A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression

Chenlei Guo, *Student Member, IEEE*, and Liming Zhang, *Senior Member, IEEE*

***Abstract*—Salient areas in natural scenes are generally regarded as areas which the human eye will typically focus on, and finding these areas is the key step 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 others, but they demand high computational cost and computing useful results mostly relies on their choice of parameters. Although some region-based approaches were proposed to reduce the computational complexity of feature maps, these approaches still were not able to work in real time. Recently, a simple and fast approach called spectral residual (SR) was proposed, which uses the SR of the amplitude spectrum to calculate the image's saliency map. However, in our previous work, we pointed out that it is the phase spectrum, not the amplitude spectrum, of an image's Fourier transform that is key to calculating the location of salient areas, and proposed the phase spectrum of Fourier transform (PFT) model. In this paper, we present a quaternion representation of an image which is composed of intensity, color, and motion features. Based on the principle of PFT, a novel multiresolution spatiotemporal saliency detection model called phase spectrum of quaternion Fourier transform (PQFT) is proposed in this paper to calculate the spatiotemporal saliency map of an image by its quaternion representation. Distinct from other models, the added motion dimension allows the phase spectrum to represent spatiotemporal saliency in order to perform attention selection not only for images but also for videos. In addition, the PQFT model can compute the saliency map of an image under various resolutions from coarse to fine. Therefore, the hierarchical selectivity (HS) framework based on the PQFT model is introduced here to construct the tree structure representation of an image. With the help of HS, a model called multiresolution wavelet domain foveation (MWDF) is proposed in this paper to improve coding efficiency in image and video compression. Extensive tests of videos, natural images, and psychological patterns show that the proposed PQFT model is more effective in saliency detection and can predict eye fixations better than other state-of-the-art models in previous literature. Moreover, our model requires low computational cost and, therefore, can work in real time. Additional experiments on image and video compression show that the HS-MWDF model can achieve higher compression rate than the traditional model.**

Manuscript received June 11, 2008; revised July 26, 2009. First published August 25, 2009; current version published December 16, 2009. This work is supported by the National Nature Science Foundation of China (60571052) and Shanghai Leading Academic Discipline Project (B112). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jenq-Neng Hwang.

C. Guo was with the Department of Electronic Engineering, Fudan University, Shanghai, 200433, China. He is now with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, 15213, USA (e-mail: cguo@andrew.cmu.edu).

L. Zhang is with the Department of Electronic Engineering, Fudan University, Shanghai 200433, China (e-mail: lmzhang@fudan.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2009.2030969

***Index Terms*—Hierarchical selectivity (HS), multiresolution wavelet domain foveation (MWDF), phase spectrum of Fourier transform (PFT), phase spectrum of quaternion Fourier transform (PQFT), receiver operating characteristic (ROC) curve, spatiotemporal saliency map, visual attention.**

I. Introduction

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

In 1890, James suggested that visual attention operates like a "spotlight" that can move around the visual field [3]. What attracts people's attention? Tresiman [4] proposed the famous feature integration theory (FIT) which described visual attention as having two stages. A set of basic visual features, such as color, motion and edges, is processed in parallel at the preattentive stage. And then, in the limited-capacity process stage, the visual cortex performs other more complex operations like face recognition and others [5]. A master map [4] or a saliency map [6] is computed to indicate the locations of salient areas. Distinctive features (e.g., luminous color, high velocity motion, and others) will "pop out" automatically in the preattentive stage, and then the salient areas become the object candidates.

Several computational models have been proposed to simulate human's visual attention, which are based on the bottom-up computational framework [7]–[23]. Itti *et al.* proposed a bottom-up model and built a system called Neuromorphic Vision C++ Toolkit (NVT) [7]. After that, following Rensink's theory [24], Walther extended this model to attend to *proto object* regions and created SaliencyToolBox (STB) [12]. He also applied the model to accomplish object recognition tasks [13]. However, the high computational cost and the choice of parameters are still the weaknesses of these models. Recently, the spectral residual (SR) approach based on Fourier Transform was proposed by [14], which does not rely on the parameters and can detect salient objects rapidly. Later, Guo *et al.* [15] manifested the fact that the phase spectrum is key to calculating the saliency map and proposed a model called phase spectrum of Fourier transform (PFT) for saliency detection. Besides these models, Bruce *et al.* proposed a model of bottom-up overt attention based on the principle of maximizing information sampled from a scene [16]. In 2004, Gao *et al.* presented

the discriminant saliency detection model which requires a discriminant saliency selection process at first (training stage), and then the saliency map can be computed by the selected features at the testing stage [17]. A graph-based visual saliency detection was proposed in 2006 [18], which can powerfully predict human fixations but demands very high computational cost; [19] and [20] proposed the region-based approaches to calculate the feature maps for their saliency models, which perform a clustering at first and compute the feature maps by these clusters to reduce computational complexity; however, their models need to set many parameters to obtain useful results, and still can not work in real time (only a few frames per second).

All these models mentioned above, however, only consider static images. Some work has been done to add the motion feature to these models (e.g., [17] and [25]) in order to perform some applications; however, the additional motion channel will increase the computational cost of the model. Incorporating motion into a saliency model without dramatically influencing its computational cost is a challenging task that motivates us to develop a novel model to generate spatiotemporal saliency map. Moreover, besides SR and our PFT model, other models require tremendous computational cost and cannot meet real-time requirements on a personal computer (PC). Therefore, how to develop a saliency model that can work in real time is another consideration of our work.

In this paper, we propose a novel quaternion representation of an image and developed a multiresolution spatiotemporal saliency detection model called phase spectrum of quaternion Fourier transform (PQFT) to compute the spatiotemporal saliency map from the image's quaternion representation. Each pixel of the image is represented by a quaternion that consists of color, intensity and motion feature. The phase spectrum of QFT is used to calculate the spatiotemporal saliency map, which considers not only salient spatial features like color, orientation and others in a single frame but also temporal feature between frames like motion. Our PQFT model is independent of prior knowledge and parameters, and experimental results show that it is fast enough to meet real-time requirements (less than 1 ms per frame in C/C++ implementation), and outperforms other state-of-the-art models in computer vision. Since the proposed PQFT model can process the image under various resolutions, we introduce the hierarchical selectivity (HS) framework based on the PQFT model, which can be used to construct a tree structure representation of an image. A novel multiresolution wavelet domain foveation (MWDF) model is presented based on this tree structure, which can improve coding efficiency in image and video compression.

The rest of this paper is organized as follows. In Section II, we introduce the quaternion representation of an image and propose a model called PQFT to calculate spatiotemporal saliency map of the image. Section III shows that our PQFT can work under various resolutions, and the Hierarchical Selectivity (HS) framework based on PQFT is presented to construct a tree structure representation of an image. Many experimental results of comparing the PQFT model with other saliency models are shown in Section IV. A MWDF model based on HS-PQFT is proposed, and its performance in image and video compression is demonstrated in Section V. The conclusions and discussions are given thereafter.

II. From Phase Spectrum to Spatiotemporal Saliency Map

In [26], Castlman pointed out that the amplitude spectrum of an image specifies how much of each sinusoidal component is present and the phase information specifies where each of the sinusoidal components resides within the image. Thus, the phase spectrum of an image represents the local information in the image. Our research in [15] showed that the locations with less periodicity or less homogeneity in an image create the "pop out" *proto objects* in the reconstruction of the image's phase spectrum, which indicates where the object candidates are located. Therefore, we proposed a saliency detection model for gray-scale image called PFT, which showed that the saliency map can be easily calculated by the phase spectrum of an image's Fourier transform when its amplitude spectrum is at nonzero constant value.

The PFT model provides the simplest and fastest way in literature to calculate the saliency map from the input image. However, it is only designed for gray-scale images and does not take into account some important cues in human early vision (e.g., color and motion), which inspires us to develop a novel quaternion representation of an image and utilize the Quaternion Fourier Transform to calculate the image's spatiotemporal saliency map. Compared with other saliency maps in [7], [8], [12], [14], [16], [17], and [18], our PQFT model considers the motion feature between sequential frames.

A. Novel Quaternion Representation of an Image

Define the input image captured at time t as $F(t)$, $t = 1, 2, \ldots, T$, where T is the total frame number. $r(t), g(t)$ and $b(t)$ are the red, green and blue channel of $F(t)$. Four broadly tuned color tunnels are created by (1) – (4) [7]

$$R(t) = r(t) - \frac{(g(t) + b(t))}{2} \tag{1}$$

$$G(t) = g(t) - \frac{(r(t) + b(t))}{2} \tag{2}$$

$$B(t) = b(t) - \frac{(r(t) + g(t))}{2} \tag{3}$$

$$Y(t) = \frac{(r(t) + g(t))}{2} - \frac{|r(t) - g(t)|}{2} - b(t). \tag{4}$$

In the human brain, there exists a "color opponent-component" system. In the center of receptive fields, neurons which are excited by one color (e.g., red) are inhibited by another color (e.g., green). Red/green, green/red, blue/yellow, and yellow/blue are color pairs which exists in human visual cortex [27]. Therefore, the color channels are designed as follows:

$$RG(t) = R(t) - G(t) \tag{5}$$

$$BY(t) = B(t) - Y(t). \tag{6}$$

The intensity channel and motion channel are calculated by (7) and (8)

$$I(t) = \frac{(r(t) + g(t) + b(t))}{3} \tag{7}$$

$$M(t) = |I(t) - I(t - \tau)| \tag{8}$$

where τ is the user-defined latency coefficient.

