



# The Assessment of Sencogi: A Visual Complexity Model Predicting Visual Fixations

Maria Laura Mele<sup>1,2,3(✉)</sup>, Silvia Colabrese<sup>1</sup>, Luca Calabria<sup>1</sup>,  
Damon Millar<sup>1</sup>, and Christiaan Erik Rijnders<sup>1</sup>

<sup>1</sup> COGISEN Engineering Company, Rome, Italy

{marialaura, silvia, luca, damon, chris}@cogisen.com

<sup>2</sup> Department of Philosophy, Social and Human Sciences and Education,  
University of Perugia, Perugia, Italy

<sup>3</sup> ECONA, Interuniversity Centre for Research on Cognitive Processing in  
Natural and Artificial Systems, Sapienza University of Rome, Rome, Italy

**Abstract.** This paper investigates whether a frequency-domain model of complexity can accurately predict human visual saliency maps. The Sencogi model uses the frequency domain to calculate maps of spatial (i.e., static) and temporal (i.e., dynamic) complexity. This study compares the complexity maps generated by Sencogi to human fixation maps obtained during a visual quality assessment task on static images. This work is the first part of an ongoing multi-step study designed to assess whether fixation maps are an accurate representation of saliency for spatio-temporal scenes. A supporting experiment confirmed that top-down factors, such as scene type, task or emotional states, did not affect human fixation maps. Results show that the Sencogi visual complexity model estimates human eye fixations of images with prediction scores that are significantly above a Chance baseline and is able to compete with a Single Observer baseline. We conclude that the Sencogi visual complexity model is able to predict human fixations in the spatial domain. The next studies will focus on the assessment of Sencogi's performance predicting visual fixations in the spatio-temporal domain.

**Keywords:** Visual complexity · Visual saliency ·  
Spatio-temporal saliency models · Computer vision

## 1 Introduction

Visual saliency is “an abstract representation of attentional priority” [1, 2]. The literature commonly uses the term “saliency” as a synonym for the total time of foveal fixations. Most computational saliency models are designed as an estimate of the overall probability that each image area influences fixations. The close relationship between fixation length and visual saliency has led to a common methodology to assess visual saliency models, by using human fixation maps as a ground truth to estimate image saliency [3].

Visual saliency processes are driven by both bottom-up and top-down perceptual processes. Bottom-up saliency is data-driven, i.e., it is based on the intrinsic features of

visual scenes and is an automatic and implicit perceptual process. Top-down saliency is guided by cognitive processes rather than the features of visual stimuli, meaning that environmental and psychological components (e.g., nature of the task, context, previous-knowledge or expectations) guide the spatio-temporal deployment of eye fixations and attention over visual scenes [4]. State of the art saliency models mostly embrace the bottom-up approach to visual saliency.

Computational saliency maps are usually compared with human fixation maps created by free-viewing tasks [5]. However, free-viewing tasks are not able to control top-down influences because there is no explicit task, thus causing idiosyncratic gaze behaviors depending on subjective factors such as expectancies, mood, language and gender [6, 7]. In a study by Köehler and colleagues, the authors studied the effects of both free viewing and targeted tasks on the ability of saliency models to predict human fixations. They showed that saliency models more accurately predicted human fixations on judgement tasks than the widely used free viewing tasks [8].

Visual saliency prediction is a topic of broad and current interest in the field of computational neuroscience and computer vision. The main challenge of visual saliency prediction is to develop architectures that are able to model how human attention behaves in visual scenes. More accurate visual saliency prediction models would benefit fields such as image and video compression [4], as in the case of the Sencogi model. Most saliency models for compression aim to predict visual saliency of images and videos from the spatial properties of the visual frames [9]. These models consider only the saliency properties within a video frame (i.e., spatial saliency), thus excluding the influence of motion between frames in saliency (spatio-temporal saliency). Until now, spatio-temporal saliency models have been mainly proposed in the cognitive science research field for modelling perceptual processes (e.g., [10]), and in spectral analysis research for extending frequency domain use of phase data (e.g., [11]).

