Quaternion-based Spectral Saliency Detection for Eye Fixation Prediction

Boris Schauerte and Rainer Stiefelhagen

Karlsruhe Institute of Technology
Institute for Anthropomatics, Vincenz-Prießnitz-Str. 3
76131 Karlsruhe, Germany

Abstract In recent years, several authors have reported that spectral saliency detection methods provide state-of-the-art performance in predicting human gaze in images (see, e.g., [1–3]). We systematically integrate and evaluate quaternion DCT- and FFT-based spectral saliency detection [3,4], weighted quaternion color space components [5], and the use of multiple resolutions [1]. Furthermore, we propose the use of the eigenaxes and eigenangles for spectral saliency models that are based on the quaternion Fourier transform. We demonstrate the outstanding performance on the Bruce-Tsotsos (Toronto), Judd (MIT), and Kootstra-Schomacker eye-tracking data sets.

1 Introduction

There are many aspects that influence the human visual attention, but probably one of the most well-studied is that visual stimuli and objects that visually stand out of the surrounding environment automatically attract the human attention and, consequently, gaze. In the last two decades, many computational models of bottom-up visual attention have been proposed that model and try to help understand this inherent attractiveness, i.e. the visual saliency, of arbitrary stimuli and objects in a scene (see, e.g., [6]). Moreover, predicting where humans look is not only an interesting question in cognitive psychology, neurophysics, and neurobiology, but it has proven to be an important information for many application areas, e.g.: for efficient scene exploration and analysis in robotics (see, e.g., [7,8]), information visualization using image retargeting (see, e.g., [9,10]), or predicting the attractiveness of advertisement (see [11]).

In recent years, starting with the work by Hou *et al.* in 2007 [12], spectral saliency models attracted a lot of interest (see, e.g., [1,3–5,7,12–15]). These approaches manipulate the image's frequency spectrum to highlight sparse salient regions (see also [16,17]) and provide state-of-the-art performance, e.g., on psychological patterns (see [12]), for salient region detection (see [12]), and spatio-temporal eye fixation prediction (see [2,5]). But, what makes these approaches particularly attractive is the unusual combination of state-of-the-art performance and computational efficiency that is inherited from the fast Fourier transform. An interesting development in this area is the use of quaternions as a holistic representation to process color images as a whole [3, 4]. The quaternion algebra makes it possible to process color images as a whole without the need to process the image channels separately and, in consequence, tear apart the color information[1]. Given the definition of (real-valued) spectral saliency models, it does not seem possible to calculate them on color images without separation of the image channels if it were not for the quaternion algebra and its hypercomplex discrete Fourier transform (DFT) and discrete cosine transform (DCT) definitions.

In this paper, we combine and extend the previous work on spectral saliency detection. Most importantly, we integrate and investigate the influence of quaternion component weights as proposed by Bian *et al.* [5], adapt the multiscale model by Peters *et al.* [1] and evaluate its effect on the quaternion-based approaches, and propose and evaluate the use of the quaternion eigenaxis and eigenangle for saliency algorithms that rely on the quaternion Fourier transform (see, e.g., [4, 12, 13]). Furthermore, we evaluate the choice of the color spaces that have been applied in previous work (see, e.g., [2]) and also address the influence of the quaternion DFT and quaternion DCT transformation axis. We evaluate all algorithms on the Bruce-Tsotsos (Toronto), Judd (MIT), and Kootstra-Schomacker data sets (see [18–20]) and analyze how well these models can predict where humans look at in natural scenes. To this end, we use the well-known area under curve (AUC) of the receiver operator characteristic (ROC) as a measure of the predictive power (see, e.g., [2]).

In summary, we are able to improve the state-of-the-art on the Toronto, and Kootstra-Schomacker data set in terms of the area under the ROC curve. On the Judd data set, our approach is outperformed by Judd's algorithm and we achieve only the second best performance. This can be explained by the fact that Judd's algorithm explicitly models the influence of higher level concepts, which are very prominent in Judd's data set. However, the evaluated spectral saliency algorithms achieve the performance at a fraction of the computational requirements of Judd's algorithm. The proposed use of the eigenaxis and eigenangle substantially improves the performance of the quaternion Fourier approach (PQFT; see [4]) and makes it the second best choice after the quaternion DCT signature saliency (QDCT; see [3]). The use of multiple scales significantly improves the results and the use of appropriately weighted color spaces is essential to achieve outstanding performance using quaternion-based spectral saliency detection.

2 Related Work

Visual saliency is a concept that has been derived from human perception and describes how likely it is that a stimulus attracts the attention (see, e.g., [6]). Many factors influence the human attention and we have to distinguish between bottom-up, data-driven as well as top-down, knowledge-driven aspects. In this paper, we consider the traditional concept of visual saliency that is primarily linked to the bottom-up attention that automatically attracts the human gaze.

[1] A simple, illustrated example and discussion of the disadvantage of processing an image's color channels separately can be found in [5, Sec. 3.1].

One of the most influential works is the feature integration theory by Treisman and Gelade [21], which is probably the first model that used several feature dimensions to calculate a *saliency map* of the image that tries to estimate how salient each image region is. Since then many different saliency models have been proposed to calculate such maps; for example, the well-known model by Itti and Koch [22], attention by information maximization (AIM) [18], saliency using natural statistics (SUN) [23], graph-based visual saliency (GBVS) [24], context-aware saliency (CAS) [9, 10], and Judd's model that is based on machine learning [19]. However, since reviewing the existing approaches is beyond the scope of this paper, we refer the interested reader to the survey by Frintrop *et al.* [6]. In the following, we summarize the most closely related work.