In sum, we have built four feature channels for the input image: two color channels, one intensity channel, and one motion channel. In the primary visual cortex, different cells (e.g., color opponent cells, intensity cells, and motion cells) can detect color, intensity and motion features at their receptive fields respectively, which are similar to the four channels designed by us. In addition, in human vision, these channels are almost independent. Therefore, the image can be represented as a quaternion image $q(t)$ shown as follows:

$$q(t) = M(t) + RG(t)\mu_1 + BY(t)\mu_2 + I(t)\mu_3 \tag{9}$$

where μ_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:

$$q(t) = f_1(t) + f_2(t)\mu_2 \tag{10}$$

$$f_1(t) = M(t) + RG(t)\mu_1 \tag{11}$$

$$f_2(t) = BY(t) + I(t)\mu_1. \tag{12}$$

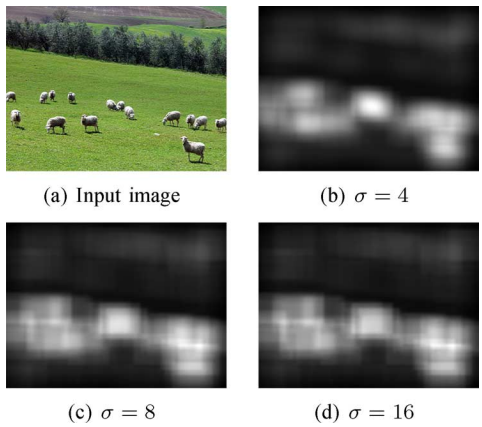
B. Calculate the Spatiotemporal Saliency Map by Quaternion Fourier Transform

The first application of Quaternion Fourier Transform to color images was published in [28], which used a discrete version of Ell's transform [29], [30]. Many image processing techniques that are dependent on Fourier transforms are applicable to hypercomplex Fourier transform according to its definitions. Later, Pei *et al.* proposed an efficient implementation of QFT [31] by using the transform defined in [32]. Since there was no definition of a Fourier transform which was applicable to color images in a holistic manner, Ell and Sangwine proposed Quaternion Fourier Transform of color images and developed the properties of the transform [33]. In addition, their study showed that the transform can be computed by using two standard complex fast Fourier transforms (FFT). Ell's work provides us with a way to utilize the quaternion representation of the image to calculate the spatiotemporal saliency map.

In (9), the quaternion representation of an image is calculated and each pixel is denoted as $q(n, m, t)$, where (n, m) is the location of each pixel and t is the time point. Thus, its QFT representation can be computed as follows:

$$Q[u, v] = F_1[u, v] + F_2[u, v]\mu_2 \tag{13}$$

$$F_i[u, v] = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu_1 2\pi((mv/M)+(nu/N))} f_i(n, m) \tag{14}$$



(a) Input image (b) $\sigma = 4$ (c) $\sigma = 8$ (d) $\sigma = 16$

Fig. 1. Saliency maps calculated by the PQFT model with various σ for the gaussian filter. σ does not affect the result much; therefore, there is a wide range to pick up a good σ. The resolution of the saliency maps is 64×64, and the saliency maps are rescaled to the size of input image (800×600) for illustration.

where (n, m) and (u, v) are the locations of each pixel in spatial and frequency domain, respectively. N and M are the image's height and width. $f_i, i \in \{1, 2\}$ is calculated by (11) and (12). Here t is omitted for simplicity.

The inverse form of (14) is calculated 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:

$$f_i(n, m) = \frac{1}{\sqrt{MN}} \sum_{v=0}^{M-1} \sum_{u=0}^{N-1} e^{\mu_1 2\pi((mv/M)+(nu/N))} F_i[u, v]. \tag{15}$$

We use (10)–(14) to calculated frequency domain representation $Q(t)$ of $q(t)$. $Q(t)$ can be represented in polar form as

$$Q(t) = \|Q(t)\| e^{\mu\Phi(t)} \tag{16}$$

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 (15) to calculate the reconstruction of $Q(t)$ denoted as $q'(t)$, which can be expressed as follows:

$$q'(t) = \rho_0(t) + \rho_1(t)\mu_1 + \rho_2(t)\mu_2 + \rho_3(t)\mu_3. \tag{17}$$

Our spatiotemporal saliency map is computed by

$$sM(t) = g * \sum_{i=0}^{3} (w_i \rho_i^2(t)) \tag{18}$$

where g is a 2-D gaussian filter with variance as σ. When $4 \leq \sigma \leq 16$, the saliency maps after filtering are very similar (Fig. 1). Set $w_i = 1$, i =0, 1, 2, 3 and (18) can be expressed as follows:

$$sM(t) = g * \|q'(t)\|^2. \tag{19}$$

Changing the permutation of the various color and motion components or choosing various μ_i, i =1, 2, 3 will not affect the saliency map in (19) if μ_i satisfies the quaternion conditions in [33], because we only use the norm of QFT reconstruction result to compute the saliency map. However, when computed by (18) with various w_i, the saliency map will be different. In this paper, we will not investigate this case because we only use (19) to calculate the spatiotemporal saliency map. In addition, we do not take the borders of images into consideration because of their discontinuity.

The spatiotemporal saliency map computed by the PQFT model 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. Therefore, it saves a lot of computational costs, and experimental results in Section IV show that the PQFT model can work in real time. Moreover, the PQFT model is independent of parameters and prior knowledge like the PFT and SR models. The PQFT model can also deal with static natural images by setting the motion channel $M(t)$ to zero.

C. Time and Space Complexity Analysis for the PQFT Model

In order to calculate the spatiotemporal saliency map of an image/frame at resolution $M \times N$, the PQFT model needs $13MN$ real addition/substraction operations to calculate the quaternion image. In the next step, both 2-D QFT and inverse 2-D QFT are required, and, therefore, $4MN \cdot \log_2(MN)$ real multiplication operations are needed [31]. Moreover, calculating the phase spectrum will cost $(5 + p)MN$ real multiplication operations and $3MN$ real addition operations, where pMN real multiplication operations are equal to MN real square root operations and p is decided by number of the iteration steps needed for the convergence of the real square root result. In the final step, computing the saliency map from the recovered quaternion representation requires $3MN$ real addition operations and $4MN$ real multiplication operations. In addition, the 2-D Gaussian filter demands k^2MN real multiplication operations and $k^2 - 1$ real addition operations, where k is the kernel width of the $k \times k$ gaussian kernel.

In sum, the total time complexity for the PQFT model to calculate the spatiotemporal saliency map is $O(MN\log_2(MN) + \kappa MN)$, where κ is the total number of required real addition/multiplication operations. If MN is small, $O(\kappa MN)$ will dominate the total time cost, and if MN is sufficient large, $O(MN\log_2(MN))$ will dominate.

The maximum space that the PQFT model needs is $4MN$ due to the storage of the quaternion representation of the image; therefore, the space complexity is $O(MN)$.

In Section IV, we will show that the PQFT model needs less than 1 ms to compute the saliency map for one image/frame in C/C++ implementation.

III. Hierarchical Selectivity Based on the PQFT Model

In this section, we will present a method to extract salient objects from the spatiotemporal saliency map at a fixed resolution. And then a discussion of how our model works under various resolutions is presented. Finally, the hierarchical selectivity (HS) framework based on the PQFT model is proposed to construct the tree structure representation of an image, which can be used to improve coding efficiency in image and video compression.

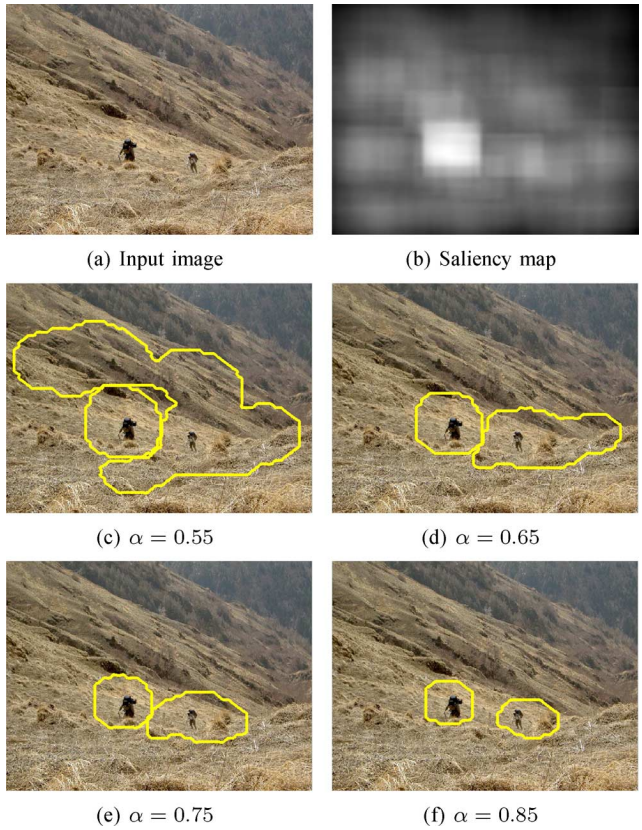


Fig. 2. Object candidate areas (OCAs) computed from the saliency map with various α. Note that the original size of the saliency map is 64×64, and it is rescaled to the size of input image (800×600) for illustration.

