Spatio-temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform

Chenlei Guo, Qi Ma and Liming Zhang
Department of Electronic Engineering, Fudan University
No.220, Handan Road, Shanghai, 200433, China
http://homepage.fudan.edu.cn/~clguo , ma.lance@gmail.com, lmzhang@fudan.edu.cn

Abstract

Salient areas in natural scenes are generally regarded as the candidates of attention focus in human eyes, which is the key stage in object detection. In computer vision, many models have been proposed to simulate the behavior of eyes such as SaliencyToolBox (STB), Neuromorphic Vision Toolkit (NVT) and etc., but they demand high computational cost and their remarkable results mostly rely on the choice of parameters. Recently a simple and fast approach based on Fourier transform called spectral residual (SR) was proposed, which used SR of the amplitude spectrum to obtain the saliency map. The results are good, but the reason is questionable.

In this paper, we propose it is the phase spectrum, not the amplitude spectrum, of the Fourier transform that is the key in obtaining the location of salient areas. We provide some examples to show that PFT can get better results in comparison with SR and requires less computational complexity as well. Furthermore, PFT can be easily extended from a two-dimensional Fourier transform to a Quaternion Fourier Transform (QFT) if the value of each pixel is represented as a quaternion composed of intensity, color and motion feature. The added motion dimension allows the phase spectrum to represent spatio-temporal saliency in order to engage in attention selection for videos as well as images.

Extensive tests of videos, natural images and psychological patterns show that the proposed method is more effective than other models. Moreover, it is very robust against white-colored noise and meets the real-time requirements, which has great potentials in engineering applications.

1. Introduction

Most traditional object detectors need training in order to detect specific object categories [1, 2, 3], but human vision can focus on general salient objects rapidly in a clustered visual scene without training because of the existence of visual attention mechanism. So human can easily deal with general object detection well, which is becoming an intriguing subject for more and more researches.

What attracts people's attention? Tresiman [4] proposed a theory which describes that visual attention has two stages. A set of basic visual features such as color, motion and edges is processed in parallel at pre-attentive stage. And then a limited-capacity process stage performs other more complex operations like face recognition and *etc*. [5]. Distinctive features (*e.g*. luminous color, high velocity motion and *etc*.) will ”pop out” automatically in the pre-attentive stage, which become the object candidates.

Several computational models have been proposed to simulate human's visual attention. Itti *et al*. proposed a bottom-up model and built a system called Neuromorphic Vision C++ Toolkit (NVT) [6]. After that, following Rensink's theory [7], Walther extended this model to attend to *proto object* regions and created SaliencyToolBox (STB) [8]. He also applied it to the object recognition tasks [9]. However, the high computational cost and variable parameters are still the weaknesses of these models. Recently, Spectral Residual (SR) approach based on Fourier Transform was proposed by [10], which does not rely on the parameters and can detect salient objects rapidly. In this approach, the difference (SR) between the original signal and a smooth one in the log amplitude spectrum is calculated, and the saliency map is then obtained by transforming SR to spatial domain. All these models mentioned above, however, only consider static images. Incorporating motion into these models is a challenging task that motivates us to develop a novel method to generate spatio-temporal saliency map.

After careful analysis, we find that SR of the amplitude spectrum is not essential to obtain the saliency map in [10]; however, the saliency map can be calculated by the image's Phase spectrum of Fourier Transform (PFT) alone. Moreover, this discovery of PFT provides an easy way to extend our work to Quaternion Fourier Transform (QFT) [11]. Each pixel of the image is represented by a

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

quaternion that consists of color, intensity and motion feature. The Phase spectrum of QFT (PQFT) is used to obtain the spatio-temporal saliency map, which considers not only salient spatial features like color, orientation and *etc*. in a single frame but also temporal feature between frames like motion.

In section 2, we propose PFT and discuss the relationship between PFT and SR. In section 3, we introduce the quaternion representation of an image and propose PQFT to obtain spatio-temporal saliency map. Many experimental results of comparing our methods with others are shown in section 4, and the conclusions and discussions are given thereafter.

2. From Phase spectrum of Fourier Transform to the Saliency Map

It was discovered that an image's SR of the log amplitude spectrum represented its innovation. By using the exponential of SR instead of the original amplitude spectrum and keeping the phase spectrum, the reconstruction of the image results in the saliency map [10]. However, we find that this saliency map can be calculated by PFT regardless of the amplitude spectrum value, which motivates us to explore the role of PFT in obtaining the saliency map.

