Predicting Human Gaze using Quaternion DCT Image Signature Saliency and Face Detection

Boris Schauerte*
*Karlsruhe Institute of Technology
http://cvhci.anthropomatik.kit.edu/~bschauer/

Rainer Stiefelhagen*†
†Fraunhofer IOSB
rainer.stiefelhagen@kit.edu

Abstract

We combine and extend the previous work on DCT-based image signatures and face detection to determine the visual saliency. To this end, we transfer the scalar definition of image signatures to quaternion images and thus introduce a novel saliency method using quaternion type-II DCT image signatures. Furthermore, we use MCT-based face detection to model the important influence of faces on the visual saliency using rotated elliptical Gaussian weight functions and evaluate several integration schemes. In order to demonstrate the performance of the proposed methods, we evaluate our approach on the Bruce-Tsotsos (Toronto) [2] and Cerf (FIFA) [3] benchmark eye-tracking data sets. Additionally, we present evaluation results on the Bruce-Tsotsos data set of the most important spectral saliency approaches. We achieve state-of-the-art results in terms of the well-established area under curve (AUC) measure on the Bruce-Tsotsos data set and come close to the ideal AUC on the Cerf data set – with less than one millisecond to calculate the bottom-up QDCT saliency map.

1. INTRODUCTION

Attention is the cognitive process of focusing the processing of sensory information onto salient – i.e. potentially relevant and thus interesting – data. This process consists of two main mechanisms: The overt attention directs the sense organs towards salient stimuli to optimize the perception and, e.g., project an object of interest onto the fovea of the eye. The covert attention focuses the mental processing of sensory information on the salient stimuli. The latter is necessary to achieve a high reactivity despite limited computational resources and has been formally shown to be a necessary mechanism, e.g., to transform the NP-complete bottom-up perceptual search task into a computationally tractable problem (see [27]).

To enable attention mechanisms in cognitive technical systems, computational models of attention are used to determine which signal components are salient and attract the attention. But, what attracts the attention? Most importantly, primitive visual features such as motion, edges, and (color) contrast attract the attention in an early, pre-attentive stage (see, e.g., [25, 26]). Therefore, contrast measures are used to compute the saliency in the corresponding feature dimensions, e.g. center-surround differences in the orientation, intensity, and color dimension [14]. Naturally, such contrast measures are related to the spatial frequency distribution in these dimensions and in recent years this has been investigated more closely (see, e.g., [21, 25]). Furthermore, there exist more complex features that influence the attention. Most importantly, it has been shown that – independent of the subject's task – faces attract the attention [3–5].

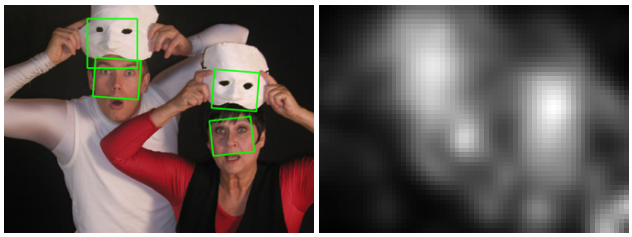

Figure 1. We use quaternion DCT image signatures and MCT face detection to calculate the visual saliency in the absence and presence of faces.

Especially in applied fields, e.g. robotics, spectral saliency approaches that operate on the Fourier and, recently, cosine (frequency) spectrum have attracted a huge interest in recent years (e.g., [10, 12, 13]). This is most likely due to their good quality in combination with their computational efficiency when compared to most other visual saliency approaches. In this contribution, we extend the recently proposed discrete cosine transform (DCT) image signature approach [12], which defines the saliency using the inverse DCT of the signs in the cosine spectrum. To this end, we use quaternions to represent and process color images in a holistic framework and, consequently, apply the quaternion DCT (QDCT) and signum function to calculate the visual saliency. Furthermore, we use the modified census transform (MCT) to detect faces and define a Gaussian

face conspicuity map to model their influence. We evaluate the proposed quaternion DCT saliency approach on the Bruce-Tsotsos data set [2] and the combination of QDCT saliency and face detections on the Cerf data set [5]. Both data sets consist of images and the respective eye-tracking data of several human subjects. Accordingly, we evaluate how well our approach predicts human eye fixations and, thus, overt attention. This makes it possible to assess the quality of the proposed model independent of any task or specific application (see, e.g., [10] or [22] for examples of task-based evaluations). As a reference, we also evaluate and present the results of the most widely applied spectral approaches on the Bruce-Tsotsos data set.

2. RELATED WORK

Throughout the last decade, computational models of attention have attracted an increasing interest in theory (e.g., [4, 14, 21, 25, 27]) and applications such as, e.g., human-robot interaction (e.g., [22]), scene exploration and analysis (e.g., [16]), and driver assistance (e.g., [17]). To this end, several computational models of attention have been developed (e.g., [1, 10–14, 30, 31]). However, reviewing them is beyond the scope of this paper. Thus, we refer the interested reader to the excellent book by Tsotsos [28] and only present work that is closely related to our approach.