The first spectral approach for visual saliency detection was presented in 2007 by Hou *et al.* [12]. Since then, several spectral saliency models have been proposed (see, e.g., [1, 4, 5, 12–15]). Hou *et al.* proposed the use of the Fourier transform to calculate the visual saliency of an image. To this end, – processing each color channel separately – the image is Fourier transformed and the magnitude components are attenuated (spectral residual). Then, the inverse Fourier transform is calculated using the manipulated magnitude components in combination with the original phase angles. The saliency map is obtained by calculating the absolute value of each pixel of this inverse transformed image and subsequent Gaussian smoothing. This way Hou *et al.* achieved state-of-the-art performance for salient region (proto-object) detection and psychological test patterns. However, although Hou *et al.* were the first to propose this method for saliency detection, it has been known for at least three decades that suppressing the magnitude components in the frequency domain highlights signal components such as lines, edges, or narrow events (see [16, 17]).

In 2008 [1], Peters *et al.* analyzed the role of Fourier phase information in predicting visual saliency. They extended the model of Hou *et al.* by linearly combining the saliency of the image at several scales. Then, they analyzed how well this model predicts eye fixations and found that "salience maps from this model significantly predicted the free-viewing gaze patterns of four observers for 337 images of natural outdoor scenes, fractals, and aerial imagery" [1].

Also in 2008 [4], Guo *et al.* proposed the use of quaternions as a holistic color image representation for spectral saliency calculation. This was possible because quaternions provide a powerful algebra that allows to realize a hypercomplex Fourier transform (see [25]), which was first demonstrated to be applicable for color image processing by Sangwine [26,27]. Thus, Guo *et al.* were able to Fourier transform the image as a whole and did not have to process each color channel separately[1]. Furthermore, this made it possible to use the scalar part of the quaternion image as 4th channel to integrate a motion component. However, in contrast to Hou *et al.*, Guo *et al.* did not use the spectral residual. Most interestingly, Guo *et al.* were able to determine salient people in videos and outperformed the models of Itti *et al.* [22] and Walther *et al.* [28]. In 2010 [13], a multiresolution attention selection mechanism was introduced, but the definition of the main saliency model remained unchanged. However, most interestingly, further

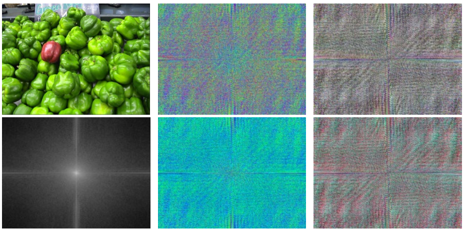

Figure 1. Visualization (see [30] on how to interpret it) of the quaternion Fourier spectrum of an example image for two transformation axes (1&2). Left-to-right, top-to-bottom: original, eigenangles (1), eigenaxes (1), magnitude (1), eigenangles (2), and eigenaxes (2). This illustration is best viewed in color.

experiments demonstrated that the approach outperformed several established approaches in predicting eye gaze on still images.

In 2009 [5], Bian *et al.* adapted the work by Guo *et al.* by weighting the quaternion components[2]. Furthermore, they provide a biological justification for spectral visual saliency models and proposed the use of the YUV color space, in contrast to the use of the previously applied intensity and color opponents (ICOPP) [4,13], and RGB [12]. This made it possible to outperform the models of Bruce *et al.* [18], Gao *et al.* [29], Walther and Koch [28], and Itti and Koch [22] when predicting human eye fixations on video sequences.

In 2012 [2], Hou *et al.* proposed and theoretically analyzed the use of the discrete cosine transform (DCT) for spectral saliency detection. They showed that this approach outperforms (in terms of the AUC) all other evaluated state-of-the-art approaches – e.g., Itti and Koch [22], AIM [18], GBVS [24], and SUN [23] – in predicting human eye fixations on the Toronto data set [18]. Furthermore, Hou *et al.* pointed out the importance of choosing an appropriate color space.

Also in 2012 [3], Schauerte *et al.* used the definition of a quaternion DCT and quaternion signatures to calculate the visual saliency and was able to outperform the real-valued approach by Hou *et al.* [2]. This way they improved the state-of-the-art in predicting where humans look in the presence and absence of faces.

3 Saliency Model

3.1 Basic Quaternion Definitions

Quaternion Algebra: Quaternions form a 4D algebra $\mathbf{H}$ over the real numbers and are – in principle – an extension of the 2D complex numbers [31]. A quaternion q is defined as $q = a + bi + cj + dk \in \mathbf{H}$ with $a, b, c, d \in \mathbf{R}$, where i, j, and k provide the basis to define the (Hamilton) product of two quaternions q_1 and q_2 ($q_1, q_2 \in \mathbf{H}$):

$$q_1 q_2 = (a_1 + b_1 i + c_1 j + d_1 k)(a_2 + b_2 i + c_2 j + d_2 k)\,, \tag{1}$$

[2] Please note that the mentioned SW approach [5] is in principle equivalent to the PFT approach by Guo *et al.* [4].