2.1. What does the phase spectrum represent?

What can be detected from the reconstruction that is calculated only by the phase spectrum of the input signal? In order to show its underlying principle, we give three one-dimensional waveforms shown in Fig.1 (left) and hope to find some intrinsic rules from their reconstructions of PFT.

Note for the following examples in Fig.1, the reconstruction is obtained by the phase spectrum alone. When the waveform is a positive or negative pulse, its reconstruction contains the largest spikes at the jump edge of the input pulse. This is because many varying sinusoidal components locate there. In contrast, when the input is a single sinusoidal component of constant frequency, there is no distinct spike in the reconstruction. Less periodicity or less homogeneity of a location, in comparison with its entire waveform, creates more "pop out".

The same rule can be applied to two-dimension signals like images as well. [12] pointed out that the amplitude spectrum specifies how much of each sinusoidal component is present and the phase information specifies where each of the sinusoidal components resides within the image. The location with less periodicity or less homogeneity in vertical or horizonal orientation creates the "pop out" *proto objects* in the reconstruction of the image, which indicates where the object candidates are located.

Please note that we do not take the borders of the signals or images into consideration because of their discontinuity.

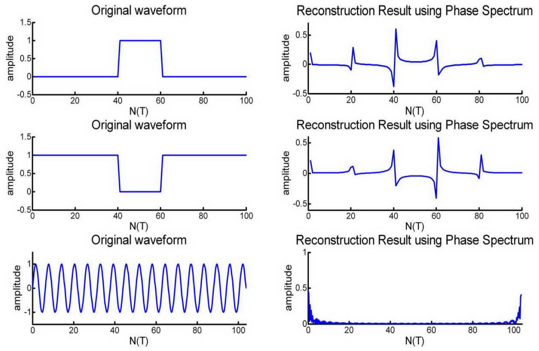

Figure 1. One dimension data examples. (left) Original data (right) Reconstructions only by the phase spectrum.

2.2. The steps of PFT approach

According to discovery above, we propose the PFT approach to calculate the saliency map. The steps (Eq.1-3) can be summarized as follows:

Given an image $I(x, y)$,

$$f(x, y) = F(I(x, y)) \tag{1}$$

$$p(x, y) = P(f(x, y)) \tag{2}$$

$$sM(x, y) = g(x, y) * \|F^{-1}[e^{i \cdot p(x,y)}]\|^2 \tag{3}$$

where F and F^{-1} denote the Fourier Transform and Inverse Fourier Transform, respectively. $P(f)$ represents the phase spectrum of the image. $g(x, y)$ is a 2D gaussian filter ($\sigma = 8$), which is the same with SR [10]. The value of the saliency map at location (x, y) is obtained by Eq.3. In SR approach, there needs to add spectral residual of the log amplitude spectrum to the square bracket of Eq. 3.

We use the intensity of the image as the input to PFT and SR in the following experiments.

2.3. Comparison between PFT and SR

It is obvious that PFT omits the computation of SR in the amplitude spectrum, which saves about 1/3 computational cost (see Section 4.2 and 4.3 for details). How different are the saliency maps? In this subsection, we analyze the saliency maps computed from PFT and SR and give some results.

We use the database of [10] as the test images. The database contains 62 natural images with resolution of around 800×600.

Define the saliency map of image i from PFT as sM_i^1 and that from SR as sM_i^2. Fig.2 shows the results of three natural images, in which sM_i^1 and sM_i^2 look the same. In order to evaluate the similarity between these saliency maps

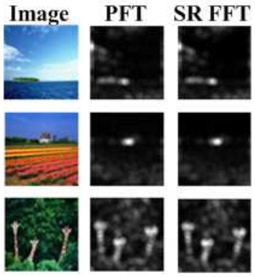

Figure 2. Test results from three input images. (left) Input images, (middle) Saliency maps from PFT, (right) Saliency maps from SR.

Scales	Max MPD	Min MPD	Average MPD
64×64	0.0342	0.0048	0.0123
128×128	0.0433	0.0061	0.0200
256×256	0.0477	0.0081	0.0217
512×512	0.0379	0.0040	0.0181

Table 1. The Maximum, Minimum and Average MPD of the saliency maps from PFT and SR in different resolutions.

in quantity, we introduce the Maximum Pixel Difference (MPD_i) of the image i as:

$$MPD_i = \max_{\substack{x=1\to w\\ y=1\to h}} |sM_i^1(x,y) - sM_i^2(x,y)| \qquad (4)$$

where (x, y) is the location of each pixel in the image. w and h are the image's width and height, respectively.