The first notable visual saliency model based on the Fourier frequency spectrum was presented by Hou *et al*. in 2007 [13]. The model is based on the inverse discrete Fourier transform (DFT) of the difference between the raw and smoothed amplitude components in the spectral domain. This is related to the well-known effect that suppressing the amplitude components of signals – also known as spectral whitening – accentuates lines, edges and other narrow events [19]. In [10], Guo *et al*. demonstrated that the smoothing operation is not the essential component and instead proposed pure spectral whitening, i.e. to set a unit amplitude; which is essentially the same as the model by Peters and Itti [21] that, in contrast to [13] and [10], considers multiple scales. Furthermore, Guo *et al*. applied the quaternion DFT to realize the spectral residual for quaternion images, which reduced the required number of focus of attention shifts to find manually specified objects in test images. This was possible, because quaternions provide a holistic representation of color images in combination with a powerful algebra (see [24]), including the quaternion Fourier transform (see [6]) as well as the recently proposed quaternion cosine transform [7]. However, in contrast to Peters and Itti [21], Hou *et al*. and Guo *et al*. did not evaluate how well their approaches predict human gaze patterns. Since then, due to their good performance and computational efficiency, these models have been applied and extended, for example, for active visual search (e.g., [16]) or spoken human-robot interaction [22]. In 2011, Hou *et al*. proposed to calculate the saliency map using the inverse cosine transform of the signs of the cosine transformed image [12]. The approach was evaluated on the Bruce-Tsotsos eye-tracking data set to determine how well it predicts human eye fixations. It was reported not just to be faster than other approaches but also to outperform several established approaches such as, most notably, the famous Itti-Koch model [14], saliency using natural statistics (SUN) [30], and graph-based visual saliency (GBVS) [11]. Related to these works, we extend the DCT-based approach to operate on quaternion images and, furthermore, we evaluate how well the established spectral saliency models predict human eye fixations.

Studies have shown that – independent of the subject's task – when looking at natural images the gaze of observers is attracted to faces (see [4, 23]). Even more, there exists evidence that the gaze of infants is attracted by face-like patterns before they can consciously perceive the category of faces [23], which may play a crucial role in social processing and development (see, e.g., [15]). This early attraction and inability to avoid looking at face-like patterns suggests that there exist bottom-up attention mechanisms for faces [4]. To model this influence, Cerf *et al*. combined traditional visual saliency models (GBVS and Itti-Koch) with face detections provided by the well-known Viola-Jones detector [3, 5]. In our presented attention system, we build on this work and, firstly, use our proposed quaternion DCT saliency model and, secondly, use MCT-based face detection, which is known to provide high performance face detections in combination with a very low false positive rate in varying illumination conditions [9]. Furthermore, considering the face detections and bottom-up visual saliency as two modalities, we investigate the influence of different multimodal combination schemes (see [18]).

3. SALIENCY MODEL

3.1. The Quaternion Discrete Cosine Transform

Quaternion Algebra and Images: Quaternions form a 4-dimensional algebra $\mathbf{H}$ over the real numbers and can be thought of as an extension of the 2-dimensional complex numbers. A quaternion x is defined as $x = a + bi + cj + dk \in \mathbf{H}$ with $a, b, c, d \in \mathbf{R}$, where i, j, and k ($i^2 = j^2 = k^2 = ijk = -1$) provide the necessary basis to define a product in $\mathbf{H}$. The corresponding Hamilton product of two quaternions x_1 and x_2 is defined as:

$$
\begin{aligned}
x_1 x_2 = \; & (a_1 + b_1 i + c_1 j + d_1 k)(a_2 + b_2 i + c_2 j + d_2 k) && (1)\\
= \; & a_1 a_2 - b_1 b_2 - c_1 c_2 - d_1 d_2 \\
& + (a_1 b_2 + b_1 a_2 + c_1 d_2 - d_1 c_2) i \\
& + (a_1 c_2 - b_1 d_2 + c_1 a_2 + d_1 b_2) j \\
& + (a_1 d_2 + b_1 c_2 - c_1 b_2 + d_1 a_2) k \,. && (2)
\end{aligned}
$$

Most notably, the Hamilton product is not commutative (e.g., note that by definition $ij = k$ while $ji = -k$). Thus, when the Hamilton product is involved, we have to consider left-sided and right-sided operations in the following (marked by L and R, respectively). A quaternion x is called real, if $x = a + 0i + 0j + 0k$, and pure imaginary, if $x = 0 + bi + cj + dk$. If we multiply an arbitrary quaternion with a real quaternion $x_3 = y$, we obtain a simple element-wise product, i.e.:

$$x_1 x_3 = a_1 y + b_1 y i + c_1 y j + d_1 y k\,. \tag{3}$$

As for complex numbers, we can define conjugate quaternions $\bar{x} = a - bi - cj - dk$ as well as the norm $|x| = \sqrt{x \cdot \bar{x}}$.

Naturally, we can represent every $M \times N \times C$ image with less than 4 channels $I_c, C \leq 4$ as a quaternion image I_Q:

$$\begin{aligned} I_\mathrm{Q} &= I_4 + I_1 i + I_2 j + I_3 k & (4)\\ &= I_4 + I_1 i + (I_2 + I_3 i) j \quad \text{(Cayley-Dickson form)} & (5) \end{aligned}$$

We represent the 4th image channel as the scalar part, because then we obtain a pure imaginary quaternion matrix for color spaces with less than 4 channels.