A. How to Detect Proto-Objects in the Spatiotemporal Saliency Map

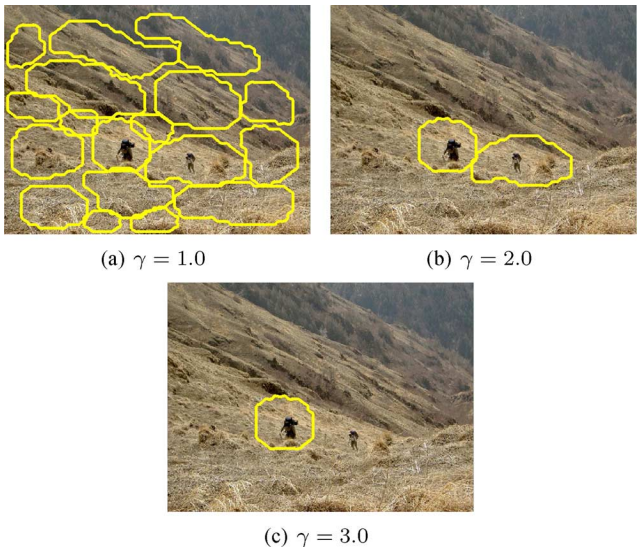
A saliency map provides information about where the proto-objects locate in an image. Therefore, the method to extract the proto-objects from a saliency map is very important. Here a method to find the i^{th} focus of attention (FoA) in our spatiotemporal saliency map is proposed and described as follows.

Suppose that the saliency map $sM(t)$ of an image $F(t)$ is calculated at time t. The i^{th} search begins from $sm_i(t)$ and $sm_1(t) = sM(t)$. Find the largest output in $sm_i(t)$, which is denoted as $O_i^{\max}$, and (x_i, y_i) is the location of $O_i^{\max}$. The i^{th} object candidate area (OCA_i) can be calculated as follows:

$$\text{Mask}_i = \{(x,y) | \alpha \cdot O_i^{\max} \le O(x,y) \le O_i^{\max}\} \quad (20)$$

$$OCA_i = \text{findArea}(\text{Mask}_i, (x_i, y_i)) \quad (21)$$

where α is the user-defined threshold to affect the size of selected region. For example, Fig. 2(a) and (b) are the input image and its corresponding saliency map at 64×64 resolution. When $\alpha = 0.55$, not only the salient objects (human beings) but also the grass around the objects are selected as salient regions



(a) $\gamma = 1.0$ (b) $\gamma = 2.0$ (c) $\gamma = 3.0$

Fig. 3. Selected candidate area under various γ. (a) When γ is too small, lots of useless areas (e.g., grass areas) are selected. (b) A suitable γ will lead to a good selection result. (c) When γ is too big, the useful salient area (e.g., human being) will be omitted. Note that $\alpha = 0.75$.

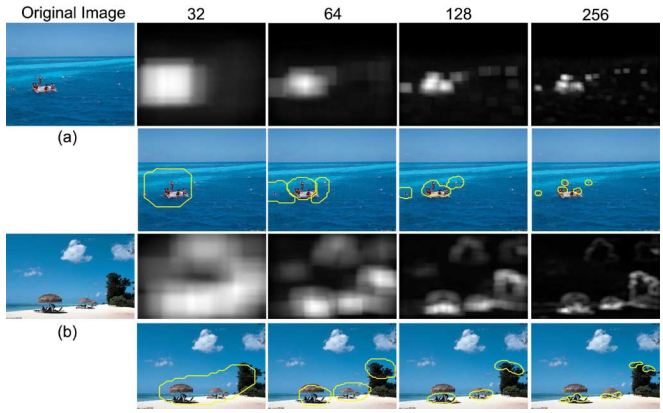


Fig. 4. Examples of the saliency maps which are computed by the PQFT model under various resolutions. Four resolutions are used to compute the saliency maps whose size is 32×32, 64×64, 128×128, 256×256. The 32×32 saliency map is the coarsest which considers a group of people (a) or summerhouses (b) as one object. However, at the finest resolution, single person in the sea (a) and the people in the summer house (b) can be detected according to the saliency map. Note that the saliency maps are rescaled to the size of input images (800×600). The images are available online at [56].

[Fig. 2(c)]. When α becomes larger and larger, less grass regions are selected and the salient areas turn smaller and smaller [Fig. 2(d)–(f)]. By lots of experiments, we find that selected regions are reasonable when $0.65 \leq \alpha \leq 0.85$. The *findArea* function is to find the 8-connected neighborhood region of (x_i, y_i) in $Mask_i$. The shape of OCA_i is decided by (20) and (21).

When OCA_i is found, the output of its area in the saliency map will be suppressed to zero which results in $sm_{i+1}(t)$. Thus, the $(i+1)^{th}$ object candidate area can be computed from the unattended area in the new saliency map. Such "inhibition of return" (IoR) was demonstrated in human visual psychophysics [35] and was used by visual attention model [7]. The search does not stop until (22) is satisfied

$$O_i^{\max} \leq \gamma \cdot E(sM(t)) \tag{22}$$

where γ is a user-defined acceptance rate and $E(\cdot)$ denotes expectation. Fig. 3 gives an example how γ affects the selection results. The input image and the saliency map are shown in Fig. 2(a) and (b). When γ is too small, lots of useless areas (e.g., grass areas) are selected [Fig. 3(a)]. However, when γ is too big, the useful salient area (e.g., the human being) will be omitted [Fig. 3(c)]. Through lots of experiments, we find that $\gamma = 2.0$ will lead to reasonable results [e.g., Fig. 3(b)].

B. How to Select the Visual Resolution?

The PQFT model can calculate the spatiotemporal saliency map under various resolutions just like human vision system (HVS). The selection of resolutions is realized by processing the input image with various sizes, which simulates the results of various view distances between the observer and the scene. Since the size of the saliency map is the same as that of the input image, a smaller size of saliency map represents a coarser resolution. Fig. 4 gives two examples of how the PQFT model works under various resolutions and the size of the saliency maps ranges from 32×32 to 256×256. Note that the saliency maps shown in Fig. 4 are rescaled to the size of input images. When the input image is processed at a coarser resolution (e.g., the size of the saliency map is 32×32), the detailed local features is depressed by global ones in the saliency map [a group of people in Fig. 4(a) or the whole summerhouses in Fig. 4(b) is regarded as only one object], just like a person looks at the scene from a long distance. However, at a finer resolution (e.g., 256×256), detailed local features are more competitive so that single person in the sea [Fig. 4(a)] or people in the summer house [Fig. 4(b)] will "pop-out" separately in the saliency map. This activity can be considered as a human looking at the objects at a fine resolution in a very careful manner.

C. Hierarchical Selectivity Based on the PQFT Model

Scholl suggested "there may be a hierarchy of units of attention, ranging from intraobject surfaces and parts to multiobject surfaces and perceptual groups" [34]. According to Scholl's theory, Sun *et al.* proposed the first HS implementation which shifts attention from one grouping to another one or from a parent grouping to its sub-grouping as well as implemented a model that can focus its attention from far to near or from coarse to fine [36]. However, their implementation relied on human subject interactions as top-down.

Motivated by Sun's model, we eliminate the interference from human subjects and propose a pure bottom-up Hierarchical Selectivity framework based on the PQFT model (PQFT-HS). The hierarchical level φ is defined in advance. In this paper, we set $\varphi = 3$ and use the saliency maps of three various resolutions from coarse to fine for one input image.

Suppose that the saliency maps sM_L, $L = 1, 2, 3$ at level L, whose size is 64×64, 128×128 and 256×256, are computed by the PQFT model. The search begins from level 1. Find the OCA_i^L from sM_L, where $i = 1 \ldots n_L$ and n_L is the total number of object candidates at level L. n_L is decided by (22), and then search continues within each OCA_i^L by its order in the saliency map sM_{L+1} at next level. The search stops when the search level $L = \varphi$.

Fig. 5. Attention shift process of the PQFT-HS model. The red (gray), yellow (white), and blue (black) rectangles represent the detected objects from the coarse-resolution saliency map to the finer-resolution ones. Note that the selection results are of erose shape decided by (20). Here, we use the rectangles to show the hierarchical selection results clearly. In addition, the size and location of the rectangles depend on the original selection results.