Table 1 shows that Maximum, Minimum and Average MPD of the entire database in four different resolutions are negligible (The maxium pixel value is 1). Thus we can deduce that SR contributes very little to the saliency map and PFT is enough to obtain the saliency map. Their slight difference in performance will be discussed in section 4.

3. From PQFT to Spatio-temporal Saliency Map

As mentioned above, PFT provides a simpler, faster way to obtain saliency map than SR. Moreover, it motivates us to develop PQFT to obtain spatio-temporal saliency map easily. Compared with the saliency maps in [6, 8, 10], our PQFT considers the motion features between sequent frames.

Our method can be divided into two stages. First, the image should be represented as a quaternion image which consists of four features. Second, PQFT needs to be calculated in order to obtain the spatio-temporal saliency map.

3.1. Create a quaternion image

Define the input image captured at time t as $F(t), t = 1 \cdots N$, where N is the total frame number. $r(t), g(t), b(t)$ are the red, green and blue channel of $F(t)$. Four broadly-tuned color tunnels are created by Eq.5 - 8 [6]:

$$R(t) = r(t) - (g(t) + b(t))/2 \qquad (5)$$

$$G(t) = g(t) - (r(t) + b(t))/2 \qquad (6)$$

$$B(t) = b(t) - (r(t) + g(t))/2 \qquad (7)$$

$$Y(t) = (r(t) + g(t))/2 - |r(t) - g(t)|/2 - b(t) \qquad (8)$$

In human brain, there exists a 'color opponent-component' system. In the center of receptive fields, neurons which are excited by one color (eg. Red) are inhibited by another color (eg. Green). Red/green, green/red, blue/yellow and yellow/blue are color pairs which exists in human visual cortex [13]. Thus the color channels are obtained as follows:

$$RG(t) = R(t) - G(t) \qquad (9)$$

$$BY(t) = B(t) - Y(t) \qquad (10)$$

The intensity channel and motion channel are calculated by Eq.11 and 12.

$$I(t) = (r(t) + g(t) + b(t))/3 \qquad (11)$$

$$M(t) = |I(t) - I(t - \tau)| \qquad (12)$$

where τ is the latency coefficient. Usually we set $\tau = 3$.

In sum, we obtain four channels of the image: two color channels, one intensity channel and one motion channel. So the image can be represented as a quaternion image $q(t)$ shown as follows (Eq.13):

$$q(t) = M(t) + RG(t)\mu_1 + BY(t)\mu_2 + I(t)\mu_3 \qquad (13)$$

where $\mu_i, i = 1, 2, 3$ satisfies $\mu_i^2 = -1$, $\mu_1 \perp \mu_2$, $\mu_2 \perp \mu_3$, $\mu_1 \perp \mu_3$, $\mu_3 = \mu_1\mu_2$.

We represent $q(t)$ in *symplectic* form:

$$q(t) = f_1(t) + f_2(t)\mu_2 \qquad (14)$$

$$f_1(t) = M(t) + RG(t)\mu_1 \qquad (15)$$

$$f_2(t) = BY(t) + I(t)\mu_1 \qquad (16)$$

3.2. Obtain Spatio-temporal Saliency Map by Quaternion Fourier Transform

Quaternion Fourier Transform (QFT) was first applied to color images by Ell and Sangwine [11]. The QFT of a quaternion image $q(n, m, t)$ can be written as:

$$Q[u, v] = F_1[u, v] + F_2[u, v]\mu_2 \qquad (17)$$

$$F_i[u, v] = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu_1 2\pi((mv/M)+(nu/N))} f_i(n, m) \qquad (18)$$

where (n, m) and (u, v) are the locations of each pixel in time and frequency domain, respectively. N and M are the image's height and width. $f_i, i \in \{1, 2\}$ is obtained by Eq.14 - 16. t is omitted for simplicity.

