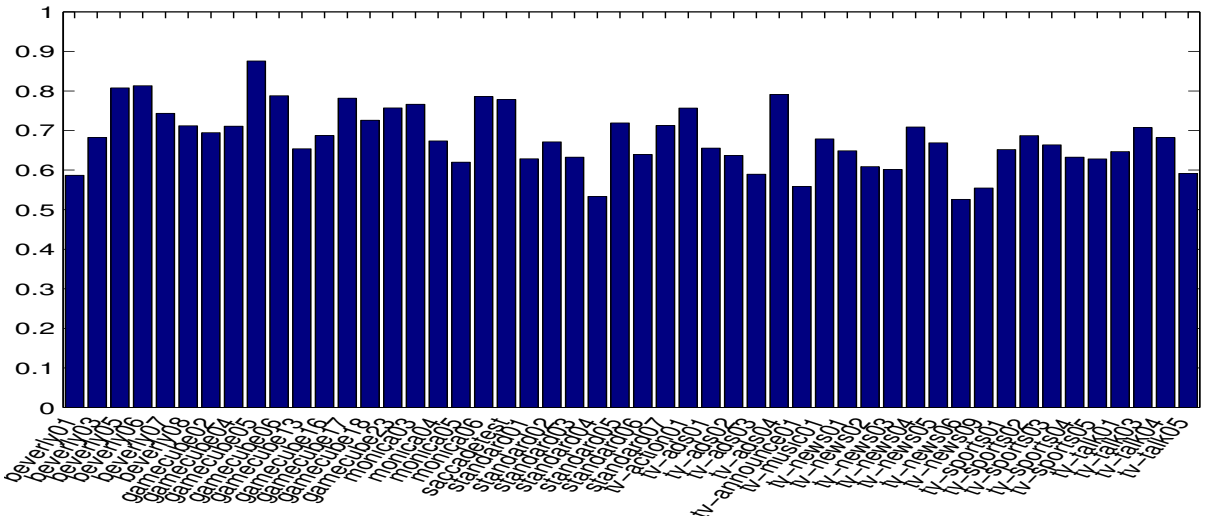

Figure 6. The AUC of the proposed method for each video from CRCNS-ORIG dataset.

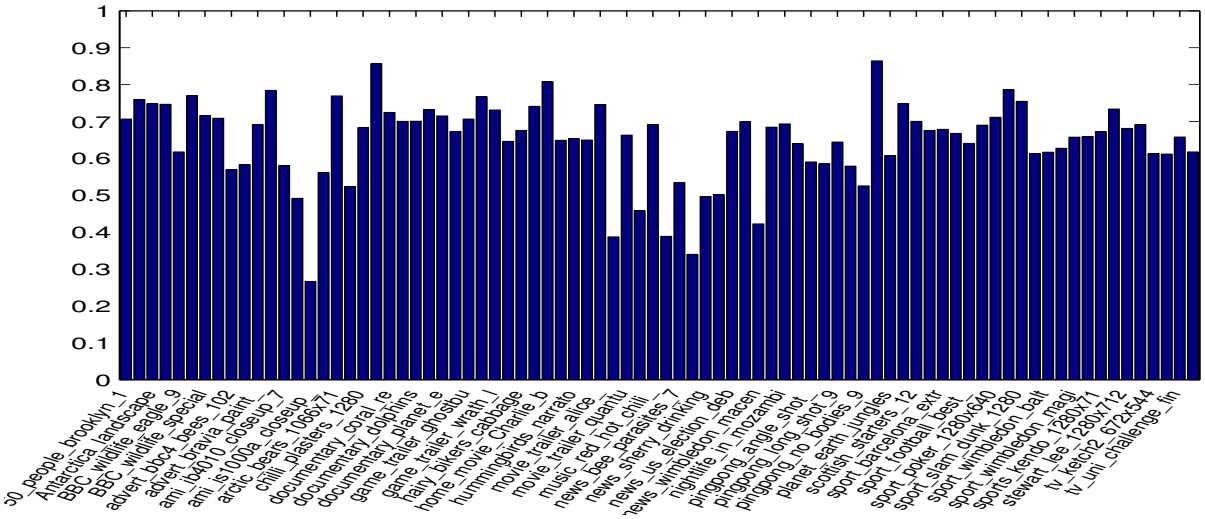

Figure 7. The AUC of the proposed method for each video from DIEM dataset.

a lot efforts from the researchers. However, most of the existing works require training stage, e.g., social force [28], sparse reconstruction [9], MPPCA [22], i.e., they need training data to initialize the model. The proposed method, instead, does **NOT** need any training stage or training data.