where $i^2 = j^2 = k^2 = ijk = -1$. Since, for example, by definition $ij = k$ while $ji = -k$ the Hamilton product is not commutative. Accordingly, we have to distinguish between left-sided and right-sided multiplications (marked by L and R, respectively, in the following). A quaternion q is called real, if $x = a + 0i + 0j + 0k$, and pure imaginary, if $q = 0 + bi + cj + dk$. We can define the operators $S(q) = a$ and $V(q) = bi + cj + dk$ that extract the scalar part and the imaginary part of a quaternion $q = a + bi + cj + dk$, respectively. As for complex numbers, we can define conjugate quaternions $\bar{q} = a - bi - cj - dk$ as well as the norm $|q| = \sqrt{q \cdot \bar{q}}$. Furthermore, we can define the quaternion scalar product $* : \mathbf{H} \times \mathbf{H} \to \mathbf{R}$

$$s = q_1 * q_2 = a_1 a_2 + b_1 b_2 + c_1 c_2 + d_1 d_2 \,. \tag{2}$$

Eigenaxis and Eigenangle: Euler's formula for the polar representation using the complex exponential generalizes to (hypercomplex) quaternion form

$$e^{\mu \varPhi} = \cos \varPhi + \mu \sin \varPhi \,, \tag{3}$$

where μ is a unit pure quaternion (see [27] and [13]). Consequently, any quaternion q may be represented in a polar representation such as:

$$q = |q| e^{\gamma \varPhi} \tag{4}$$

with the norm $|q|$, its *eigenaxis* γ

$$\gamma = f_\gamma(q) = \frac{V(q)}{|V(q)|} \,, \tag{5}$$

and the corresponding *eigenangle* $\varPhi$

$$\varPhi = f_\varPhi(q) = \arctan\left(\frac{|V(q)| \operatorname{sgn}(V(q) * \gamma)}{S(q)}\right) \tag{6}$$

with respect to the eigenaxis γ, which is a unit pure quaternion, and where $\operatorname{sgn}(\cdot)$ is the signum function (see [27]). The eigenaxis γ specifies the quaternion direction in the 3-dimensional space of the imaginary, vector part and can be seen as being a generalization of the imaginary unit of complex numbers. Analogously, the eigenangle $\varPhi$ corresponds to the argument of a complex number.

3.2 Quaternion Images

Every image $\mathbf{I} \in \mathbb{R}^{M \times N \times C}$ – with at most 4 color components, i.e. $C \leq 4$ – can be represented using a $M \times N$ quaternion matrix

$$\begin{align} \mathbf{I}_\mathrm{Q} &= \mathbf{I}_4 + \mathbf{I}_1 i + \mathbf{I}_2 j + \mathbf{I}_3 k \tag{7} \\ &= \mathbf{I}_4 + \mathbf{I}_1 i + (\mathbf{I}_2 + \mathbf{I}_3 i) j \quad \text{(symplectic form)}, \tag{8} \end{align}$$

where $\mathbf{I}_c$ denotes the $M \times N$ matrix of the cth image channel. It is common to represent the (potential) 4th image channel as the scalar part (see, e.g., [27]), because when using this definition it is possible to work with pure quaternions for the most common color spaces such as, e.g., RGB, YUV and Lab.

Weighted Quaternion Color Components: Naturally, as done by Bian *et al.* [5] and also related to the recent trend to learn feature dimension weights (see, e.g., [32] and [19]), we can model the relative importance of the color space components for the visual saliency by introducing a quaternion component weight vector $\mathbf{w} = [w_1 \; w_2 \; w_3 \; w_4]^{\mathrm{T}}$ and adapting Eq. 7 appropriately:

$$\mathbf{I}_{\mathrm{Q}} = w_4\mathbf{I}_4 + w_1\mathbf{I}_1 i + w_2\mathbf{I}_2 j + w_3\mathbf{I}_3 k \,. \tag{9}$$

In case of equal influence of each color component, i.e. uniform weights, Eq. 7 is a scaled version of Eq. 9, which is practically equivalent for our application.

3.3 Quaternion Transforms and Transformation Axis

Quaternion Discrete Fourier Transform: We can transform a $M \times N$ quaternion matrix $\mathbf{f}$ using the definition of the quaternion Fourier transform $\mathscr{F}^{\mathrm{L}}_{\mathrm{Q}}$ [30]:

$$\mathscr{F}^{\mathrm{L}}_{\mathrm{Q}}[\mathbf{f}](u,v) = \mathbf{F}^{\mathrm{L}}_{\mathrm{Q}}(u,v) \tag{10}$$

$$\mathbf{F}^{\mathrm{L}}_{\mathrm{Q}}(u,v) = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\eta 2\pi((mv/M)+(nu/N))} \mathbf{f}(m,n) \,,$$

see Fig. 1 for an example. The corresponding inverse quaternion discrete Fourier transform $\mathscr{F}^{-\mathrm{L}}_{\mathrm{Q}}$ is defined as:

$$\mathscr{F}^{-\mathrm{L}}_{\mathrm{Q}}[\mathbf{F}](m,n) = \mathbf{f}^{\mathrm{L}}_{\mathrm{Q}}(m,n) \tag{11}$$

$$\mathbf{f}^{\mathrm{L}}_{\mathrm{Q}}(m,n) = \frac{1}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} e^{\eta 2\pi((mv/M)+(nu/N))} \mathbf{F}(u,v) \,.$$

Here, η is a unit pure quaternion that serves as an axis and determines a direction in the color space. Although the choice of η is arbitrary, it is not without consequence (see [30, Sec. V]) and can influence the results. For example, in RGB a good axis candidate would be the "gray line" and thus $\eta = (i + j + k)/\sqrt{3}$. In fact, as discussed by Ell and Sangwine [30], this would decompose the image into luminance and chrominance components.