The inverse form of Eq.18 is obtained by changing the sign of the exponential and summing over u and v, instead of n and m. The inverse quaternion Fourier transform can be described as follows:

$$f_i(n,m) = \frac{1}{\sqrt{MN}} \sum_{v=0}^{M-1} \sum_{u=0}^{N-1} e^{\mu_1 2\pi((mv/M)+(nu/N))} F_i[u,v] \tag{19}$$

We use Eq.14 - 18 to obtain frequency domain representation $Q(t)$ of $q(t)$. $Q(t)$ can be represented in polar form as:

$$Q(t) = \|Q(t)\| e^{\mu\Phi(t)} \tag{20}$$

where $\Phi(t)$ is the phase spectrum of $Q(t)$ and μ is a unit pure quaternion.

Set $\|Q(t)\| = 1$, and then $Q(t)$ only contains the phase spectrum in frequency domain. Then we use Eq.19 to calculate the reconstruction of $Q(t)$ as $q'(t)$, which can be expressed as follows:

$$q'(t) = a(t) + b(t)\mu_1 + c(t)\mu_2 + d(t)\mu_3 \tag{21}$$

Our spatio-temporal saliency map is obtained by Eq.22.

$$sM(t) = g * \|q'(t)\|^2 \tag{22}$$

where g is a 2D gaussian filter ($\sigma = 8$).

The spatio-temporal saliency map using PQFT considers the features such as motion, color, intensity and orientation mentioned in literature. These features are represented as a quaternion image, which means that they are processed in a parallel way. Thus, it saves a lot of computational costs and is fast enough to meet real-time requirements. We can show in Section 4 that PQFT is better in performance than other models. Moreover, PQFT is independent of parameters and prior knowledge like PFT and SR.

The spatio-temporal saliency map can also deal with static natural images by setting motion channel $M(t)$ to zero.

4. Experimental Results

To evaluate the performance of our approach, four kinds of experiments are designed to compare our PQFT and PFT with SR, NVT and STB.

We set the saliency maps' resolution of PQFT, PFT and SR to 64×64 in all the experiments. The resolution of NVT and STB's saliency maps is adjusted by the programs themselves. For NVT and STB, we use the default parameters.

All the tests were run at MATLAB 2007a on Linux platform. The PC is equipped with P4 3G and 1G Memory. Please note that NVT is a C implementation and all the others are implemented by MATLAB.

4.1. How to evaluate the saliency maps?

Two aspects should be considered to evaluate the performance of saliency maps. One aspect is the computational cost. The other is the number of the correct objects detected in the images or videos, because saliency maps provide the locations of salient object candidates. Many approaches have been introduced to extract the objects or focus on the objects by the saliency map [6, 8, 10]. In order to give a fair result, let NVT and STB use their mechanisms to find the objects. As for PQFT, PFT and SR, the first n largest output in the saliency map is denoted as O_i^{max}, where $i = 1 \cdots n$ and (x_i, y_i) is the location of the i^{th} Focus of Attention (FoA). The i^{th} object candidate area can be obtained by Eq.23 and 24:

$$Mask_i = \{(x,y) | \alpha \cdot O_i^{max} \leq O(x,y) \leq O_i^{max}\} \tag{23}$$

$$Rgn_i = findArea(Mask_i, (x_i, y_i)) \tag{24}$$

where α is the threshold to affect the size of region. The smaller the α is, the coarser the selected area is. In all the experiments of this paper, we set $\alpha = 0.75$. The *findArea* function is to find the 8-connected neighborhood of (x_i, y_i) in $Mask_i$.

The candidates of correct objects are taken from a voting strategy in test videos and images labeled by unaffiliated volunteers (Here we use 4 persons). Only those labels that the majority agrees on are considered to be ”correct objects”. In the experiments, every model is allowed to select the first five fixations in the input image. If the model finds the objects which agree with the ”correct objects”, it is considered as a successful search. Otherwise, it will be considered as a failure. The number of correct objects detected is not the only criterion to evaluate the quality of the saliency map. The selection order is another important aspect, which can distinguish the quality of the saliency maps if they find the same number of objects. The less fixations a saliency map needs, the better it is.

4.2. Video Sequences

We use a video (988 images in total) captured at 15f/s with the resolution of 640×480 to test the performance of each methods. Fig.3 shows the selection results and orders of five methods in six frames. PQFT will select the salient people in the center of the frame at first, while other methods paid attention to the less salient trees or buildings. The test results of all frames are shown in Table 2 (Fig.4 is the bar view of Table 2). The performance of PQFT is the best because it can detect 2.52 objects *per* frame and can always select the salient objects in the frames within the first four fixations (see Table 2).