Spatial models for video saliency do not take into account the fact that regions of consecutive frames have a spatio-temporal correlation that may affect saliency (e.g., a moving object has higher salience) [12]. Spatio-temporal models that process both within-frames and between-frames information in dynamic visual scenes are rarely used for compression applications (e.g., [12–16]) because discriminating salient motion from noise is still one of the most challenging issues in this field [17].

This work is part of an ongoing multi-step study for the assessment of Sencogi. It describes the assessment of the spatial component of Sencogi on static images. Sencogi is a bottom-up spatio-temporal model for visual compression applications (see [19–21]). The model calculates visual complexity maps using a frequency-domain remapping of the visual complexity of blocks of pixels. The aim of the model is to predict visual fixations on scenes in the spatial and temporal domain. The model has been validated on subjective quality scores with both images and videos for compression applications [19–21]. In this study, the evaluation of the spatial component of Sencogi model was conducted by comparing Sencogi visual complexity maps with human fixation maps derived from a judgment task on an image database. Commonly used metrics for the assessment of saliency models were used to compare Sencogi's maps with human fixation maps. Two additional models were computed to compare the Sencogi model performance with (i) a Chance baseline with visual fixations maps

randomly assigned and (ii) a Single Observer baseline with visual fixations maps as predicted by a model computing a single observer performance.

Saliency maps are affected by top-down factors such as interest and mood. Some of the factors that influence top-down salience can be measured using biometric sensors, to determine whether the top-down motivations of all participants are similar. Eye fixations, EEG correlates of decision making and motivation, and facial expressions of emotional states have been studied. Frontal EEG asymmetry in alpha oscillations is an important decision-making marker related to approach/withdraw processes [22]. The Frontal Alpha Asymmetry (FAA) power indicates subjects' approach/withdraw from unexpected or affective stimuli. Facial expressions coded by the Facial Action Coding System (FACS) [23] are patterns of involuntary muscular movements related to affective valence (positive, negative and neutral) and the seven universal basic emotions proposed by Ekman, i.e., joy, sadness, anger, contempt, disgust, fear, and surprise [23].

### 1.1 Sencogi: A Visual Complexity Model

Sencogi is a frequency-domain algorithm which extracts from input images (or videos) a number of features and combines them linearly. It models the variations of brightness between pixels of a visual frame in order to create a map called a "visual complexity map". Sencogi's visual complexity map represents predictions of which parts of the input image is perceptually relevant and might be affected by visual distortions if heavily compressed. The visual complexity model is able to estimate the degree that a portion of the input image is visually relevant. Visual complexity maps can be used to compress images, by removing data from areas that have low visual complexity.

In order to avoid visual degradation due to information removal, Sencogi's architecture is comprised of four independent visual complexity maps. They are simultaneously computed on the real-time visual scene stream and combined to drive the compression task. The four visual complexity maps are: (1) "Static image visual complexity", measuring visual complexity within each video frame. (2) "Spatio-temporal saliency", measuring visual complexity between multiple frames. (3) "Delta-quality", that detects when quality changes to the image quality introduce artefacts that might be subjectively perceived by the human visual system, thus affecting scene saliency [24]. (4) "Pixel noise detection" that differentiates between scene motion and sensor pixel noise.

The four visual complexity maps are weighted by tunable thresholds, and then combined to return a final visual complexity map driving compression. In this paper, the result of this combination, i.e., the final visual complexity map, has been tested. More details on Sencogi model can be found in [25].

## 2 Method

The study evaluates the performance of Sencogi, a visual complexity model for predicting human visual fixations. Typically, fixation prediction models are compared to human fixation datasets provided by MIT [26]. However, MIT ground truth datasets are based on free viewing tasks. As Sect. 1 explains, explicit tasks are better at controlling

top-down influences, to avoid subjective factors. A judgement task has been used as described in Sect. 2.3. An analysis investigating whether scene type and cognitive factors related to the test affect human performance was performed through eye tracking and EEG. As visual fixations can be affected by emotional states [6], an emotion recognition method based on facial expressions was used to ascertain whether the test affected emotions.