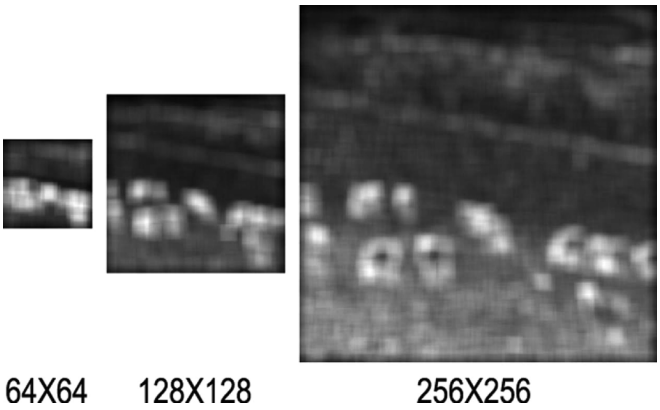


Fig. 6. Three saliency maps that are used in the hierarchical search of the image shown in Fig. 5. The size is 64 × 64 (coarse), 128 × 128 (fine), and 256 × 256 (finest), respectively.

Fig. 5 illustrates how the PQFT-HS model works. In the image, a lot of sheep are eating grass on a grassland. Three saliency maps are computed which are shown in Fig. 6. At the coarse resolution (64 × 64), the sheep are divided into two groups. Each group can not be separated and is considered as an object. Then the selection process continues within the two groups at a finer resolution, respectively. Finally each sheep can be detected at the finest resolution. Fig. 7 shows the tree structure representation of the image established by the PQFT-HS model. The arrows represents the search order. This representation is very useful in image and video compression, which will be discussed in Section VI.

IV. Experimental Results

To evaluate the performance of our approach, four kinds of experiments are designed to compare our proposed PQFT model and PFT models with the SR, NVT and STB models. The PQFT, PFT and SR models are frequency domain approaches and others are spatial domain ones.

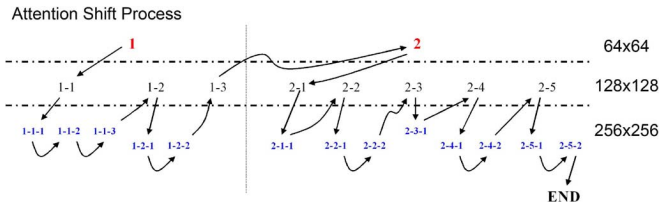


Fig. 7. Tree structure representation of the sheep image. The arrow represents the search order.

As visual selection is coarser than visual resolution [34], it is reasonable to set a constant coarse resolution for the images. The saliency maps' resolution of the PQFT, PFT and SR models is set to 64×64 in our data sets. The resolution of NVT and STB's saliency maps is adjusted by the programs themselves. The variance σ of the 2-D gaussian filter for the PQFT, PFT and SR models is set to 8. For NVT and STB, default parameters are applied. The user-defined latency coefficient τ of the motion channel for the PQFT model is set to 3.

All the tests were run on Linux platform. The PC is equipped with P4 3.0 G and 1-G Memory. NVT is implemented in C/C++, and STB is implemented in MATLAB. The PQFT, PFT and SR models are implemented in both C/C++ and MATLAB.

A. How to Evaluate the Saliency Map?

1) Extract Salient Objects in Videos and Images: The saliency map provides the locations of salient object candidates; therefore, an effective saliency map should provides as much information as possible. In the past, many approaches have been introduced to extract or focus on the proto-objects by the saliency map [7], [12], [14]. In order to give a fair result, let NVT and STB use their own proto-object extraction method to find the objects. As for the PQFT, PFT and SR models, we use the strategy described in Section III-A. Note that we do not use the user-defined threshold ($\gamma \cdot E(sM(t))$ in (22) to stop the visual search in order to make our results comparable to NVT [7] and STB [12].

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 fewer fixations a saliency map needs, the better it is.

Moreover, our research goal is to develop a spatiotemporal saliency detection model that can work in real time. Thus, the computational cost is another important metric to evaluate every model.

2) Predict the Human Eye Fixations: Receiver operating characteristic (ROC) curve is recently used to evaluate the saliency map's ability to predict human eye fixations in natural images [37], which is also used in [16], [17], and [18]. Given

a threshold value, A saliency map that is computed from a saliency model can be divided into the fixation region (target) and nonfixation region (background). All participants' eye fixations in the entire data set are combined in a binary map (the fixation point's value is 1 and the rest is 0). And then the map is convolved with a decaying gaussian kernel, which results in the subject saliency map ("ground truth"). Thus, those nonzero points in the subject saliency map are regarded as true target points and the rest as true background points. The percentage of true target points that fall into the fixation region given by a saliency model is true positive rate (TPR), and the percentage of true background points that fall into the fixation region is false positive rate (FPR). The ROC is the curve of TPR versus FPR by varying the threshold. The area beneath the curve is called ROC area. The larger the ROC area is, the better the prediction power of a saliency map is. The perfect prediction of an ROC area is 1 and chance performance occurs at an area of 0.5.

Considering the diversity of the fixation point from different subjects, another ROC metric called intersubject ROC curve is proposed by [18], which is to value the consistency of the subject saliency map with that from different saliency models.

In our experiments, both of the metrics are used to evaluate the prediction power of our saliency model in the natural images.

B. Video Sequence

We use a video (988 images in total) captured at 15 f/s with the resolution of 640×480 to evaluate the performance of each model by the total number of detected objects and the search order. Fig. 8 shows the selection results and orders of five models in six frames. The PQFT model selected the salient people in the center of the frame at first, while other models always paid attention to the less salient trees or buildings. The test results of all frames are shown in Table I. The performance of the PQFT model is the best because it can detect more objects in each frame than any other model (2.52 objects *per* frame in Table I).

Table II shows the average time to calculate the saliency map *per* frame by five models. The PFT model is the fastest and the PQFT model ranks third in speed among the five models. However, in the C/C++ implementation, the PQFT model needs less than 1 ms to process one single frame which is fast enough to work in real-time.

This experiment is designed only to show the advantage of our PQFT model to extract salient objects in the video because it considers the motion feature between frames. Other models do not have such capacity. Moreover, we do not provide the comparison with optical flow approaches [38], [39] for the following reasons. First, our aim is to quickly compute the saliency map (SM) in order to find the region of interest (RoI) that may contain the object candidates in videos or natural images. SM can help reduce search regions and save search time for object detection/recognition. Although optical flow can detect moving objects precisely, they fail to detect still RoI in videos or natural images. Second, The computational complexity of optical flow approaches is much higher (at least 1.2 s/frame in [39]) than ours (less than 1 ms/frame in average when implemented in C/C++). Therefore, it does not match our intent for real-time applications.

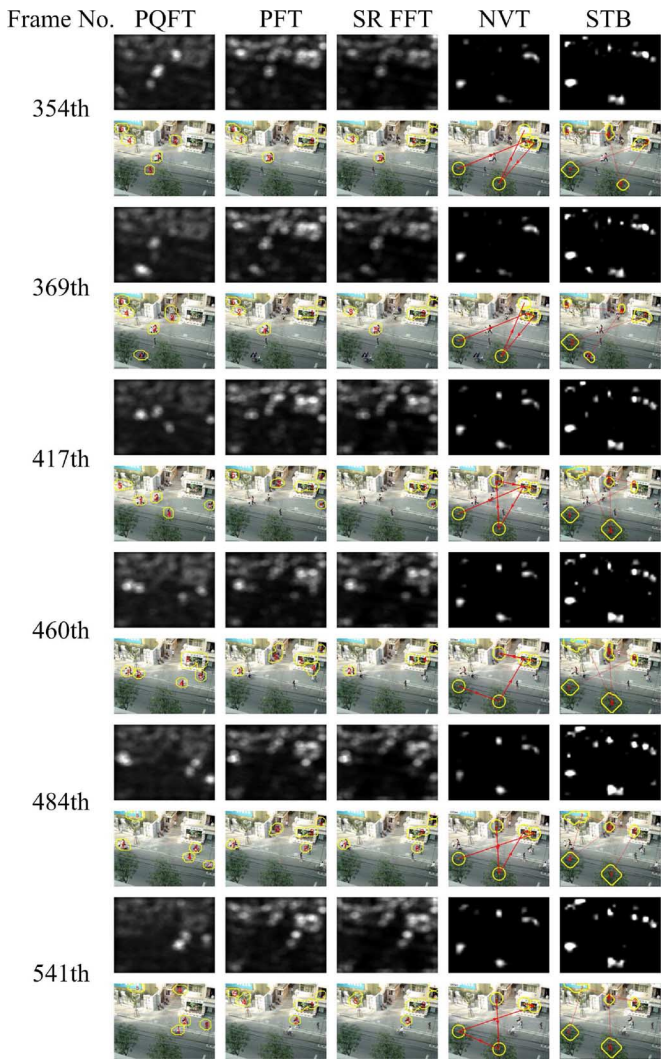


Fig. 8. Selection results and orders of five models in the test video. The images are available online at [56].

TABLE I
NUMBER OF CORRECT OBJECTS DETECTED AT EACH FIXATION 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

Note that ANODF represents the Average Number of Object Detected *per* Frame.