Quaternion Discrete Cosine Transform: Similarly, it is possible to define a quaternion discrete cosine transform:

$$\mathscr{D}^{\mathrm{L}}_{\mathrm{Q}}[\mathbf{f}](u,v) = \frac{2}{\sqrt{MM}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \eta f(m,n) \beta^{M}_{u,m} \beta^{N}_{v,n} \tag{12}$$

$$\mathscr{D}^{-\mathrm{L}}_{\mathrm{Q}}[\mathbf{F}](m,n) = \frac{2}{\sqrt{MM}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \eta F(u,v) \beta^{M}_{u,v} \beta^{N}_{m,n} \tag{13}$$

with $\beta^{M}_{u,m} = \cos\left[\frac{\pi}{M}(m + \frac{1}{2})u\right]$ (see, e.g., [3]). However, as can be seen when comparing Eq. 10 and 12, the definition of $\mathscr{D}^{\mathrm{L}}_{\mathrm{Q}}$ is substantially different from

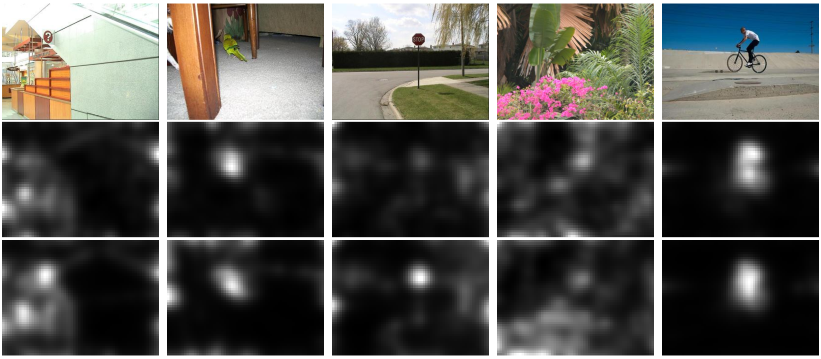

Figure 2. Example images (top) from the data sets (see Sec. 4.1) that illustrate the difference of the PQFT (middle) and proposed EigenPQFT (bottom) saliency maps.

$\mathscr{F}_\mathrm{Q}^\mathrm{L}$, because the factors $\beta_{u,m}^{M}$ are real-valued instead of the hypercomplex terms of $\mathscr{F}_\mathrm{Q}^\mathrm{L}$, as discussed by Schauerte *et al.* [3]. However, both definitions share the concept of a unit pure quaternion η that serves as a transformation axis.

3.4 Eigenaxis and -angle Spectral Saliency

Similar to the real-numbered definition of the spectral residual by Hou *et al.* [12], let $\mathbf{A}_\mathrm{Q}$ denote the amplitude, $\mathbf{E}_\gamma$ the eigenaxes, and the eigenangles $\mathbf{E}_\Theta$ (see Sec. 3.1) of the quaternion image $\mathbf{I}_\mathrm{Q}$:

$$\mathbf{E}_\gamma(x,y) = f_\gamma(\mathbf{I}_\mathrm{Q}(x,y)) \tag{14}$$

$$\mathbf{E}_\Theta(x,y) = f_\Theta(\mathbf{I}_\mathrm{Q}(x,y)) \tag{15}$$

$$\mathbf{A}_\mathrm{Q}(x,y) = |\mathbf{I}_\mathrm{Q}(x,y)| \, . \tag{16}$$

Then, we calculate the log amplitude and a low-pass filtered log amplitude using a Gaussian filter h_{σ_A} with the standard deviation σ_A to obtain the spectral residual R_Q:

$$\mathbf{L}_\mathrm{Q}(x,y) = \log\, \mathbf{A}_\mathrm{Q}(x,y) \tag{17}$$

$$\mathbf{R}_\mathrm{Q}(x,y) = \mathbf{L}_\mathrm{Q}(x,y) - \left[\mathbf{h}_{\sigma_\mathrm{A}} * \mathbf{L}_\mathrm{Q}\right](x,y) \, . \tag{18}$$

Finally, we can calculate the *Eigen Spectral Residual* (*EigenSR*) saliency map S_ESR using the spectral residual R_Q, the eigenaxis E_γ, and the eigenangle E_Θ:

$$S_\mathrm{ESR} = \mathscr{S}_\mathrm{ESR}(\mathbf{I}_\mathrm{Q}) = \mathbf{h}_{\sigma_\mathrm{S}} * \left|\mathscr{F}_\mathrm{Q}^{-\mathrm{L}}\left[e^{\mathbf{R}_Q + \mathbf{E}_\gamma \circ \mathbf{E}_\Theta}\right]\right| , \tag{19}$$

where $\circ$ denotes the Hadamard product and $\mathbf{h}_{\sigma_\mathrm{S}}$ is a Gauss filter with standard deviation σ_S. If σ_A approaches zero, then the spectral residual $\mathbf{R}_Q$ will become $\mathbf{0}$, i.e. $\lim_{\sigma_\mathrm{A} \to 0^+} \mathbf{R}_\mathrm{Q}(x,y) = 0$, in which case we refer to the model as the *Eigen Spectral Whitening* model (*EigenSW* or *EigenPQFT*).