The 2-D Quaternion DCT: Following the definition of the quaternion DCT in [7], we can transform the $M \times N$ quaternion matrix I_Q:

$$\mathrm{QDCT}^\mathrm{L}(p,q) = \alpha_p^M \alpha_q^N \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} u_\mathrm{Q} I_\mathrm{Q}(m,n) \beta_{p,m}^M \beta_{q,n}^N \tag{6}$$

$$\mathrm{QDCT}^\mathrm{R}(p,q) = \alpha_p^M \alpha_q^N \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_\mathrm{Q}(m,n) \beta_{p,m}^M \beta_{q,n}^N u_\mathrm{Q}\,, \tag{7}$$

where u_Q is a unit (pure) quaternion, i.e. $u_\mathrm{Q}^2 = -1$, that serves as DCT axis. In accordance with the definition of the traditional type-II DCT, we define α and N as follows[1]:

$$\alpha_p^M = \begin{cases} \sqrt{\frac{1}{M}} & \text{for} \quad p = 0 \\ \sqrt{\frac{2}{M}} & \text{for} \quad p \neq 0 \end{cases} \tag{8}$$

$$\beta_{p,m}^M = \cos\left[\frac{\pi}{M}(m + \frac{1}{2})p\right]. \tag{9}$$

Consequently, the corresponding inverse quaternion DCT is defined as follows:

$$\mathrm{IQDCT}^\mathrm{L}(m,n) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p^M \alpha_q^N u_\mathrm{Q} C_\mathrm{Q}(p,q) \beta_{p,q}^M \beta_{m,n}^N \tag{10}$$

$$\mathrm{IQDCT}^\mathrm{R}(m,n) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p^M \alpha_q^N C_\mathrm{Q}(p,q) \beta_{p,q}^M \beta_{m,n}^N u_\mathrm{Q}\,. \tag{11}$$

The choice of the axis u_Q is arbitrary (see [6]) and we will use $u_\mathrm{Q} = -\sqrt{1/3}i - \sqrt{1/3}j - \sqrt{1/3}k$ in the following.

[1] It is not necessary to define the case in α that handles $p = 0$ (it makes the DCT-II matrix orthogonal). Even more, it is possible to operate without normalization, which results – irrelevant for saliency – in a scale change.

Figure 2. Quaternion CIE Lab image channels before (top; left-to-right: original color image, L, a, b) and after the quaternion axis transform $u_\mathrm{Q} I_\mathrm{Q}$ with $u_\mathrm{Q} = -\sqrt{1/3}i - \sqrt{1/3}j - \sqrt{1/3}k$ (bottom; left-to-right: scalar, i, j, k).

Implementation and Runtime Aspects: In the following, we will use the left-sided QDCT^L and $\mathrm{IQDCT}^\mathrm{L}$ as default QDCT and IQDCT, respectively, unless stated otherwise. Since α and β are real, the Hamilton product is drastically simplified (see Eq. 3). Thus, the additional runtime required to compute the quaternion DCT – compared to the traditional real DCT – is relatively low, because the only computationally complex quaternion operations are the axis multiplications $u_\mathrm{Q} I_\mathrm{Q}$ and $u_\mathrm{Q} C_\mathrm{Q}$ for the QDCT^L and $\mathrm{IQDCT}^\mathrm{L}$, respectively (see Fig. 2). Considering the relatively low-resolution saliency maps that are common for spectral saliency approaches, the additional runtime for the axis multiplication is relatively low. Furthermore, we can – basically – interpret $u_\mathrm{Q} I_\mathrm{Q}$ and $u_\mathrm{Q} C_\mathrm{Q}$ as real 4-channel image (see Eq. 4 and 5) and use 4 real DCTs and IDCTs, respectively, to take advantage of highly optimized DCT or DFT implementations (see also [20]) such as, for example, provided by the FFTW [8].

3.2. Quaternion DCT Image Signatures

DCT Signatures: The visual saliency based on DCT image signatures S_DCT for a multi-channel image I is defined as follows [12, Sec. II]:

$$\begin{aligned} S_\mathrm{DCT}(I) &= g * \sum_c [T(I_c) \circ T(I_c)] \quad \text{with} & (12)\\ T(I_c) &= \mathrm{IDCT}(\mathrm{sgn}(\mathrm{DCT}(I_c))\,, & (13) \end{aligned}$$

where I_c is the c'th image channel, $\circ$ denotes the Hadamard – i.e. element-wise – product, sgn is the signum function, and g is typically a Gaussian smoothing filter. Most notably, it has been formally shown that the DCT image signatures, i.e. $\mathrm{sgn}(\mathrm{DCT}(I_c))$, suppress the background and are likely to highlight sparse (salient) features and objects [12].