Table 3 shows the average time to calculate the saliency map *per* frame for the five methods. PFT approach is the

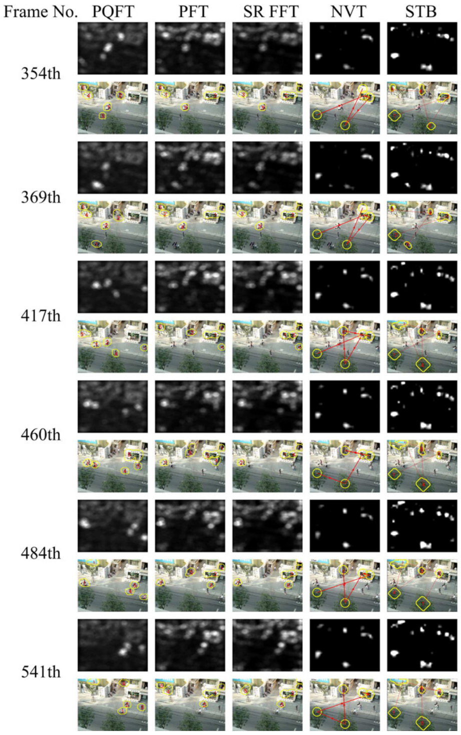

Figure 3. Selection results of five models in the test video.

Model	1st	2nd	3rd	4th	5th	ANODF
PQFT	921	704	405	274	182	2.52
PFT	132	349	283	260	262	1.30
SR	142	208	229	234	218	1.04
NVT	24	56	70	91	75	0.32
STB	138	58	58	58	63	0.38

Table 2. The number of correct objects detected at each fixation in the test video. Note that ANODF represents the Average Number of Object Detected *per* Frame.

fastest and PQFT ranks third in speed among the five methods, but it can surely meet real-time requirements.

Please note that this experiment is designed only to show the advantage of our PQFT approach to extract salient objects in the video because it considers motion feature between frames. Other models do not have such capacity.

4.3. Natural Images

To fairly test the five methods, 100 natural images with resolution around 800×600 are used as a test set, which are also used in [6, 10]. Fig.6 and Table 4 show the number of correct objects detected within the first five fixations by five methods respectively, and it is obvious that our PQFT can select more salient objects in these images and use fewer fixations than the other methods. Fig.5 gives the selection results and orders of five models in four natural images, which shows that our spatio-temporal saliency map can detect the salient animals, people and castle at the first fixation and find more interesting objects in each scene than other models. However, other models can only find a part of these objects and need more fixations to detect. Please note that PFT performs a little better than SR but saves about $1/3$ computational cost, which suggests that SR may not be necessary to calculate the saliency map (Table 4 and 5).

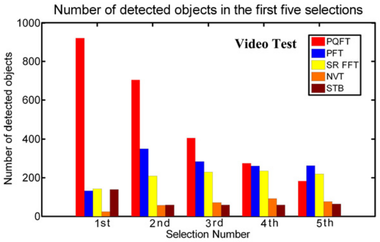

Figure 4. Number of correct objects detected within the first five fixations in the test video.

Model	Average Time Cost (s)
PQFT	0.0565
PFT	0.0106
SR	0.0141
NVT	0.4313
STB	3.5337

Table 3. Average time cost *per* frame in the test video.

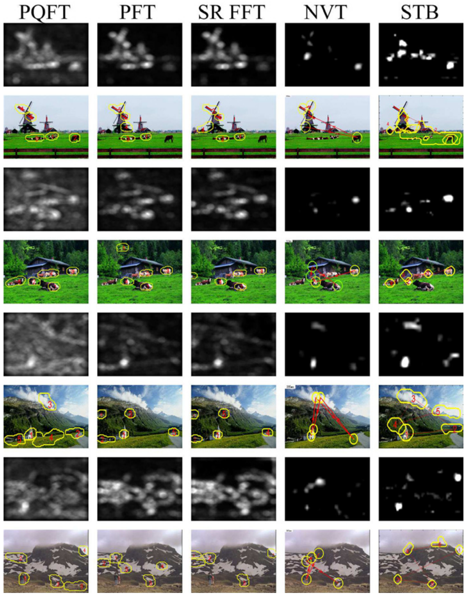

Figure 5. Selection results of five models in four natural images.