If the input image is a single-channel image, then the quaternion definitions and equations are reduced to their real-valued counterparts, in which case Eq. 19 is identical to the single-channel real-numbered definition by Hou *et al.* [12]. Our EigenSR and EigenPQFT definition that is presented in Eq. 19 differs from Guo's PQFT [4] definition in two aspects: First, it – in principle – preserves Hou's spectral residual definition [12]. Second, it relies on the combination of the eigenaxes and eigenangles instead of the combination of a single unit pure quaternion and the corresponding phase spectrum (see [13, Eq. 16] and [4, Eq. 20]), see Fig. 2 for an illustration.

3.5 Multiple Scales

The above saliency definitions only consider a fixed, single scale (see, e.g., [2–5, 13]). But, the scale is an important parameter when calculating the visual saliency and an integral part of many saliency models (see, e.g., [6]). For spectral approaches the scale is (implicitly) defined by the resolution of the image I_{Q} (see, e.g., [33]). Consequently, as proposed by Peters and Itti [1], it is possible to calculate a multiscale saliency map $\mathbf{S}^M$ by combining the spectral saliency of the image at different image scales. Let $\mathbf{I}_{\mathrm{Q}}^m$ denote the quaternion image at scale $m \in M$, then

$$S^M = \mathscr{S}^M(\mathbf{I}_{\mathrm{Q}}) = h_{\sigma_{\mathrm{M}}} * \sum_{m \in M} \phi_R(\mathscr{S}(\mathbf{I}_{\mathrm{Q}}^m)) \,, \tag{20}$$

where ϕ_R rescales the matrix to the target saliency map resolution R and $h_{\sigma_{\mathrm{M}}}$ is an additional, optional Gauss filter.

4 Evaluation and Discussion

4.1 The Data Set, Algorithms, and Measures

Data Sets: To evaluate the considered saliency algorithms, we use the following eye-tracking data sets: The Toronto data set by Bruce and Tsotsos (see [18]), which consists of 120 images (681 × 511 pixels) and the corresponding eye-tracking data of 20 subjects that were free-viewing each image for 4 seconds. The data set by Kootstra and Schomacker [20], which consists of 100 images (1024 × 768 pixels) and eye-tracking data of 31 subjects that free-viewed the images. The images are subdivided into 5 categories and were selected from the McGill calibrated color image database [34]. Furthermore, we use the data set by Judd et al. (see [19]), which is also known as MIT data set and is the largest publicly available data set. It consists of 1003 images (varying resolution and aspect ratio) and eye-tracking data of 15 viewers.

Algorithms: We evaluate the following spectral saliency methods[3]: spectral residual (SR) [12], pure Fourier transform aka. spectral whitening (PFT) [4,13],

[3] Please note that our reference implementations are publicly available and free (BSD License) to allow for fair benchmarking and evaluation by other authors in the future.

PFT at multiple scales (ΔPFT) [1], pure quaternion Fourier transform (PQFT) [4,13], DCT signature (DCT) [2], and quaternion DCT signature (QDCT) [3]. We mark algorithms that use multiple scales with a preceding Δ (imagine a stylized image pyramid). To serve as a reference, we use the following algorithms[4] as baselines: the Itti and Koch model [22], graph-based visual saliency (GBVS) [24], context-aware saliency (CAS) [9, 10], attention using information maximization [18], and Judd's model [19].

We evaluate how well the proposed algorithms perform for all color spaces that have been applied in the closely related literature, i.e.: red-green-blue (RGB) (see, e.g., [2, 12]), CIELAB (Lab) (see [2]), intensity and red-green/blue-yellow color opponents (ICP) (see [4,13]), and YUV which consists of the luma Y and the two chrominance components U and V (see [5]).

ΔPQFT, EigenPQFT, ΔEigenPQFT, EigenSR, ΔEigenSR, and ΔQDCT are methods that are first introduced and evaluated in this paper. Except for the PQFT in combination with YUV that was proposed by Bian *et al.* [5], the influence of the quaternion component weight has not been evaluated for any quaternion-based algorithm.

Evaluation Measure: We follow the evaluation procedure described by Hou *et al.* [2] and use the center-bias corrected area under the receiver operating characteristic curve as performance measure. As has been shown in prior art (see, e.g., [33]), the human gaze is often biased toward the center of an image. To remove this bias (see [2] and [35]), we define a positive and a negative sample set of eye fixation points for each image. The positive sample set contains the fixation points of all subjects on that image. The negative sample set contains the union of all eye fixation points across all images from the same data set, but excluding the samples of the positive sample set. Each saliency map can be thresholded and the resulting binarized (saliency) map can be considered to be a binary classifier that tries to separate positive and negative samples. Thus, for each threshold, the true positive rate is the proportion of the positive samples that fall in the positive region of the binarized (saliency) map. Analogously, the false positive rate can be calculated by using the negative sample set. Sweeping over all thresholds leads to the receiver operating characteristic (ROC) curves. The area under the ROC curves is a widely used compact measure for the ability of the saliency map to predict human eye fixations. Chance would lead to an AUC of 0.5, values < 0.5 indicate a negative correlation, and values > 0.5 indicate a positive correlation (perfect prediction is 1.0).