The result on UMN abnormal dataset is shown in Tab. 2, where we compute the frame-level true positive rate and false positive rate then compute the area under the ROC (Fig. 9). Fig. 10 shows the result for videos of three scenes, where we plot saliency value of each frame and show some sample frames. The result on UCSD dataset is shown in Tab. 3, where we report frame-level equal-error rate (EER) [25]. Fig. 11 shows the ROC for UCSD dataset with the proposed method; Fig. 8 shows eight samples frames, where red color highlights abnormal regions. We can find that, without training data, the proposed method still outperforms several state-of-arts in the literature, e.g., social force, MPPCA.

4. Conclusion and Discussion

In this paper, we proposed a novel approach for detecting spatiotemporal saliency, which was simple to implement and computationally efficient. The proposed approach was inspired by recent development of spectrum analysis based visual saliency approaches, where phase information was used for constructing the saliency map of the image. Recognizing that the computed saliency map captured the region of human's attention for dynamic scenes, we proposed two algorithms utilizing this saliency map for two important vision tasks. These approaches were evaluated on several well-known datasets with comparisons to the state-of-arts in the literature, where good results were demonstrated.

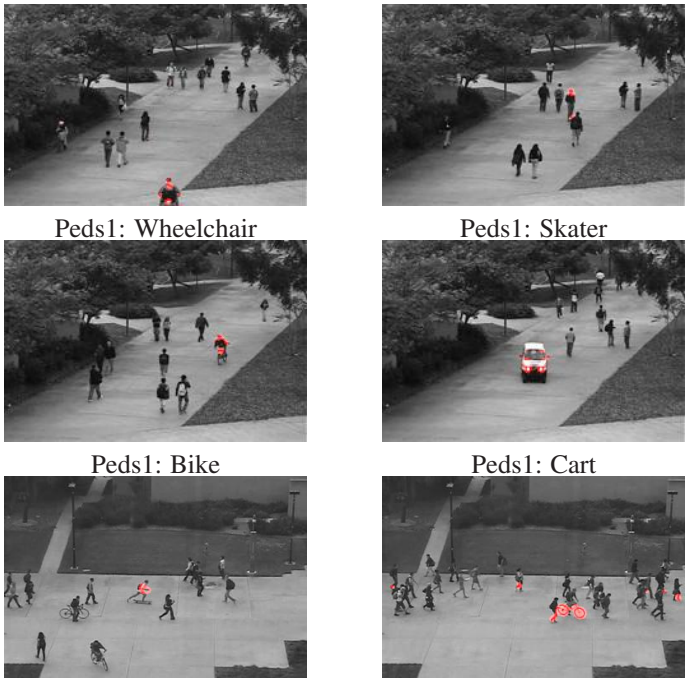

Figure 8. Some sample results for the UCSD datasets, where the red color highlights the detected abnormal region, i.e., the saliency value of the pixel is higher than four times of the mean saliency value of the video. Please see the figures in color print.

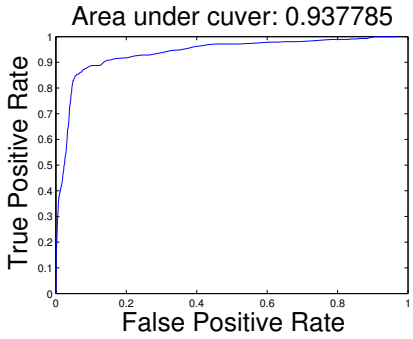

Figure 9. The ROC for the UMN dataset computed with the propose method.

5. Acknowledgment

The work was supported in part by an ARO grant (#W911NF1410371) and an ONR grant (# N00014-15-1-2722). Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of ARO or ONR.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *PAMI*, 30(3):555 –560, march 2008. 6

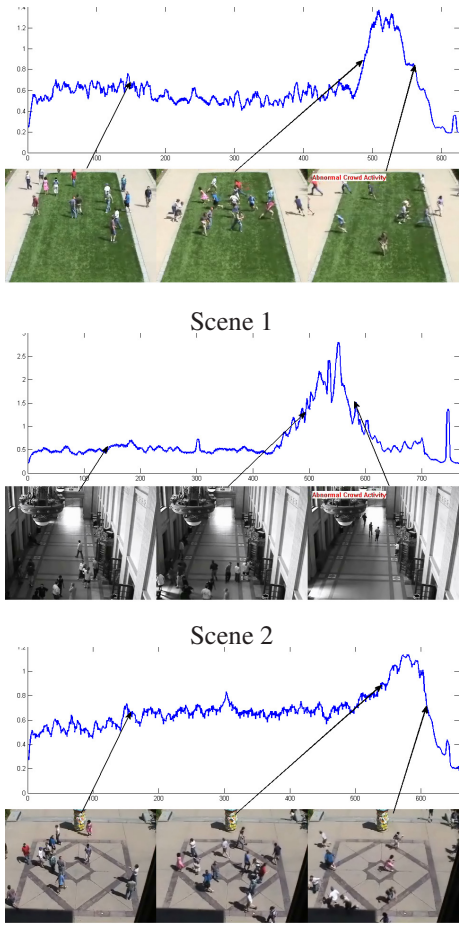

Figure 10. Some sample results for the UMN datasets, where we pick one video for each scene. The top is the saliency value (Y-axis) for each frame (X-axis) and bottom are sample frames picked from different frames (as shown by the arrow).