Model	1st	2nd	3rd	4th	5th	Total
PQFT	88	55	30	23	11	207
PFT	79	49	27	19	22	196
SR	73	43	32	21	18	187
NVT	81	43	23	19	10	176
STB	70	48	27	9	10	164

Table 4. The number of correct objects detected at each fixation in the database of 100 natural images.

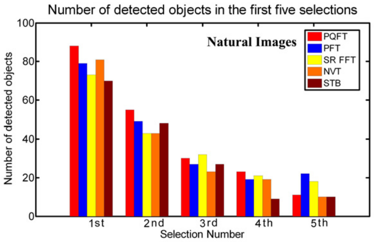

Figure 6. Number of correct objects detected within the first five fixations in the database of 100 natural images.

Model	Average Time Cost (s)
PQFT	0.0597
PFT	0.0099
SR	0.0159
NVT	0.7440
STB	4.7395

Table 5. Average time cost *per* image in the database of 100 natural images.

4.4. Natural Images with White-colored Noises

In this experiment, we hope to test the performance of five models when the images are stained by white-colored noise. We use the image [6], in which two people stand in front of a snow-covered mountain (as shown in the last row of Fig.5). Two kinds of noise patches (3×3 and 5×5) are randomly put into the test image with the intensity σ ranging from 0.1 to 0.8. Fig. 7 only gives the results when the 5×5 noise patch's intensity σ ranges from 0.1 to 0.4. Fig 8 shows the number of steps needed by PQFT, NVT and STB to attend to the salient people in all these noisy images. PFT and SR are not shown because they are very sensitive to the noise and cannot attend to the salient people at all (see Fig.7 for details). PQFT works best because it needs only one fixation to focus on the salient object under any circumstance. Other models need more fixations and even meet failures in some cases. The results show that our PQFT is very robust against noise when properties of the noise (such as color) do not conflict with the main feature of the target.

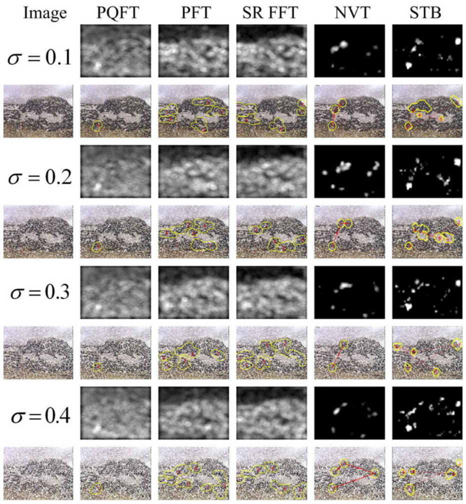

Figure 7. Selection results of the test Image with 5×5 noise patch whose density σ ranges from 0.1 to 0.4.

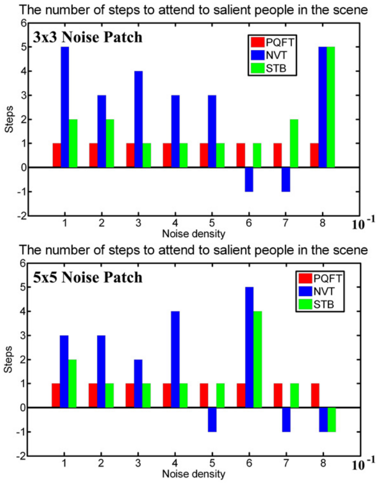

Figure 8. The number of steps to attend to salient people in the image with 3×3 (up) and 5×5 (down) noise patch. $Step = -1$ means the failure to attend the object within the first five fixations.

4.5. Psychological Patterns

Psychological patterns are widely used in attention experiments not only to explore the mechanism of visual search but also to test the effectiveness of saliency map [4, 5]. We used 13 patterns to test the models and the results are shown in Fig.9,10, 11 and 12.

In Fig.9, the first image is a salient color pattern. Our PQFT successfully find the red bar at the first fixation, but other models fails to find the target. The second and third image are salient orientation patterns, PQFT, PFT and SR find the targets immediately, NVT finds them but needs three fixations to attend to the horizonal pattern. STB fails to find them all. The fourth and fifth images are the patterns that are both salient in color and orientation, which should be the easiest task. PQFT and NVT can find the targets by only one fixation. PFT and STB both find the salient red vertical bar but fail to detect the salient red horizonal bar. SR fails to find any of the salient targets above.