4.2 Experimental Results and Discussion

We kept the image resolution fixed at 64×48 pixels in the evaluation, because in preparatory pilot experiments this resolution has constantly shown to provide

[4] When available, we used the publicly available reference implementation from the authors. For the Itti and Koch model we used the implementation by Harel [24], which achieved a better performance than the iLab Neuromorphic Vision Toolkit (iNVT).

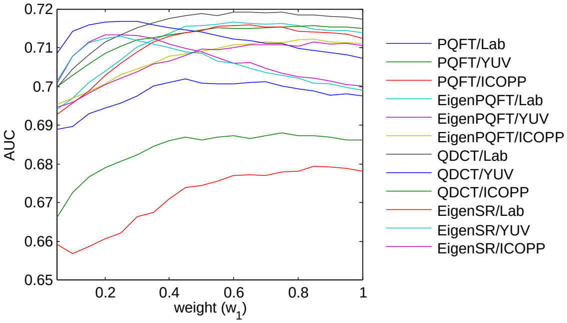

Figure 3. Example of the influence of quaternion color component weights on the AUC performance for QDCT, EigenPQFT, EigenSR, and PQFT on the Toronto data set. Using the default quaternion transformation axis $\eta = (i + j + k)/\sqrt{3}$.

very good results on all data sets and is the resolution most widely used in the literature (see, e.g., [2,3]). For multiscale approaches 64×48 pixels is consequently the base resolution. For the Gaussian filtering of the saliency maps, we use the fast recursive filter implementation by Geusebroek *et al.* [36].

The achievable performance depends substantially on the data set, see Tab. 1. We can rank the data sets by the maximum area under the ROC curve that spectral algorithms achieved and obtain the following descending order: Toronto, Judd, and Kootstra. This order can most likely be explained with the different characteristics of the images in each data set. Two image categories are dominant within the Toronto data set: street scenes and objects. Furthermore, the images have relatively similar characteristics. The Judd data set contains many images from two categories: images that depict landscapes and images that show people. The second category is problematic for low-level approaches that do not consider higher-level influences on visual attention such as, e.g., the presence of people and faces in images. This also is the most important aspect why Judd's model excels on this data set. The Kootstra data set is the data set with the highest data variability. It contains five image categories, close-up as well as landscape images, and images with and without a strong photographer bias.

RGB is the color space with the worst performance. While Lab provides the best performance on the Toronto data set, it is outperformed by YUV and ICOPP on the Kootstra and Judd data set. Since YUV is the best color model on the Kootstra and Judd data set and is close to the performance of Lab on the Toronto data set, we advise the use of the YUV color space.

Within each color space and across all data sets the performance ranking of the algorithms is relatively stable, see Tab. 1. We can observe that with naive parameters the performance of the quaternion-based approaches may be lower than the performance of their real-valued counterparts. The extent of this effect depends on the color space as well as on the algorithm. For example, for QDCT this effect does not exist on the RGB and ICOPP color spaces and for Lab only on the Kootstra data set. However, over all data sets this effect is most apparent for

	Area under the Receiver Operating Characteristic curve (mean)											
	Toronto				Kootstra				Judd			
Method	Lab	YUV	ICP	RGB	Lab	YUV	ICP	RGB	Lab	YUV	ICP	RGB
	Weighted Color Space and Non-Standard Axis											
ΔQDCT	.7201	.7188	.7174	.7091	.6104	.6125	.6110	.6007	.6589	.6751	.6712	.6622
QDCT	.7195	.7170	.7158	.7066	.6085	.6119	.6106	.5994	.6528	.6656	.6623	.6552
ΔEigenPQFT	.7183	.7160	.7144	.7035	.6053	.6082	.6064	.5963	.6527	.6658	.6617	.6559
EigenPQFT	.7180	.7137	.7122	.7006	.6058	.6073	.6063	.5934	.6483	.6611	.6568	.6493
ΔEigenSR	.7175	.7153	.7133	.7014	.6050	.6077	.6056	.5941	.6508	.6649	.6603	.6534
EigenSR	.7162	.7129	.7112	.6990	.6038	.6068	.6044	.5912	.6467	.6601	.6554	.6470
ΔPQFT	.7085	.6969	.6927	.6930	.5943	.5994	.5922	.5868	.6467	.6503	.6429	.6468
PQFT	.7042	.6881	.6826	.6891	.5930	.5970	.5913	.5861	.6404	.6416	.6379	.6398
PQFT/Bian	.7035	.6880	.6817	.6884	.5928	.5961	.5911	.5861	.6404	.6411	.6375	.6396
	Uniform Color Space Weights and Standard Axis											
ΔQDCT	.7191	.7107	.7070	.7088	.6050	.6036	.6078	.6002	.6539	.6648	.6618	.6620
QDCT	.7180	.7079	.7039	.7056	.6036	.6005	.6079	.5987	.6517	.6572	.6552	.6551
ΔEigenPQFT	.7148	.7030	.7024	.7026	.6005	.5963	.6045	.5959	.6490	.6530	.6548	.6556
EigenPQFT	.7141	.7006	.6982	.7006	.5984	.5939	.6023	.5934	.6461	.6496	.6518	.6491
ΔEigenSR	.7142	.7135	.7006	.7013	.6003	.5951	.6028	.5937	.6477	.6504	.6534	.6531
EigenSR	.7132	.6998	.6969	.6988	.5975	.5930	.6007	.5909	.6448	.6486	.6502	.6466
ΔPQFT	.7022	.6925	.6868	.6927	.5803	.5826	.5877	.5850	.6431	.6441	.6380	.6465
PQFT	.6974	.6858	.6796	.6884	.5788	.5808	.5860	.5846	.6368	.6368	.6271	.6396
	Non-Quaternion Spectral Baseline Algorithms											
DCT	.7137	.7131	.7014	.6941	.6052	.6089	.6049	.5907	.6465	.6604	.6556	.6461
ΔPFT	.7177	.7170	.7079	.7014	.6072	.6107	.6084	.5945	.6502	.6601	.6583	.6523
PFT	.7140	.7120	.7025	.6958	.6057	.6079	.6058	.5908	.6445	.6590	.6572	.6446
SR	.7156	.7144	.7051	.6983	.6059	.6090	.6061	.5916	.6462	.6599	.6573	.6461
	Baseline Algorithms (algorithm-specific feature spaces)											
CAS	.6921				.6034				.6622			
AIM	.6986				.5749				.6662			
Judd	.6847				.5793				.7696			
GBVS	.6703				.5791				.6539			
Itti	.6492				.5672				.6433			