## 2.1 Apparatus

The Sencogi Visual Complexity model was evaluated by three psychophysiological tools synchronized together using a biometric research platform called iMotions. The evaluation test was conducted with an MSI laptop computer, CPU Intel(R) Core(TM) i7-7820HK CPU @2.90 GHz, RAM 32.0 GB. A second screen was used at native resolution (HP 22vx IPS Monitor 546 × 305 mm, 21.5-in. LED backlight, full HD 1920 × 1080) to present images to the participants. The eye tracker, the electroencephalography (EEG) and the facial expression recognition software specifications are detailed as follows.

**Eye Tracker.** A Tobii X3, with 120 Hz gaze sample rate, 0.4° accuracy, 0.24° precision, headbox size 500 × 400 mm, affixed to the second screen.

**Electroencephalograph.** We used a 256 Hz B-Alert (Advanced Brain Monitoring, CA, USA) X10 EEG, a wireless EEG headset recording up to 9 channels of monopolar EEG based on the international 10–20 system [27] and one optional channel of electrocardiography (ECG) data. The FAA index was calculated as the difference between the alpha EEG power right electrode (in this study, F4) and the alpha EEG power left (in this study, the F3 electrode) [28].

**Facial Expression Recognition Software.** The Affectiva Affdex technology was used. The Affdex algorithm uses an index of facial expressions, called Facial Action Coding System (FACS) [23] to classify the fundamental actions of muscles involved in facial expressions, called Action Units [29]. The Action Units processed by Affdex are unconscious psychomotor processes corresponding to emotional valence rated from −100 (Negative valence) to 100 (Positive valence), and the seven universal human emotions rated from 0 (absent) to 100 (present). Measures were computed with a 20% confidence threshold. The facial expressions that fall outside the threshold were assigned with a “neutral” valence or “lack of facial expressions”.

## 2.2 Methods for Data Representation

**Human Visual Fixations Representation.** Human fixations can be represented in different formats, either preserving their discrete nature or smoothing the fixation locations to have a continuous representation, called a heatmap. Common practice is to use a Gaussian to blur the points and spread them to mimic visual acuity at one degree [30].

**Sencogi Visual Complexity Representation.** Sencogi maps have a blocky nature: the model measures the visual complexity per region of neighboring pixels. Those regions, hereafter called macroblocks, can have configurable size (8 × 8, 16 × 16 or 32 × 32,

depending on the compression). Each pixel within a macroblock has the same visual complexity value. Visual complexity values are independently computed for each macroblock, meaning that visual complexity can greatly vary among nearby macroblocks.

**Common Representation.** To fairly compare Sencogi maps to visual fixation heatmaps, a common representation was needed. First, the Gaussian heatmaps were computed according to Le Meur and colleagues [30], then the Gaussian heatmaps were transformed into a macroblock version by adopting the maximum heatmap value within the covered area. Gaussian heatmaps are commonly compared to the blurred version of saliency maps, however blurring a Sencogi map would lead to the loss of neighboring values' independence. In order to not blend independently-computed information, we opted for a subsequent blocky transformation on the visual fixation heatmaps. Heatmaps were computed with the code provided by Bylinskii and colleagues [26], using a cut off value ( $fc = 8$ ) which corresponds to a Gaussian width of 37 pixels (roughly  $1^\circ$  of visual angle). Each Sencogi macroblock dimension ( $8 \times 8$ ,  $16 \times 16$ ,  $32 \times 32$ ) falls within the Gaussian width, thus the blocky transformation does not introduce artefacts that may affect the information conveyed by heatmaps values due to exceeding  $1^\circ$  of visual angle.

### 2.3 Metrics

A Chance baseline and a Single Observer baseline were computed. The Chance baseline was calculated by assigning a random value to image pixels and using it as a saliency map [26]. We used the Chance baseline as a reference bound for each metric and macroblock combination (see Sect. 4) in order to calculate whether Sencogi predicts eye fixations of participants significantly above chance. The Single Observer baseline was computed by calculating how well a fixation map from a baseline of a single human observer predicts all the subjects of the test, averaged over all images and five variations of observers per image. The Single Observer baseline was computed by following the procedure described by Judd and colleagues [31]. The Single Observer baseline was created in order to understand how a fixation map of one observer predicts fixations of the other observers' fixations. The literature shows that the Single Observer baseline provides values "about as accurate as the best saliency models" ([31], p. 12). The performance of Sencogi Visual Complexity models in predicting the average visual fixations of our cohort, was compared with both the Chance baseline and Single Observer baseline.