Quaternion DCT Image Signatures: The signum function for quaternions can be considered as the quaternion "direction" and is defined as follows:

$$\mathrm{sgn}(x) = \begin{cases} \frac{x_0}{|x|} + \frac{x_1}{|x|} i + \frac{x_2}{|x|} j + \frac{x_3}{|x|} k & \text{for} \quad |x| \neq 0 \\ 0 & \text{for} \quad |x| = 0\,. \end{cases} \tag{14}$$

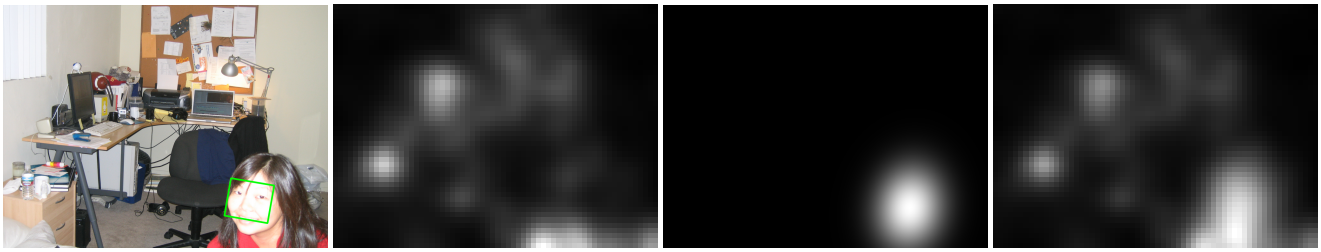

Figure 3. An example for the combined saliency model (the image is part of the Cerf data set). Left-to-right: the image with marked face detections, the quaternion DCT saliency map, the face conspicuity map, and the resulting linearly combined saliency map.

Given that definition, we can now easily transfer the single-channel definition of the DCT signature and derive the visual saliency S_{QDCT} using the quaternion DCT signature

$$S_{\mathrm{QDCT}}(I_{\mathrm{Q}}) = g * [T(I_{\mathrm{Q}}) \circ \bar{T}(I_{\mathrm{Q}})] \quad \text{with} \quad (15)$$

$$T(I_{\mathrm{Q}}) = \mathrm{IQDCT}^{\mathrm{L}}(\mathrm{sgn}(\mathrm{QDCT}^{\mathrm{L}}(I_{\mathrm{Q}}))) \,. \quad (16)$$

3.3. Face Detection and the Face Conspicuity Map

In [5] and [3], each detected face is modeled in the face conspicuity map by a circular 2-D Gaussian weight function with the standard deviation $\sigma = \sqrt{(w+h)/4}$, where w and h is the width and height, respectively, of the Viola-Jones face detection's bounding box. We extend this model in two ways: First, we allow an in-plane rotation θ of the face bounding boxes provided by the MCT detectors. Then, we use an elliptical 2-D Gaussian weight function g_0, where σ_u and σ_v is the standard deviation in the direction parallel and orthogonal, respectively, to the orientation θ:

$$g_0(u,v,\sigma_u,\sigma_v) = \frac{1}{\sqrt{2\pi}\sigma_u} \exp\left\{-\frac{1}{2}\frac{u^2}{\sigma_u^2}\right\} * \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left\{-\frac{1}{2}\frac{v^2}{\sigma_v^2}\right\}, \quad (17)$$

where the u-axis corresponds to the direction of θ and the v-axis is orthogonal to θ, i.e.

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \hat{\theta}(x) \\ \hat{\theta}(y) \end{pmatrix} . \quad (18)$$

Accordingly, we can calculate the face conspicuity map S_{F}

$$S_{\mathrm{F}}(x,y) = \sum_{1 \le i \le N} g_0(\hat{\theta}(x-x_i), \hat{\theta}(y-y_i), \sigma_{u,i}, \sigma_{v,i}, \theta) \,, \quad (19)$$

where (x_i, y_i) is the detected center of face i with orientation θ_i and the standard deviations $\sigma_{u,i}$ and $\sigma_{v,i}$. Since, the width and height of the bounding box may not be directly equivalent to the optimal standard deviation, we calculate σ_u and σ_v by scaling w and h with the scale factors s_w and s_h that we experimentally determined for our MCT detectors.

3.4. Multimodal Integration

Interpreting the calculated visual saliency map S_{Q} and the face detections represented in S_{F} as two separate modalities, we have to consider several (biologically) plausible multimodal integration schemes (see [18]):

Linear: We can use a linear combination

$$S_{+} = w_{\mathrm{Q}} S_{\mathrm{Q}} + w_{\mathrm{F}} S_{\mathrm{F}} \quad (20)$$

as applied in [5] and [3]. In contrast to [3], we analyze the weight space in order to determine weights that provide optimal performance in practical applications. Therefore, we normalize the value range of the saliency map S_{Q} and use a convex combination, i.e. $w_{\mathrm{Q}} + w_{\mathrm{F}} = 1$ with $w_{\mathrm{F}}, w_{\mathrm{S}} \in [0,1]$. From an information-theoretical point of view, the linear combination is optimal in the sense that the information gain equals the sum of the unimodal information gains [18].