TABLE II
AVERAGE TIME COST *PER* FRAME IN THE TTEST VIDEO

Model	MATLAB (ms)	C/C++ (ms)
PQFT	56.5	< 1
PFT	10.6	< 1
SR	14.1	< 1
NVT	N/A	431.3
STB	3533.7	N/A

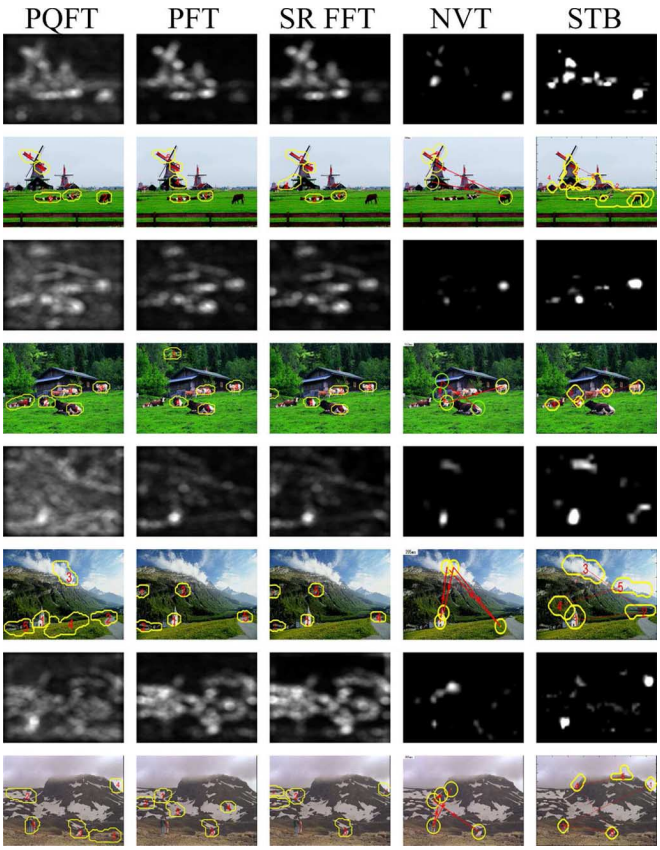


Fig. 9. Selection results and fixation orders of five models in four natural images. The images are available online at [56].

TABLE III
NUMBER OF CORRECT OBJECTS DETECTED AT EACH FIXATION IN THE DATA SET OF 100 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

C. Natural Images

To fairly test the five models, 100 natural images with resolution around 800×600 are used as a test set, which are also used in [7] and [14]. Table III show the number of correct objects detected within the first five fixations by five models respectively, and it is obvious that our PQFT model can select more objects in these images and use fewer fixations than the other models. Fig. 9 gives the selection results and orders of five models in four natural images, which shows that our spatiotemporal 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. Table IV shows the average time each model needs to process an image. Our PQFT model only need 59.7 ms in MATLAB and less than 1 ms in C/C++ to compute the saliency map of a natural image.

TABLE IV
AVERAGE TIME COST *PER* IMAGE IN THE DATA SET OF 100 NATURAL IMAGES

Model	MATLAB (ms)	C/C++ (ms)
PQFT	59.7	< 1
PFT	9.9	< 1
SR	15.9	< 1
NVT	N/A	744.0
STB	4739.5	N/A

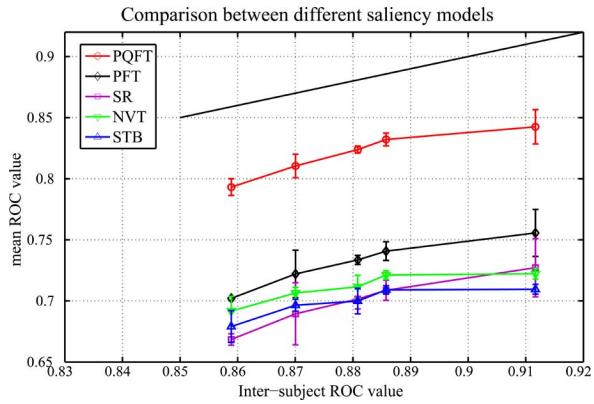


Fig. 10. Mean ROC value of five models computed for each range of intersubject ROC values in our data set.

TABLE V
ROC AREAS FOR DIFFERENT SALIENCY MODELS ACCORDING TO ALL HUMAN FIXATIONS IN OUR DATA SET

Model	PQFT	PFT	SR	NVT	STB
ROC area	0.8328	0.7433	0.6977	0.7508	0.7415

D. Prediction of Eye Fixations on Natural Images

The spatiotemporal saliency map computed by the PQFT model can be used to predict human eye fixations. In order to evaluate the performance, the saliency maps should be compared to the eye fixations of human subjects in a visual search task. Here we use two data sets. One is ours which use fixation data collected from 12 subjects (aged from 24 to 55) and 100 natural images in the above experiment. The other one is that in [16] using fixation data collected from 20 subjects and 120 natural images. The main reason why we use the data set of [16] is that we want to make our results comparable to those in [16] and [17].

Fig. 10 shows the mean ROC value (ROC area) which is computed for each range of intersubject ROC values in our data set. The oblique line above is the upper bound of performance. It is clear that the prediction power of our PQFT model is over the other approaches, which is the closest to the upper boundary line (in Fig. 10). Table V presents average ROC area for all models according to all human fixation data across the entire data set. Our PQFT model still exhibits the best performance because its ROC area is the largest.

Table VI shows the average ROC area of the PQFT, PFT, RS, NVT, discriminant saliency (DS) and Information Maximization (Informax) models with respect to all human fixations in the data set of [16]. The ROC area of our PQFT model is still the largest, which demonstrates that the strong prediction power of the PQFT model over other models. Note that the results of DS

TABLE VI
ROC AREAS FOR DIFFERENT SALIENCY MODELS WITH RESPECT TO ALL HUMAN FIXATIONS IN THE DATA SET OF[16]

Model	PQFT	PFT [15]	SR [14]	GBVS [18]	Itti et al. [8]	DS [17]	Informax [16]
ROC area	0.8241	0.7520	0.7183	0.8110	0.6932	0.7694	0.7277

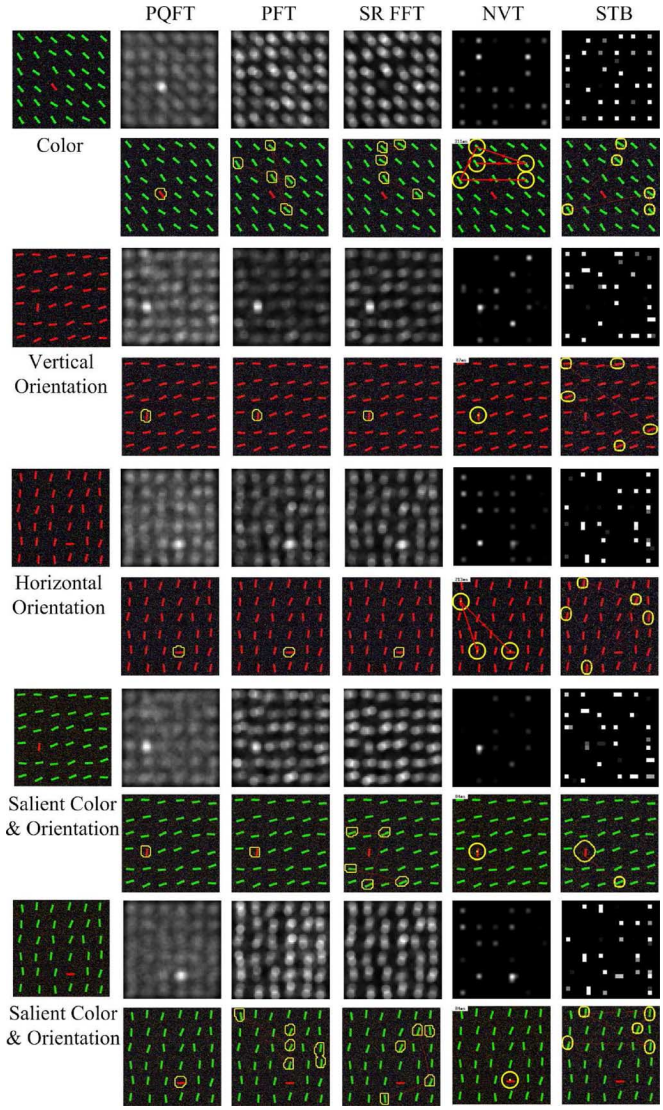


Fig. 11. Search results of salient color or orientation patterns.