In Fig.10, NVT and STB can find all the patterns within five fixations. Please note that they are capable of finding the target in closure pattern because of the denser intensity of the enclosed pieces, but not because of enclosure. PQFT, PFT and SR fail in the closure search, which is one common limitation of these methods.

Our PQFT and PFT can attend to the missing vertical black bar at the first fixation in Fig. 11, which agrees with human behavior. However, other methods fail in this test. Please note that the saliency maps by PFT and SR are quite different in this case although they look very similar in other tests.

Fig.12 shows that all the models cannot perform conjunction search effectively because it is believed that conjunction search needs thinking and prior knowledge (top-down) and all these models only considers bottom-up information.

In sum, our PQFT method performs best because it encounters only three detection failures (one in closure pattern and two in conjunction search) and needs only one fixation to detect the targets among all the other patterns, which shows that our spatio-temporal saliency map provides effective information to detect salient visual patterns.

5. Conclusions and Discussions

We proposed a method called PQFT to calculate spatio-temporal saliency maps which is used to detect salient objects in both natural images and videos. [10] discovered SR and used it to detect *proto objects*; however, our work indicates that the saliency map can be easily obtained when the amplitude spectrum of the image is at any nonzero constant value. Thus the phase spectrum is critical to the saliency map and the experimental results show that PFT is better and faster than SR, especially in the test of psychological patterns (Fig.11). As SR still preserves the phase spectrum, we doubt whether SR is a necessary step or not.

The effect of the phase spectrum provides us with a very easy way to extend PFT to PQFT which considers not only color, orientation and intensity but also the motion feature between frames. We incorporate these features as a quaternion image and process them in parallel, which is better than processing each feature separately, because some features can not ”pop out” if they are projected into each dimension. As a result, the spatio-temporal saliency map that PQFT produces can deal with videos, natural images and psychological patterns better than the other state-of-the-art models, which shows that PQFT is an effective saliency detection method. In addition, it is very robust against white-colored noise and is fast enough to meet real-time requirements. Our PQFT is independent of parameters and prior knowledge as well. Considering its good performance, is it possible that human vision system does the same function like PQFT or PFT? We hope that more work can be done to discover the role of phase spectrum in early human vision.

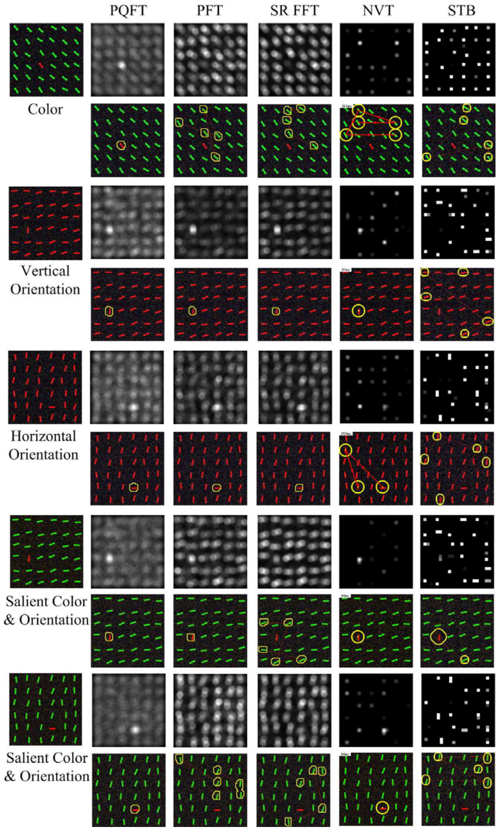

Figure 9. Search results of salient color or orientation patterns.

Comparing with human vision, our methods still have some limitations. Firstly, our methods can not deal with the closure pattern well up to now; however, human can find these patterns in a very short time. Secondly, our experimental results show the strong robustness of PQFT against white-colored noise; however, if the noise is very similar to the salient feature of the target, our spatio-temporal saliency map will fail to detect the target. Finally, Wang *et al*. suggested that people could perform effective conjunction search [14], and our experimental results showed that all the models including our own failed in these patterns. We will do more work to explore these unsolved issues.