Sub-Linear (Late Combination): In late combination schemes, no true cross-modal integration occurs. Instead, the candidate fixation points from the unimodal saliency maps compete against each other. Given saliency maps, we can use the maximum to realize such a late combination scheme, resulting in a sub-linear combination:

$$S_{\max} = \max\{S_{\mathrm{Q}}, S_{\mathrm{F}}\} \,. \quad (21)$$

Supra-Linear (Early Interaction): Early interaction assumes that there has been cross-modal sensory interaction at an early stage, before the saliency computation and focus of attention selection, which imposes an expansive nonlinearity. As alternative model, this can be realized using a multiplicative integration of the unimodal saliency maps:

$$S_{\circ} = S_{\mathrm{Q}} \circ S_{\mathrm{F}} \,. \quad (22)$$

Quaternion Face Channel: From a technical perspective, if the image's color space has less than 4 channels, we can also use the quaternion scalar part to explicitly represent faces and obtain an integrated saliency map S_{QF}:

$$S_{\mathrm{QF}} = S_{\mathrm{QDCT}}(I_{\mathrm{QF}}) \quad \text{with} \quad (23)$$

$$I_{\mathrm{QF}} = S_{\mathrm{F}} + I_{\mathrm{Q}} = S_{\mathrm{F}} + I_1 i + I_2 j + I_3 k \,. \quad (24)$$

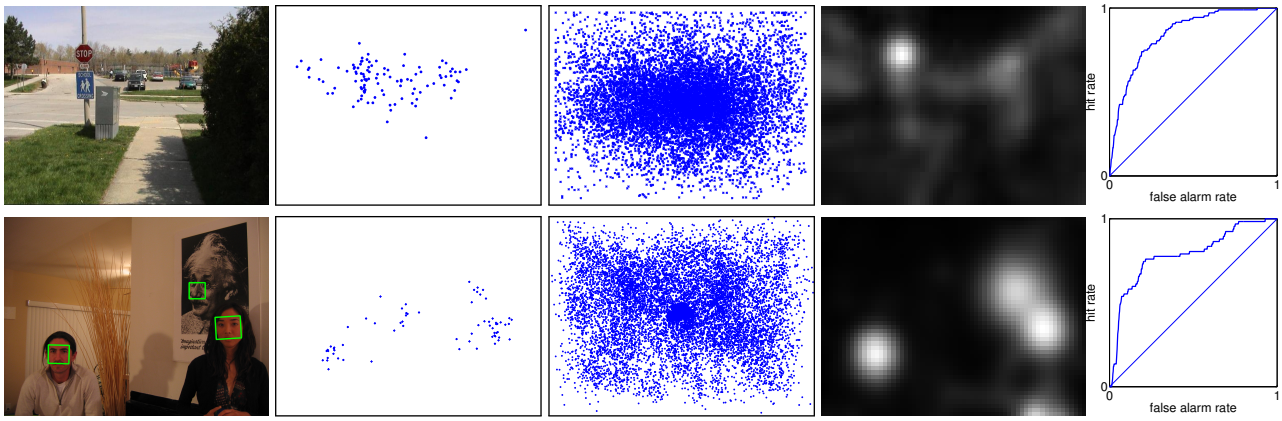

Figure 4. Example images from the Bruce-Tsotsos (top row) and Cerf (bottom row) data set. Left-to-right: image, human eye fixations (positive samples), combined eye fixations for the other images in the data set (negative samples), quaternion DCT saliency combined with the face map, and the ROC curve.

4. EVALUATION

4.1. Algorithms, Data Sets, and Measures

Algorithms: In [12], it has been shown that DCT image signatures outperform the currently leading visual saliency models in predicting human eye fixations; most importantly, the classical Itti-Koch model (Itti) [14], Graph-based Visual Saliency (GBVS) [11], Attention based on Information Maximization (AIM) [1], and Saliency using Natural Statistics (SUN) [30]. Thus, in order to assess the performance of the proposed quaternion DCT image signatures (QDCT), we compare it to DCT image signatures (DCT) [12] and, extending the experimental results reported in [12], to the most widely used spectral approaches. The additional spectral algorithms are: spectral whitening (PFT) [10], quaternion-based spectral whitening (PQFT) [10], and spectral residual (SR) [13]. Additionally, we evaluate the publicly available implementation of Itti and GBVS, because they have been applied in combination with face detection in [11] and, thus, serve as reference in that context.

Since the color space has a considerable impact on the performance of the spectral saliency approaches (see [12]), we will perform the evaluation of the spectral algorithms for the Red-Green-Blue (RGB), CIE L*a*b* (Lab), and Intensity and Red-Green/Blue-Yellow Color Opponents (ICOPP) color space (see [10]). The latter was proposed in [10] for usage with PQFT and PFT and in [12] it was shown that Lab provides a better performance than RGB in combination with DCT image signatures.

For all spectral saliency algorithms, the saliency map is calculated at an image resolution of 64×48. Images and saliency maps are represented using double precision to avoid the influence of quantization and varying value ranges.

Data Sets: In order to evaluate the algorithms in the absence of faces, we use the Bruce-Tsotsos data set [2], see Fig. 4. The data set contains 120 images (681×511 px) with eye-tracking data of 20 subjects. The subjects had no assigned task, i.e. "free-viewing", and saw each image for 4 seconds. To evaluate the performance of our approach in the presence of faces, we use the Cerf data set [3], see Fig. 4. The data set consists of eye-tracking data (2 seconds, "free-viewing") of 9 subjects for 200 (1024×768 px) images of which 157 contain one or more faces. Additionally, the data set provides perfect annotations of the location and size of faces in the images, which we use to evaluate the influence between perfect, i.e. manual, and automatic face detection.