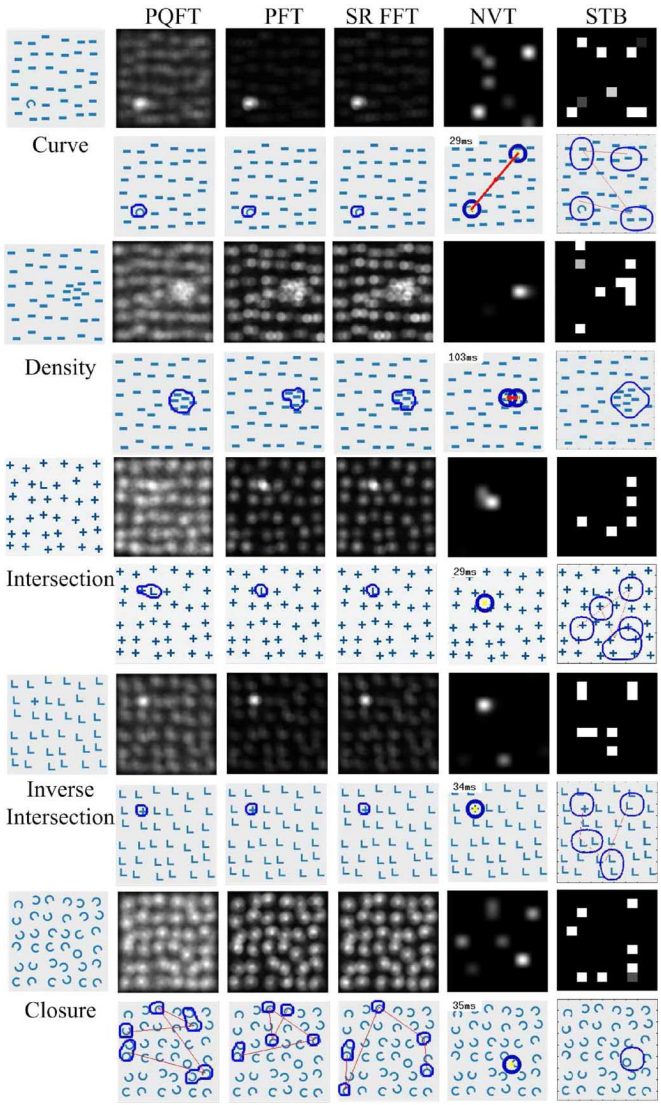


Fig. 12. Search results of salient orientation patterns.

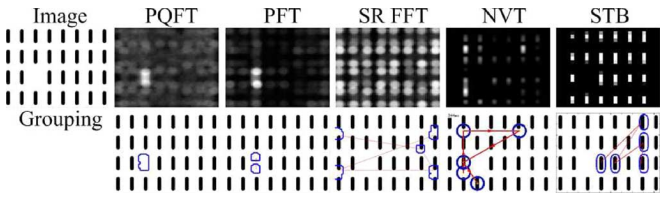


Fig. 13. Only PQFT and PFT can find the missing item.

and Informax model are collected from [16] and [17], respectively, in order to make their results fair. Here the resolution of the spatiotemporal saliency map is 32×32.

E. Psychological Patterns

Psychological patterns are widely used in visual attention experiments not only to explore the mechanism of visual search but also to test the effectiveness of saliency map [4], [5]. We used 17 patterns to compare the PQFT model with others, and the results are shown in Figs. 11–15.

In Fig. 11, the first image is a salient color pattern. Our PQFT model 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, the PQFT, PFT and SR models 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 salient in both color and orientation, which should be the easiest task. The PQFT model and NVT can find the targets by only one fixation. The PFT model and STB both find the salient red vertical

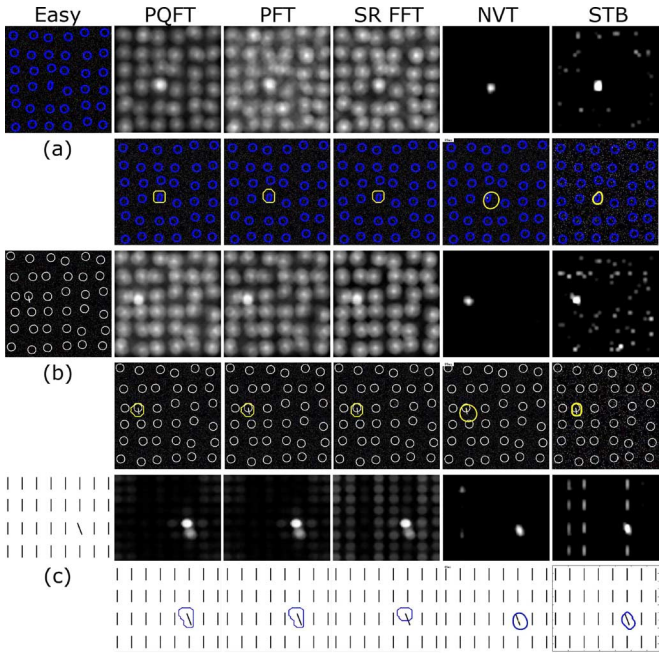


Fig. 14. Search Asymmetries. The patterns are easy for human to detect.

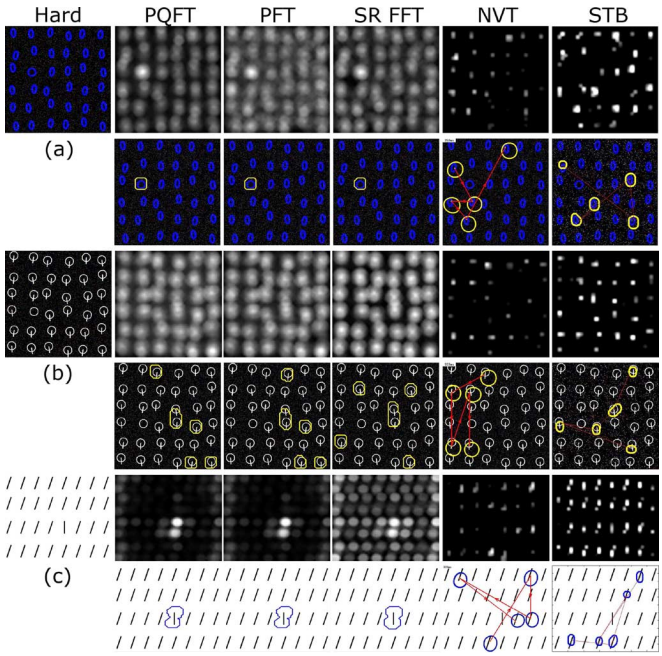


Fig. 15. Search Asymmetries. The patterns above is not so easy as those in Fig. 14 for human to detect.

bar but fail to detect the salient red horizonal bar. The SR model fails to find any of the salient targets above.

In Fig. 12, NVT can find all the patterns within five fixations and STB fails in the inverse intersection pattern. The PQFT, PFT and SR models fail in the closure search, which is one common limitation of these frequency domain models. Note that NVT and STB are capable of finding the target in closure pattern because of the denser intensity of the enclosed pieces, but not because of enclosure.

Our PQFT and PFT models can attend to the missing vertical black bar at the first fixation in Fig. 13, which agrees with human behavior. However, other models fail in this test. Note that the saliency maps by the PFT and SR models are quite different in this case although they look very similar in other tests.

Figs. 14 and 15 presents the search asymmetries reproduced by five saliency models. The patterns in Fig. 14 are much easier to be detected by human beings than those in Fig. 15. For example, finding the ellipse among many circles (requires shorter reaction time) is easier than finding a circle among many ellipses for humans. Moreover, finding a letter "Q" among many letters "O" and finding a tilted line among vertical ones are easier than the opposite. In a saliency detection task, a saliency model should overcome the search asymmetry that is observed in humans [40]. The PQFT, PFT, and SR models can detect two hard patterns [Fig. 15(a) and (c)] but only fails in one pattern Fig. 15(b) and other models fail in all cases, which shows the effectiveness of the PQFT, PFT and SR models in saliency detection. The patterns in Figs. 14 and 15 are used in [11] and [21].

In sum, our PQFT model performs best because it encounters only two detection failures [one in closure pattern Fig. 12, one in Fig. 15(b)] and needs only one fixation to detect the targets among all the other patterns, which shows that our spatiotemporal saliency map provides effective information to detect salient visual patterns.

V. Applications in Image and Video Coding

In the previous section, lots of experiments have proven that our PQFT model can extract salient proto-objects from videos and images fast and effectively, and is able to predict human eye fixations in a natural image better than other models in literature. Considering its powerful performance, we extends our model to improve coding efficiency of image and video compression in order to show its potential in engineering field.

A. Related Work

When human beings look at natural images or video clips, only a small region around the center of his eye fixation is captured at high resolution with logarithmic resolution falloff with eccentricity because of the nonuniform distribution of photoreceptor on the human retina [41]. Therefore, there is no need to compress the images or videos with uniform quality. The importance of priority coding has already been acknowledged by the engineering field as well, e.g., JPEG 2000 standard has included and implemented the region-of-interest (RoI) coding in its drafts [42], [43]. However, how to select the RoI still remains an open problem.