Table 1. Area under the receiver-operator characteristic curve (ROC AUC) of the evaluated algorithms (0.5 is chance level, > 0.5 indicates positive correlation, and an AUC of 1.0 represents perfect prediction). In the default configuration, the algorithms use a uniform color component weight (except for the approach by Bian *et al.*, which is denoted as PQFT/Bian), and the uniform transform axis $\eta = (i + j + k)/\sqrt{3}$. In the optimized configuration, we use appropriate quaternion component weights and non-standard axes. The overall optimal quaternion component weight vectors are roughly $[0.5\ 1\ 1\ 0]^{\mathrm{T}}$, $[0.2\ 1\ 1\ 0]^{\mathrm{T}}$, and $[0.8\ 1\ 1\ 0]^{\mathrm{T}}$ for Lab, YUV, and ICOPP, respectively. The best overall performance can be achieved with a smoothing filter standard deviation (σ_s and σ_M for single-scale and multiscale, respectively) of roughly 0.039 (image widths). The optimal axis depends on the algorithm as well as on the color space. But, interestingly, for RGB the most reliable axis is in fact the "gray line", see Sec. 3.3. In general, the influence of the quaternion FFT and DCT transformation axis is secondary to the influence of the quaternion component weights. But, a badly chosen quaternion transformation axis can substantially degrade the performance.

the YUV color space. But, the YUV color space is also the color space that profits most from non-uniform quaternion component weights with an optimal weight vector around $[0.2\ 1\ 1\ 0]^{\mathrm{T}}$, which indicates that the unweighted influence of the luminance component is too high. When weighted appropriately, as mentioned before, we achieve the overall best results using the YUV color space.

The influence of quaternion component weights is substantial, see Tab. 1, and depends on the color space. As discussed it is most important for the YUV color space. However, it is also substantial for Lab and ICOPP. Most importantly, the best weights are relatively constant over all data sets. Accordingly, we advise the use of the following weight vectors $[0.5\ 1\ 1\ 0]^{\mathrm{T}}$, $[0.2\ 1\ 1\ 0]^{\mathrm{T}}$, and $[0.8\ 1\ 1\ 0]^{\mathrm{T}}$ for Lab, YUV, and ICOPP, respectively. In general, the influence of the quaternion transformation axis is secondary to the influence of the quaternion component weights. However, a badly chosen axis can significantly degrade the performance.

The influence of multiple scales varies depending on the data set. For the Toronto data set the influence is small, which can be explained by the fact that the resolution of 64×48 pixels is nearly optimal for this data set (see [2]). On the Kootstra data set the influence is also relatively small, which may be due to the heterogeneous image data. The highest influence of multiple scales can be seen on the Judd data set (e.g., compare ΔQDCT with QDCT).

We achieve the best single-scale as well as multiscale performance with the quaternion DCT approach. With respect to their overall performance we can rank the algorithms as follows: QDCT, EigenPQFT, EigenSR, and PQFT. With the exception of Judd's model on the Judd data set (discussed earlier), quaternion-based spectral approaches and especially QDCT are able to outperform the baseline methods on all three data sets. Furthermore, the proposed EigenPQFT is a substantial improvement over Guo's PQFT and is the 2nd best quaternion-based approach. It is also able to achieve a better performance than the low-level baseline algorithms on all three data sets. In consequence, using quaternion-based saliency detection, quaternion component weights and multiple scales, we are able to improve the state-of-the-art in predicting human gaze points.

The spectral approaches inherit the $O(N \log_2 N)$ arithmetic complexity of the discrete fast Fourier transform and in practice also benefit from highly optimized fast Fourier implementations. Thus, without going into any detail, the (quaternion) FFT- and DCT-based models that we evaluated can be implemented to operate in less than one millisecond (single-scale) on an off the shelf PC (Intel Core i5 @ 3GHz). This is an important aspect for practical applications and a fraction of the computational requirements of most other approaches such as, e.g., Judd (more than 30,000× slower on Judd's data set), CAS (more than 40,000× slower), and AIM (more than 100,000× slower).