AUC Measure: We use the well-known receiver operating characteristic (ROC) area under curve (AUC) measure (see, e.g., [12, 25, 30]) to assess the performance of the saliency algorithms[2]. Each saliency map can be thresholded and then considered to be a binary classifier that separates positive samples – composed of the fixation points of all human subjects on that image – from negative samples – composed of the fixations of all subjects on all other images in the data set, which accounts for the center bias [25] –, see Fig. 4. Accordingly, we can sweep over all thresholds to calculate the ROC curve for each saliency map and then calculate the area beneath the ROC curve. The area under the curve (AUC) provides a good measure to assess how ac-

[2]Please note that our results differ (slightly) from the reported results in [12, Table II]. However, we were in contact with the authors and discussed our results. Our evaluation measure implementation is in accordance with the benchmark implementation by A. Borji and L. Itti that is available at https://sites.google.com/site/saliencyevaluation. Furthermore, please note that we report the – center bias corrected – AUC and not the normalized AUC (nAUC), which may lead to confusion and needs to be considered when reading and interpreting the results reported in some other papers. The nAUC may be calculated by dividing the calculated AUC by the ideal AUC [31].

curately the saliency map predicts the human eye fixations on the image. An AUC value of 0.5 corresponds to chance, a value > 0.5 indicates positive correlation, and 1.0 corresponds to a perfect prediction of eye fixations.

As a performance baseline for each data set, we can calculate an ideal AUC that measures how well the fixations of one subject can be predicted by the fixations of the other subjects. The ideal AUC for the Bruce-Tsotsos and Cerf data set is 0.878 and 0.786, respectively (see [31] and [3]).

4.2. Results on the Bruce-Tsotsos data set

As can be seen in Tab. 1, QDCT Lab signatures provides competitive if not higher performance – highest average AUC and lowest AUC standard deviation – than the evaluated competitors. Furthermore, as the other spectral approaches, it outperforms established models such as GBVS, AIM, Itti, and SUN. However, the mean AUCs of the spectral approaches (QDCT, DCT, PQFT, PFT, and SR) are at a very close range, which is not surprising, because they are all based on (roughly) the same principle. Interestingly, it can be seen that the spectral residual does have a positive influence compared to pure spectral whitening. But, to our surprise, the visual saliency based on the quaternion Fourier transform (PQFT) is considerably weaker, which is even more surprising in the light of fact that quaternion DCT Lab signatures achieve the best results. In our opinion, the most likely cause for these results is the more complex quaternion basis change and complex DFTs that are necessary to compute the quaternion Fourier transform.

As can be seen in Tab. 1, the Lab color space is the foundation to achieve the best performance of each spectral saliency model. QDCT saliency achieves the best results among the spectral models for the RGB and Lab color space. It fails in doing so for ICOPP[3] where the Spectral Residual achieves the best results. This could be due to the axis transform, which in case for RGB leads to $u_Q I_Q \simeq (R+G+B) - (G+B)i - (R+B)j + (R-G)k$. Thus, we have the intensity in the scalar part and linear combinations of the channels in the vector part, which is similar to ICOPP and could explain the results for RGB and ICOPP. But, QDCT on RGB achieves higher values than any algorithm on the ICOPP color space. For practical applications, this makes QDCT-RGB an interesting choice, because it operates on RGB and avoids explicit color space conversions.

Method	Color Space	AUC (mean)	AUC (std)	σ
QDCT	Lab	**0.7183**	0.0856	0.0438
DCT [12]	Lab	0.7124	0.0919	0.0438
PQFT [10]	Lab	0.6978	0.0956	0.0422
PFT [10]	Lab	0.7135	0.0891	0.0383
SR [13]	Lab	0.7159	0.0908	0.0398
QDCT	ICOPP	0.7041	0.0909	0.0352
DCT [12]	ICOPP	0.7007	0.0913	0.0359
PQFT [10]	ICOPP	0.6796	0.0989	0.0453
PFT [10]	ICOPP	0.7026	0.0898	0.0352
SR [13]	ICOPP	**0.7052**	0.0897	0.0414
QDCT	RGB	**0.7061**	0.0907	0.0391
DCT [12]	RGB	0.6923	0.0941	0.0391
PQFT [10]	RGB	0.6882	0.0986	0.0414
PFT [10]	RGB	0.6960	0.0961	0.0422
SR [13]	RGB	0.6981	0.0932	0.0383
GBVS[2] [11]		0.6718	0.1019	0.0221
Itti[2] [14]		0.6488	0.1106	0.0383
AIM† [1]		0.7000		
GBVS† [11]		0.6782		
SUN† [30]		0.6751		
Itti† [14]		0.6524		
Ideal		0.878		

Table 1. AUC performance of the evaluated algorithms on the Bruce-Tsotsos data set [2] (mean and standard deviation). σ denotes the standard deviation of the applied Gaussian smoothing filter. (†): as reported by Hou *et al.* in [12] for the optimal σ.

4.3. Results on the Cerf data set

Now we can use the optimal parameters determined in Sec. 4.2 – most importantly σ and the color space – to evaluate the influence of faces, see Tab. 2. Since GBVS was reported to perform better than Itti-Koch when combined with face detections [5], we compare our system to GBVS.