In the past years, many approaches have been proposed to find the RoI for image/video compression. Gaze-contingent approaches compress the videos according to the eye position of human subjects captured by an eye-tracking device [44]–[49]. Some other approaches use specific filters, which can mimic the known properties of HVS, to select the RoI [50]–[53]. However, all these approaches depend on human interaction or prior knowledge, and, therefore, they are impractical. In 2004, Itti proposed a visual attention model based on [7] and applied it

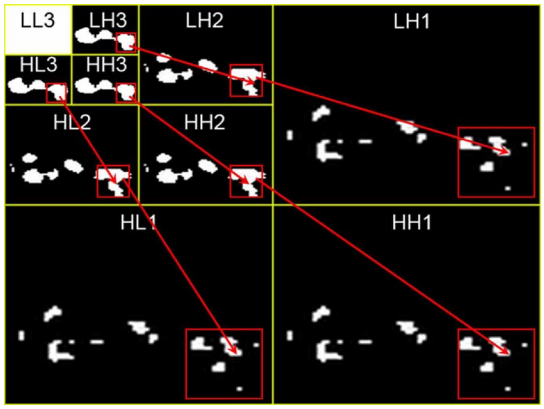


Fig. 16. Multiresolution wavelet domain foveation model.

to the video compression (MPEG-1 and MPEG-4). This visual attention approach does not rely on human interaction or prior knowledge. However, the weakness of this model lies in its high computational cost (see Section IV for details). As our saliency model outperforms the state-of-the-art models in predicting human eye fixations and is fast enough to meet real time requirements, we propose a new RoI selection model based on the PQFT model for image and video compression.

B. Multiresolution Wavelet Domain Foveation Model

With the help of multiresolution property in the PQFT model, a tree structure representation of a natural image can be constructed by the HS architecture (Section III). Based on this representation, we propose a MWDF model to automatically find the RoI in an input image or a video frame.

The wavelet domain foveated weighting (WDFW) model was proposed in [53], which uses specific HVS filters (e.g., skin detector to find the human face) to find the RoI (mask) and then applies the same mask to every level of DWT transform. Distinct from the WDFW model, our MWDF model have different masks under various wavelet transform level and these masks are detected at multiresolution by the PQFT model. Since specific HVS filters can not give a satisfactory RoI selection results in general data sets, we use our PQFT model to find the RoI for the WDFW model at the resolution of 64×64 in the following experiments for fairness.

Suppose that we have constructed the tree structure representation of an image by the PQFT-HS model, the leaf nodes at the finest resolution are combined as a mask for LH1, HL1 and HH1 sub-bands. Then in the next wavelet domain level, a coarser mask in the PQFT-HS tree is applied for LH2, HL2 and HH2 sub-bands. The depth of the tree is equal to the level of wavelet transform that is applied to image coding. The masks will never applied to LL sub-band. Fig. 16 gives an example of how the HS-MWDF model works on the image shown in Fig. 5. We keep all the coefficients in the RoI (white region in Fig. 16) and omit the coefficients below the bit plane threshold (BPT) in the background region (black region in Fig. 16).

Fig. 17 shows the compression results of the sheep image using the WDFW model and our HS-MWDF model. You can see that the sheep are very clear in both images. The main difference lies in the grass around the sheep. In Fig. 17(a), these grasses are also as clear as the sheep, which wastes the bits in the compressed file. However, in Fig. 17(b), the grass near the sheep is more blurred than the sheep but clearer than the background grassland. As a result, the file size of Fig. 17(a) is 328 Kb and that of Fig. 17(b) is only 261 Kb, which shows that our HS-MWDF model can save more bits and achieve higher precompression rate (The precompression rate is computed using lossless JPEG 2000 compression).

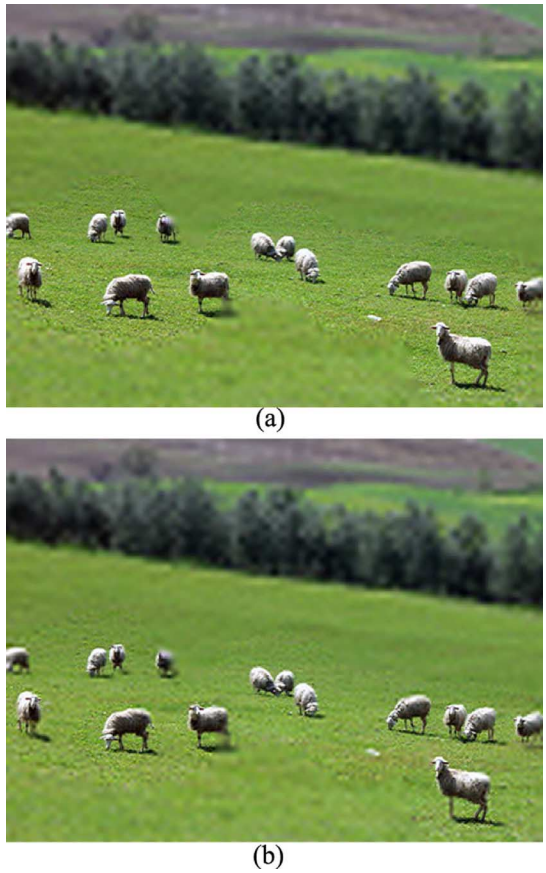


Fig. 17. Precompression results of the sheep image using WDFW (a) and HS-MWDF (b). As a result, the file size of (a) is 328 Kb and that of (b) is only 261 Kb. (The original image is 628 Kb).

C. Evaluate the Performance of the HS-MWDF Model in Image and Video Compression

In order to evaluate our model's influence on image and video compression, we use it as a front end before standard image compression [JPEG 2000] and video compression. For the PQFT-HS model, we set $\alpha = 0.75$ and $\gamma = 2$.

1) Image Compression: Table VII presents the comparison between the WDFW model and the HS-MWDF model on images' precompression rate in our data set (100 natural images with resolution around 800×600). The precompression rate is calculated by lossless JPEG 2000 codec. Table VII shows our HS-MWDF model can achieve higher precompression rate than the WDFW model under different fovea strategies. Here v fov means that we only use the first v OCAs found by the PQFT

Fig. 18. Compression results of two natural images in our data set. JPEG2000 is used and the compression rate of the images are 200.

Fig. 19. Compression results of the video. The moving people are very clear while the background (trees and roads) is blurred.

TABLE VII
COMPARISON BETWEEN WDFW AND HS-MWDF ON IMAGES' PRE-COMPRESSION RATE IN OUR DATA SET

Model	Auto fov	1 fov	3 fov	5 fov
WDFW	1.48	1.76	1.53	1.46
HS-MWDF	1.72	1.89	1.75	1.71

model and auto foveas (Auto fov) means we use (22) to decide the number of foveas.

Fig. 18 shows the compression results of two natural images in our data set. JPEG2000 is used as a encoder and the compression rate of the images are 200. Within RoI, the targets (human beings) in the compression results of the WDFW model (middle) and the HS-MWDF model (right) are much clearer than those without RoI selection (left). The results from the WDFW and HS-MWDF models are not distinguishable by human eyes in the RoI; however, the file size generated from the HS-MWDF model is smaller. Note that the good performance of the WDFW model relies on the accurate RoI selection by our PQFT model.

2) Video Compression: Table VIII shows the size of compressed files using different video compression standard (auto fov). The WDFW and HS-MWDF models are applied as the front end before the video coder respectively. We used the video clip in Section IV-B as the test video. Our HS-MWDF model can achieve higher compression rate (The compressed file is the smaller) than the WDFW model, which demonstrates that our proposed HS-MWDF model can help improve video's coding efficiency. Some of the video compression results are shown in Fig. 19. The RoIs (moving people) are clear and the background is blurred (trees and roads), which reduces the size of the compressed file.

3) Evaluate the Quality of the Attention-Based Compressed Image/Video: L. Itti pointed out that the quality of the attention-based compressed image/video should be evaluated by how well the results of the saliency model match the human fixation data [25]. The experimental results in Section IV show that our PQFT model has the best performance to detect objects and can best predict eye fixations among the state-of-the-art

TABLE VIII
COMPRESSED VIDEO FILE SIZE USING DIFFERENT COMPRESSION STANDARDS (AUTO FOV)

Model	MPEG-4	H.264
None	11.4Mb	8.88Mb
WDFW	8.53Mb	7.20Mb
HS-MWDF	7.07Mb	5.98Mb

saliency models [7], [8], [14], [16]–[18]. In addition, our PQFT model can work in real time (less than 1 ms per image/frame in C/C++ implementation); however, other models ([7], [8], [14], [16]–[18]) cannot do well. Therefore, our model is better than other models for the RoI selection of image/video compression.

VI. CONCLUSIONS AND DISCUSSIONS

Our work in [15] manifested the fact that the saliency map of an image can be calculated when the amplitude spectrum of the image is at a nonzero constant value. Thus, we proposed a saliency detection model for gray-level image called PFT, which is fast and simple.

The discovery of the phase spectrum's effect on the saliency map provides us with an easy way to extend the PFT model to the PQFT model which considers not only color, orientation and intensity within a single frame 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 spatiotemporal saliency map that the PQFT model produces can deal with videos, natural images and psychological patterns better than other state-of-the-art models, which shows that the PQFT model is an effective saliency detection model. In addition, our PQFT model is independent of parameters and prior knowledge, and is fast enough to meet real-time requirements as well (less than 1 ms per image/frame in C/C++ implementation).