The potential of our work lies in the engineering fields, which can be extended to the application like object recognition, video coding and *etc*. In addition, as our model only considers bottom-up information, it is necessary to add top-

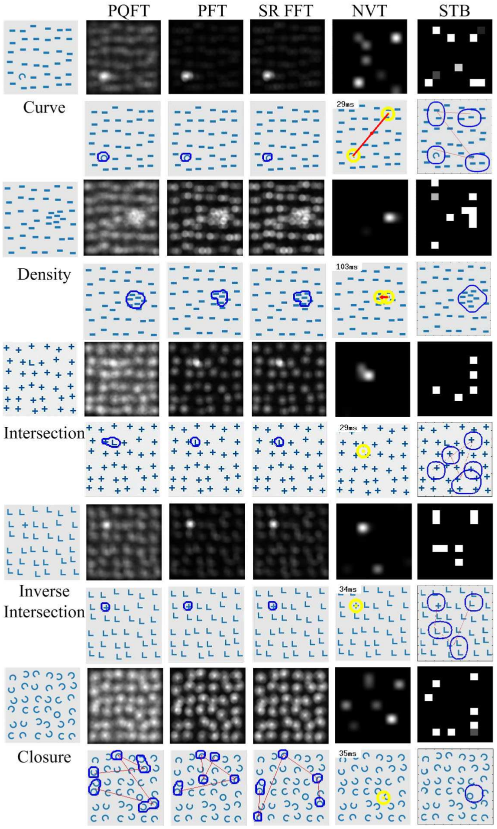

Figure 10. Search results of salient orientation patterns.

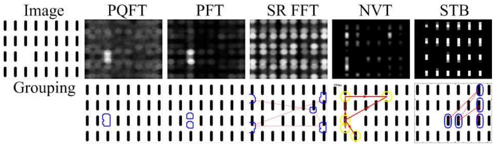

Figure 11. Only PQFT and PFT can find the missing item.

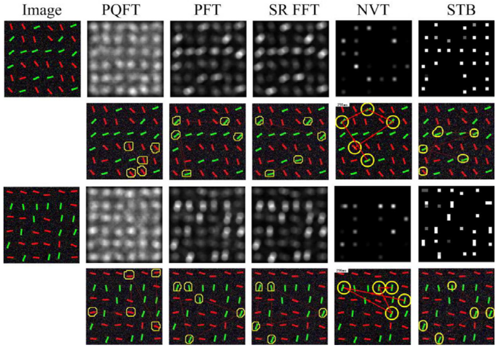

Figure 12. Conjunction search result of five models.

down signals (*e.g.* visual memory) for developing an effective vision system in robot application.

References

[1] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *Proc. CVPR*, 2, 2003. 1

[2] T. Liu, J. Sun, N. Zheng, X. Tang and H. Shum Learning to Detect A Salient Object. *Proc. CVPR*, 2007. 1

[3] D. Gao and N. Vasconcelos Integrated learning of saliency, complex features, and objection detectors from cluttered scenes. *Proc. CVPR*, 2005. 1

[4] A. Treisman and G. Gelade. A Feature-Integration Theory of Attention. *Cognitive Psychology*, 12(1):97-136, 1980. 1, 6

[5] J. Wolfe. Guided Search 2.0: A Revised Model of Guided Search. *Psychonomic Bulletin & Review*, 1(2):202-238, 1994. 1, 6

[6] L. Itti, C. Koch, E. Niebur, et al. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254-1259, 1998. 1, 3, 4, 5, 6

[7] R. Rensink. Seeing, sensing, and scrutinizing. *Vision Research*, 40(10-12):1469-87, 2000. 1

[8] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*. 19, 1395-1407, 2006. 1, 3, 4

[9] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional Selection for Object Recognition - a Gentle Way. *Lecture Notes in Computer Science*, 2525(1):472-479, 2002. 1

[10] X. Hou and L. Zhang. Saliency Detection: A Spectral Residual Approach. *Proc. CVPR*, 2007. 1, 2, 3, 4, 5, 7

[11] T. Ell and S. Sangwin. Hypercomplex Fourier Transforms of Color Images. *IEEE Transactions on Image Processing*, 16(1):22-35, 2007. 1, 3

[12] K. Castleman. Digital Image Processing. Prentice-Hall, New York, 1996. 2

[13] S. Engel, X. Zhang, and B. Wandell. Colour Tuning in Human Visual Cortex Measured With Functional Magnetic Resonance Imaging. *Nature*, vol.388, no.6,637, pp.68-71, July 1997. 3

[14] D.L. Wang, A. Kristjansson, and K. Nakayama. Efficient visual search without top-down or bottom-up guidance. *Perception & Psychophysics*, vol. 67, pp. 239-253. 7