It can be seen in Tab. 2 that the performance on the Cerf data set without the face detections is notably worse than the performance on the Bruce-Tsotsos data set, which was expected due to the faces attracting the visual focus of attention. Furthermore, the face conspicuity map has a considerable predictive power with an AUC of 0.659, which has already been reported in [5]. Interestingly, we can also observe the phenomenon that the AUC is higher (0.665) when using automatic face detection (Face$^{\text{MCT}}$) instead of optimal bounding boxes calculated from the manually annotated face regions (Face*). This can be explained by the fact that false positives usually occur on complex image patches that are also likely to attract the attention. The linear combination of the bottom-up visual saliency and the face conspicuity map significantly increases the results: from an AUC of 0.704 to 0.769 for QDCT and from 0.664 to 0.727 for GBVS, using the annotated face regions (Face*). If we use MCT-based face detections (Face$^{\text{MCT}}$) instead, the AUC for QDCT becomes 0.764 with our proposed scaled elliptical Gauss model (Proposed) and 0.754 with the circular Gauss model by Cerf *et al.* (Cerf). The performance difference can be explained by the false negative and false posi-

[3] ICOPP (I, RG, BY) for an RGB image is calculated as follows: Intensity $I = (r+g+b)/3$, Red-Green $RG = [r-(g+b)/2] - [g-(r+b)/2]$, and Blue-Yellow $BY = [b-(r+g)/2] - [(r+g)/2|r-g|/2-b]$.

Method	Face Model	AUC (mean)	AUC (std)
QDCT-Lab + Face*	Proposed	**0.7693**	0.0864
QDCT-Lab + Face*	Cerf	0.7676	0.0902
GBVS + Face*	Proposed	0.7274	0.1031
GBVS + Face*	Cerf	0.7268	0.1030
QDCT-Lab + Face^{MCT}	Proposed	**0.7639**	0.0884
QDCT-Lab + Face^{MCT}	Cerf	0.7540	0.0851
GBVS + Face^{MCT}	Proposed	0.7218	0.0966
GBVS + Face^{MCT}	Cerf	0.7132	0.0968
Face*	Proposed	0.6597	0.1160
Face*	Cerf	0.6593	0.1153
Face^{MCT}	Proposed	**0.6648**	0.1178
Face^{MCT}	Cerf	0.6368	0.1040
QDCT-Lab		0.7044	0.1071
GBVS		0.6641	0.1022
Ideal		0.786	

Table 2. AUC performance of the evaluated algorithms (with optimal parameters) on the Cerf data set [3] (mean and standard deviation). (*): manual annotations used as face detections.

tive – but, consider the effect observed for the unimodal face map – MCT face detections, since it remains roughly equivalent for GBVS. The only minor performance difference between the face conspicuity map models for optimal bounding boxes (Face*) suggests that the elliptical model only provides a minor advantage over the circular model. However, the performance difference when using MCT-based face detections suggests that the scaling has a considerable influence on the results. Finally, if we use the ideal ROCs and AUC to calculate the normalized AUC (nAUC), we obtain an nAUC of 0.972 which is also higher than the recently reported 0.962 that was obtained using an optimally weighted Itti-Koch with center bias model [31, see Table 1].

The chosen multimodal integration scheme (see Sec. 3.4) has a considerable influence on the performance, see Fig. 5. The linear combination achieves the best performance with an AUC of 0.769 at $w_Q = 0.62$ for QDCT and 0.727 for GBVS ($w_Q = 0.75$). This is closely followed by the maximum operator (late combination) that achieves an AUC of 0.761 and 0.719 for QDCT and GBVS, respectively. Consequently, especially for practical applications, we propose the use of the weighted scheme, because it provides a higher performance than max for a relatively large value range of w_Q (see Fig. 5). However, this close performance on a single data set makes it hard to conclusively decide between the late and linear combination scheme. Although integrating the face conspicuity map in the quaternion image does not perform equally well with an AUC of 0.721, it is better than unimodal and outperforms the supra-linear combination that achieves an AUC of 0.658 and 0.661 for QDCT and GBVS, respectively. Thus, this combination performs worse than each unimodal map, which could be expected,

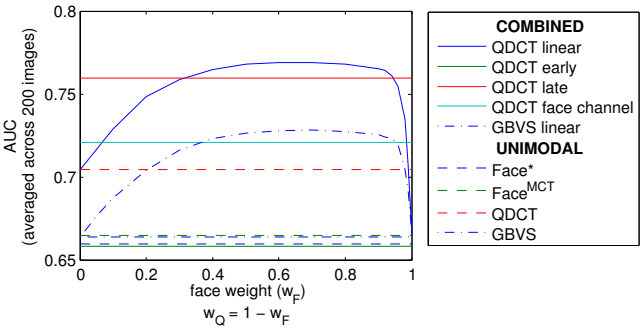

Figure 5. Illustration of the average AUC in dependency of the chosen multimodal integration (see Sec. 3.4) on the Cerf data set [3] using manually annotated face regions and our scaled elliptical face model to calculate the combined saliency maps.

because it implies a logical "and".

4.4. Runtime Aspects

In practical applications, we use a hard-coded 64×48 real DCT-II and DCT-III – the latter is used to calculate the inverse – implementation and are able to calculate the bottom-up saliency map S_{QDCT} in approximately 0.4 ms (excluding the time for resizing, which depends on the raw image resolution, but including the time for anisotropic Gauss filtering) on an Intel Core i5 with 2.67 Ghz (single-threaded; double-precision). This makes our implementation around factor $20 - 50\times$ faster than previously reported for spectral approaches (see [12] and [10]) and (substantially) faster than sophisticated implementations of most other approaches, e.g. the Neuromorphic Vision C++ Toolkit (NVT) Itti-Koch model reference implementation or – especially – its multi-GPU implementation by T. Xu *et al.* (see [29, Table II]), while providing a state-of-the-art quality (see Tab. 1 and 2). Please note that our implementation is publicly available and free (BSD License).