Since the PQFT model can calculate the spatiotemporal saliency map of the input image at multiresolution, we propose a pure bottom-up Hierarchical Selectivity (HS) framework based on the PQFT model. A tree structure representation of an image is constructed by PQFT-HS, by which a multiresolution Wavelet Domain Foveation (HS-MWDF) model is developed as a front end before the image/video encoder. Our experimental results show that the model can improve coding efficiency in image and video compression tasks, and, therefore, achieve higher compression rate than the traditional WDFW model under different fovea strategies.

Comparing with human vision, our saliency model still has some limitations. First, the PQFT model can not deal with the closure pattern well up to now; however, human beings can find these patterns in a very short time. Second, we only applied our PQFT-based HS-MWDF model as the front end before the image and video compression tasks in this paper. Since the PQFT model is fast and can best predict the human fixation, we can insert our model into the image/video encoders to improve the efficiency of embedded encoders like EZW, SPIHT and EBCOT in still image/video compression. We will do more work to investigate these issues, and some biological explanations for our proposed model will be given in the future.

We believe that the potential of our work lies in many other engineering fields besides image and video compression, such as image retrieval and object recognition [13], [54], [55]. In addition, as our model only considers bottom-up information, it is necessary to add top-down signals (e.g., visual memory) for developing a humanoid vision system for robots.

ACKNOWLEDGMENT

The authors would like to thank D. Walther and J. Harel for sharing the codes and the discussions and N. D. Bruce and J. K. Tsotsos for the eye fixation database.

REFERENCES

[1] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," presented at the CVPR, 2003.
[2] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," presented at the CVPR, 2007.
[3] W. James, *The Principles of Psychology*. New York: Holt, 1890.
[4] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psych.*, vol. 12, no. 1, pp. 97–136, 1980.
[5] J. Wolfe, "Guided search 2.0: A revised model of guided search," *Psychonomic Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.
[6] C. Koch and S. Ullman, "Shifts in selection in visual attention: Toward the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.
[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[8] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, pp. 1489–1506, 2000.
[9] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Proc. NIPS*, 2005, pp. 547–554.
[10] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. CVPR*, 2005, pp. 631–637.
[11] L. Itti, "Models of Bottom-Up and Top-Down Visual Attention," Ph.D. dissertation, California Inst. chnol., Pasadena, 2000.
[12] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, pp. 1395–1407, 2006.
[13] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition–A gentle way," *Lecture Notes in Computer Science*, vol. 2525, no. 1, pp. 472–479, 2002.
[14] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," presented at the CVPR, 2007.
[15] C. L. Guo, Q. Ma, and L. M. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," presented at the CVPR, 2008.
[16] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," presented at the NIPS, 2005.
[17] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," presented at the NIPS, 2007.
[18] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," presented at the NIPS, 2006.
[19] G. Backer, B. Mertsching, and M. Bollmann, "Data- and model-driven gaze control for an active-vision system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1415–1429, Dec. 2001.
[20] M. Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," *IEEE Trans. Image Process.*, vol. 17, no. 5, pp. 5, 633–644, May 2008.
[21] S. Frintrop, "OCUS: A Visual Attention System for Object Detection and Goal-Directed Search," Ph.D. dissertation, 2006.
[22] O. L. Meur, O. L. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 802–817, 2006.
[23] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modelling visual attention via selective tuning," *Artif. Intell.*, vol. 78, no. 1-2, pp. 507–545, 1995.
[24] R. Rensink, "Seeing, sensing, and scrutinizing," *Vis. Res.*, vol. 40, no. 10-12, pp. 1469–1487, 2000.
[25] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[26] K. Castleman, *Digital Image Processing*. New York: Prentice-Hall, 1996.
[27] 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, Jul. 1997.
[28] S. J. Sangwine, "Fourier transforms of colour images using quaternion, or hypercomplex, numbers," *Electron. Lett.*, vol. 32, no. 21, pp. 1979–1980, Oct. 1996.
[29] T. A. Ell, "Hypercomplex Spectral Transforms," Ph.D. dissertation, Univ. Minnesota, Minneapolis, 1992.
[30] T. A. Ell, "Quaternion-Fourier transforms for analysis of two-dimensional linear time-invariant partial-differential systems," in *Proc. IEEE Conf. Decision and Control*, San Antonio, TX, Dec. 15–17, 1993, vol. 1–4, pp. 1830–1841.
[31] S.-C. Pei, J.-J. Ding, and J.-H. Chang, "Efficient implementation of quaternion Fourier transform, convolution, and correlation by 2-D complex FFT," *IEEE Trans. Signal Process.*, vol. 49, no. 11, pp. 2783–2797, Nov. 2001.
[32] S. J. Sangwine and T. A. Ell, J. M. Blackledge and M. J. Turner, Eds., "The discrete Fourier transform of a colour image," in *Proc. Image Processing II Mathematical Methods, Algorithms and Applications*, Chichester, U.K., 2000, pp. 430–441.
[33] T. Ell and S. Sangwin, "Hypercomplex Fourier transforms of color images," *IEEE Trans. Image Process.*, vol. 16, pp. 22–35, 2007.
[34] B. J. Scholl, "Objects and attention: The state of the art," *Cognition*, vol. 80, pp. 1–46, 2001.
[35] M. I. Posner and Y. Cohen, "Components of visual orienting," in *Attention and Performance*, H. Bouma and D. G. Bouwhuis, Eds. Hillsdale, NJ: Erlbaum, 1984, vol. 10, pp. 531–556.
[36] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artif. Intell.*, vol. 146, no. 1, pp. 77–123, May 2003.
[37] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, pp. 643–659, 2005.
[38] B. Horn and B. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1-3, pp. 185–203, 1981.
[39] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," presented at the ECCV, 2004.
[40] A. Treisman and S. Gormican, "Feature analysis in early vision: Evidence from search asymmetries," *Psych. Rev.*, vol. 95, pp. 14–58, 1988.
[41] B. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer, 1995.
[42] V. Sanchez, A. Basu, and M. K. Mandal, "Prioritized region of interest coding in JPEG2000," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 9, pp. 1149–1155, Sep. 2004.
[43] *JPEG 2000 Part I: Final Draft International Standard (ISO/IEC FDIS15444–1)*, , Aug. 2000, ISO/IECJTC1/SC29/WG1 N1855.
[44] P. T. Kortum and W. S. Geisler, "Implementation of a foveated imagecoding system for bandwidth reduction of video images," in *Proc. SPIE*, 1996, vol. 2657, pp. 350–360.
[45] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Improving the performance of MPEG coders using adaptive regions of interest," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 12, pp. 928–934, Dec. 1998.
[46] S. Lee, A. C. Bovik, and Y. Y. Kim, "Low delay foveated visual communications over wireless channels," in *Proc. Int. Conf. Image Processing*, 1999, pp. 90–94.
[47] U. Rauschenbach and H. Schumann, "Demand-driven image transmission with levels of detail and regions of interest," *Comput. Graph.*, vol. 23, no. 6, pp. 857–866, 1999.
[48] E. M. Reingold and L. C. Loschky, "Saliency of peripheral targets in gaze-contingent multiresolutional displays," *Behavior Res. Meth., Instrum. Comput.*, vol. 34, no. 4, pp. 491–499, 2002.
[49] D. J. Parkhurst and E. Niebur, "Variable-resolution displays: A theoretical, practical, and behavioral evaluation," *Human Factors*, vol. 44, no. 4, pp. 611–629, 2002.
[50] W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image using a model of the human visual system," in *Proc. ICPR*, Aug. 1998, pp. 701–704.
[51] F. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression," in *Proc. Picture Coding Symp.*, Apr. 2001, pp. 101–104.
[52] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions- of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.
[53] Z. Wang, L. G. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.
[54] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proc. CVPR*, 2006, pp. 2049–2056.
[55] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottomup attention useful for object recognition?," in *Proc. CVPR*, 2004, pp. 37–44.
[56] [Online]. Available: http://homepage.fudan.edu.cn/~clguo

Chenlei Guo (S'08) received the B.S. degree and M.S. degree in electronic engineering from Fudan University, Shanghai, China, in 2005 and 2008. He is currently pursuing the Ph.D. degree at Carnegie Mellon University, Pittsburgh, PA.

His primary research interest lies in modeling biologically-plausible computational visual attention, object detection/recognition, and brain computer interface.

Liming Zhang (M'92–SM'03) received the undergraduate degree in physics from Fudan University, Shanghai, China, in 1965.

From 1986 to 1988, she was a visiting scholar at the Electrical Engineering Department, University of Notre Dame, South Bend, IN. In 1996, she was a senior visiting Scholar at Munich Technology University, Munich, Germany. Now she is a full professor and Doctoral Advisor in the Department of Electronics, Fudan University, and a leader of the Image and Intelligence Laboratory. Since 1986, she has been engaged in artificial neural network, machine learning, feature selection, and pattern recognition of image and objects, including face recognition, brain-like robots, etc. Her group has accomplished more than ten projects supported by climbing program, national key project, natural sciences foundation, Shanghai Science and Technology Committee, etc. She has published more 120 papers on important national and international journals and conference proceedings concentrated on pattern recognition, machine learning, and neural networks.