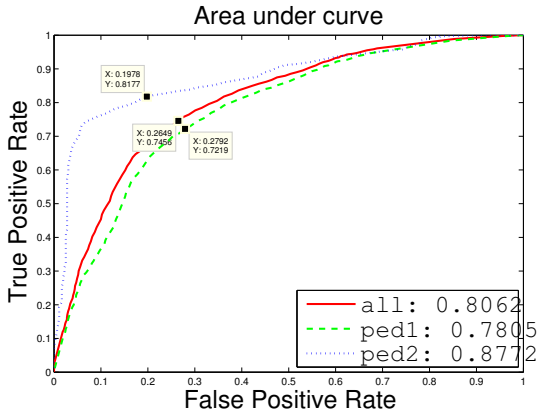

Figure 11. The ROC for the UCSD dataset computed with the propose method.

[2] A. Antoniou. *Digital signal processing*. McGraw-Hill Toronto, Canada:, 2006. 2, 3

[3] D. M. Beck and S. Kastner. Stimulus context modulates com-

petition in human extrastriate cortex. *Nature neuroscience*, 2005. 2

[4] P. Bian and L. Zhang. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. *NIPS*, pages 251–258, 2009. 4, 6

[5] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. 2012. 6

[6] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, pages 155–162, 2005. 6

[7] L. Chen, Q. Zhang, P. Zhang, and B. Li. Instructive video retrieval for surgical skill coaching using attribute learning. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6, June 2015. 1

[8] X. Chen and K. S. Candan. GI-NMF: group incremental non-negative matrix factorization on data streams. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1119–1128, 2014. 1

[9] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449 –3456, june 2011. 6, 7

[10] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Decorrelation and distinctiveness provide with human-like saliency. In *NIPS*, pages 343–354. Springer, 2009. 6

[11] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, 2008. 4

[12] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 6

[13] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *PAMI*, pages 194–201, 2012. 2, 3, 4

[14] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1 –8, 2007. 6

[15] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *NIPS*, 21:681–688, 2008. 6

[16] D. E. Huber and C. G. Healey. Visualizing data with motion. In *Visualization, 2005. VIS 05. IEEE*, pages 527–534. IEEE, 2005. 2, 4, 5

[17] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *NIPS*, 18:547, 2006. 6

[18] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Optical Science and Technology*, pages 64–78. International Society for Optics and Photonics, 2004. 1

[19] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, nov 1998. 1

[20] R. Itti, Laurent; Carmi. Eye-tracking data from human volunteers watching complex video stimuli. Online, 2009. 5

[21] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV 2009*, pages 2106 –2113, 29 2009-oct. 2 2009. 6

[22] J. Kim and K. Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *CVPR 2009*, pages 2921 –2928, june 2009. 6, 7

[23] J. Li, M. D. Levine, X. An, X. Xu, and H. He. Visual saliency based on scale-space analysis in the frequency domain. *PAMI*, pages 996–1010, 2013. 2

[24] X. Li, S. Huang, K. S. Candan, and M. L. Sapino. Focusing decomposition accuracy by personalizing tensor decomposition (PTD). In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 689–698, 2014. 1

[25] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975 –1981, june 2010. 6, 7

[26] M. Mancas. *Computational attention: Modelisation and application to audio and image processing*. PhD thesis, PhD. Thesis, University of Mons, 2007. 6

[27] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *IJCV*, 82(3):231–243, 2009. 6

[28] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR 2009*, pages 935 –942, june 2009. 6, 7

[29] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011. 5

[30] R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino. Optimizing interaction force for global anomaly detection in crowded scenes. In *Computer Vision Workshops (ICCV Workshops)*, pages 136 –143, nov. 2011. 1

[31] V. Reddy, C. Sanderson, and B. Lovell. Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *CVPRW*, pages 55 –61, june 2011. 6

[32] Q. Tian and B. Li. Simultaneous semantic segmentation of a set of partially labeled images. In *IEEE Winter Conference on Applications of Computer Vision*, 2016. 1

[33] A. Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003. 6

[34] Y. Wang, Y. Hu, S. Kambhampati, and B. Li. Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 473–482, 2015. 1

[35] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li. Unsupervised sentiment analysis for social media images. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 2378–2379, 2015. 1

[36] L. Zhaoping and P. Dayan. Pre-attentive visual selection. *Neural Networks*, 19(9):1437–1439, 2006. 2

[37] D. Zhou, J. He, K. S. Candan, and H. Davulcu. MUVIR: multi-view rare category detection. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4098–4104, 2015. 1