5. CONCLUSION

We presented a novel approach to determine the visual saliency of an image using quaternion DCT signatures and MCT face detection. To this end, we introduced the type-II quaternion DCT and demonstrated how to transfer the scalar, real DCT signatures to quaternion images. Then, we integrated information about detected faces in the saliency map, because – as discussed – the presence of faces has a considerable influence on the visual saliency. To demonstrate the performance of the proposed bottom-up quaternion DCT saliency approach in the absence of faces, we evaluated the approach as well as the most widely applied spectral saliency algorithms on the Bruce-Tsotsos eye-tracking data set. Furthermore, we evaluated the performance of our approach in the presence of faces on the Cerf eye-tracking data set and, additionally, evaluated sev-

eral multimodal combination schemes. In summary, on both data sets we were able to achieve higher results in predicting where humans look than previously reported.

Acknowledgments

This work is supported by the German Research Foundation (DFG) within the Collaborative Research Program SFB 588 "Humanoide Roboter". The authors thank M. Fischer for providing his MCT detector implementation. Furthermore, the authors would like to thank X. Hou for the helpful discussion about the results on the Bruce-Tsotsos data set. Last but not least, the authors thank S. J. Sangwine for his time and an insightful discussion about the QDCT.

References

[1] N. Bruce and J. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24, 2009. 138, 141, 142

[2] N. D. B. Bruce and J. K. Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems*, pages 155–162, 2006. 137, 138, 141, 142

[3] M. Cerf, E. P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9, 2009. 137, 138, 140, 141, 142, 143

[4] M. Cerf, P. Frady, and C. Koch. Subjects' inability to avoid looking at faces suggests bottom-up attention allocation mechanism for faces. In *Soc. Neurosci.*, 2008. 137, 138

[5] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems*, 2007. 137, 138, 140, 142

[6] T. Ell and S. Sangwine. Hypercomplex fourier transforms of color images. *IEEE Trans. Image Process.*, 16(1):22–35, 2007. 138, 139

[7] W. Feng and B. Hu. Quaternion discrete cosine transform and its application in color template matching. In *Int. Cong. Image and Signal Processing*, pages 252–256, 2008. 138, 139

[8] M. Frigo and S. G. Johnson. The design and implementation of FFTW3. *Proc. IEEE*, 93(2):216–231, 2005. 139

[9] B. Fröba and A. Ernst. Face detection with the modified census transform. In *Proc. Int. Conf. Auto. Face Gesture Rec.*, 2004. 138

[10] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2008. 137, 138, 141, 142, 143

[11] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, 2007. 138, 141, 142

[12] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 99, 2011. 137, 138, 139, 141, 142, 143

[13] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2007. 137, 138, 141, 142

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998. 137, 138, 141, 142

[15] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch. Gen. Psychiatry*, 59(9):809–816, 2002. 138

[16] D. Meger, P.-E. Forssén, et al. Curious George: An attentive semantic robot. In *IROS Workshop: From sensors to human spatial concepts*, 2007. 138

[17] T. Michalke, J. Fritsch, and C. Goerick. A biologically-inspired vision architecture for resource-constrained intelligent vehicles. *Computer Vision and Image Understanding*, 114(5):548–563, 2010. 138

[18] S. Onat, K. Libertus, and P. König. Integrating audiovisual information for the control of overt attention. *Journal of Vision*, 7(10), 2007. 138, 140

[19] A. Oppenheim and J. Lim. The importance of phase in signals. *Proc. IEEE*, 69(5):529–541, 1981. 138

[20] 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.*, 49(11):2783–2797, nov 2001. 139

[21] R. Peters and L. Itti. The role of fourier phase information in predicting saliency. *Journal of Vision*, 8(6):879, 2008. 137, 138

[22] B. Schauerte and G. A. Fink. Focusing computational visual attention in multi-modal human-robot interaction. In *Proc. Int. Conf. Multimodal Interfaces*, 2010. 138

[23] C. Simion and S. Shimojo. Early interactions between orienting, visual sampling and decision making in facial preference. *Vision Research*, 46(20):3331–3335, 2006. 138

[24] O. N. Subakan and B. C. Vemuri. A quaternion framework for color image smoothing and segmentation. *Int. J. Comp. Vis.*, 91:233–250, 2011. 138

[25] B. Tatler, R. Baddeley, and I. Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005. 137, 138, 141

[26] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cog. Psy.*, 12(1):97–136, 1980. 137

[27] J. K. Tsotsos. Behaviorist intelligence and the scaling problem. *Artif. Intell.*, 75:135–160, 1995. 137, 138

[28] J. K. Tsotsos. *A Computational Perspective on Visual Attention*. The MIT Press, 2011. 138

[29] T. Xu, T. Pototschnig, et al. A high-speed multi-GPU implementation of bottom-up attention using CUDA. In *Proc. Int. Conf. Robot. Autom.*, 2009. 143

[30] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008. 138, 141, 142

[31] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3):1–15, 2011. 138, 141, 142, 143