5 Conclusion

We analyzed quaternion-based spectral saliency approaches which use the quaternion Fourier or cosine spectrum to calculate the visual saliency. Using the Toronto, Kootstra and Judd eye-tracking data sets, we are able to show the suitability of these approaches to predict human gaze in images and demonstrated the influence of multiple scales, color space weights, and quaternion transformation axes. We integrated these aspects and presented a consistent and comprehensive evaluation of spectral saliency models for predicting human eye fixations. We were able to achieve the best results for low-level approaches (i.e., without explicitly modeling the influence of higher level aspects such as, e.g., faces) on all three data sets in terms of the area under the receiver-operator characteristic curve.

In our opinion, the most important future work for spectral saliency detection would be to define an appropriate color/feature space. As can be seen in the results the color space has an important influence, but color space weighting can only be seen as an intermediate step toward a more appropriate feature space.

Acknowledgements. This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

1. Peters, R., Itti, L.: The role of fourier phase information in predicting saliency. Journal of Vision **8** (2008) 879
2. Hou, X., Harel, J., Koch, C.: Image signature: Highlighting sparse salient regions. IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012) 194–201
3. Schauerte, B., Stiefelhagen, R.: Predicting human gaze using quaternion dct image signature saliency and face detection. In: Proc. Workshop on the Applications of Computer Vision. (2012)
4. Guo, C., Ma, Q., Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In: Proc. Int. Conf. Comp. Vis. Pat. Rec. (2008)
5. Bian, P., Zhang, L.: Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In: Advances in Neural Information Processing Systems. (2009)
6. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundation: A survey. ACM Trans. Applied Perception **7** (2010) 6:1–6:39
7. Meger, D., Forssén, P.E., Lai, K., et al.: Curious george: An attentive semantic robot. Robotics and Autonomous Systems **56** (2008) 503–511
8. Schauerte, B., Kühn, B., Kroschel, K., Stiefelhagen, R.: Multimodal saliency-based attention for object-based scene analysis. In: Proc. Int. Conf. Intell. Robots Syst. (2011)
9. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: Proc. Int. Conf. Comp. Vis. Pat. Rec. (2010)
10. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Trans. Pattern Anal. Mach. Intell. (2012)
11. Ma, Z., Qing, L., Miao, J., Chen, X.: Advertisement evaluation using visual saliency based on foveated image. In: Proc. Int. Conf. Multimedia and Expo. (2009)
12. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: Proc. Int. Conf. Comp. Vis. Pat. Rec. (2007)

13. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Process. **19** (2010) 185–198
14. Achanta, R., Süsstrunk, S.: Saliency detection using maximum symmetric surround. In: Proc. Int. Conf. Image Process. (2010)
15. Li, J., Levine, M.D., An, X., He, H.: Saliency detection based on frequency and spatial domain analysis. In: Proc. British Mach. Vis. Conf. (2011)
16. Oppenheim, A., Lim, J.: The importance of phase in signals. Proc. IEEE **69** (1981) 529–541
17. Huang, T., Burnett, J., Deczky, A.: The importance of phase in image processing filters. IEEE Trans. Acoust., Speech, Signal Process. **23** (1975) 529–542
18. Bruce, N., Tsotsos, J.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision **9** (2009) 1–24
19. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: Proc. Int. Conf. Comp. Vis. (2009)
20. Kootstra, G., Nederveen, A., de Boer, B.: Paying attention to symmetry. In: Proc. British Mach. Vis. Conf. (2008)
21. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cog. Psy. **12** (1980) 97–136
22. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20** (1998) 1254–1259
23. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision **8** (2008)
24. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems. (2007)
25. Ell, T.: Quaternion-fourier transforms for analysis of two-dimensional linear time-invariant partial differential systems. In: Int. Conf. Decision and Control. (1993)
26. Sangwine, S.J.: Fourier transforms of colour images using quaternion or hypercomplex, numbers. Electronics Letters **32** (1996) 1979–1980
27. Sangwine, S., Ell, T.: Colour image filters based on hypercomplex convolution. IEEE Proc. Vision, Image and Signal Processing **147** (2000) 89–93
28. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks **19** (2006) 1395–1407
29. Gao, D., Mahadevan, V., Vasconcelos, N.: On the plausibility of the discriminant center-surround hypothesis for visual saliency. Journal of Vision **8** (2008) 1–18
30. Ell, T., Sangwine, S.: Hypercomplex fourier transforms of color images. IEEE Trans. Image Process. **16** (2007) 22–35
31. Hamilton, W.R.: Elements of Quaternions. University of Dublin Press. (1866)
32. Zhao, Q., Koch, C.: Learning a saliency map using fixated locations in natural scenes. Journal of Vision **11** (2011) 1–15
33. Judd, T., Durand, F., Torralba, A.: Fixations on low-resolution images. Journal of Vision **11** (2011)
34. Olmos, A., Kingdom, F.A.A.: A biologically inspired algorithm for the recovery of shading and reflectance images. Perception **33** (2004) 1463–1473
35. Tatler, B., Baddeley, R., Gilchrist, I.: Visual correlates of fixation selection: Effects of scale and time. Vision Research **45** (2005) 643–659
36. Geusebroek, J.M., Smeulders, A., van de Weijer, J.: Fast anisotropic gauss filtering. IEEE Trans. Image Process. **12** (2003) 938–943