The performance assessment of both the Sencogi model and the Single Observer baseline was measured by four of the most used saliency evaluation metrics chosen according to the recommendations provided by Bylinskii and colleagues in a recent paper [26]. Although using only one metric is considered enough for assessing one saliency model, the same metric may be not appropriate for comparing its performance with a model designed with different assumptions and for different applications, as in this study. For our purposes, we chose to use metrics that differ in the way they consider fixation locations (i.e., location based or distribution based) and how they compare saliency maps with fixation maps (i.e., similarity or dissimilarity). Three similarity measures, i.e., Linear Correlation Coefficient (CC), Shuffled AUC (sAUC),

and Similarity (Sim), and one dissimilarity measure, i.e., Kullback–Leibler divergence (KL) were used. CC is a correlation coefficient metric returning a measure of the linear correlation between two maps. A CC value of one reflects a total correlation between the maps. KL is a nonlinear measure often used in machine learning and neuroscience to calculate the overall dissimilarity between two probability distributions. KL range is from zero to infinity, where a value close to zero indicates a higher similarity. Area Under the ROC curve (AUC) uses fixation and non-fixation points. The shuffled-AUC (sAUC) uses gaze fixations from other images to sample non-fixation distribution. While in AUC the sampling of non-fixation points is random, in sAUC there is more control on the sampling strategy, meaning that points are taken from regions of other images that are highly fixated. sAUC values of one indicates a perfect correlation.

Sim measures the similarity between two distributions, which are considered as histograms. Sim is calculated as the sum of the minimum values at each pixel, after normalizing the input maps [26]. Sim values close to one indicate a higher similarity of distributions, while Sim values close to zero indicate no correlation.

These are the most commonly used metrics for the evaluation of saliency maps and they are reported in the MIT Saliency Benchmark [32].

## 2.4 Procedure

A selection of forty images taken from the MIT1003 database [33] was used. The database consists of a collection of 1003 random images from Flickr creative commons and LabelMe [33]. The images were selected by dimension (selected pictures were all  $1024 \times 768$  pixels), format (images were all landscape format) and scene type (ten pictures were selected according to four scene types: closeup, indoor, outdoor and subjects). The selected pictures were used for an image quality judgment test using the Single Stimulus Continuous Quality Scale [34] procedure. The test consists of showing a series of single images and asks participants to rate their subjective perceived quality on a slider numerically marked from one to one hundred and divided into five equal portions, which are labelled as “Bad”, “Poor”, “Fair”, “Good”, and “Excellent”. The image quality ratings that the participants provide in their judgement task are not used - its role is only to provide a judgement task for the participants. The test was administered to participants using the iMotions platform. Subsequent images were interleaved with a central fixation stimulus displayed for 300 ms. Each image was displayed for 5000 ms at screen native resolution. After the image was displayed, the participant was asked to rate the perceived quality on a dim-grey page lasting at least 300 ms. The experimental task of judging the quality of forty images lasted about 10 min. The test was conducted under an artificial constant dim light. Participants were seated at about 600 mm from the screen and equipped with the EEG headset. Before starting the test, the electrodes impedance was assessed, EEG benchmark data was acquired, and a 16-point eye-tracking calibration was conducted. The whole pre-setting phase required about 20 min per participant.

### 3 Results

This part reports the results of comparisons between the Sencogi model, a Chance baseline and a Single Observer baseline, compared to human eye fixation data obtained from our cohort on a selection of pictures taken from MIT1003 database. Results on eye tracking data, facial expression recognition and electroencephalography are also described.

#### 3.1 Subjects

An experimental test with twenty right-handed subjects was conducted. Participants (40% female, mean age 28.75 years old) declared normal or corrected-to-normal vision, no color blindness, and no professional experience in the field of video systems. Participants were rewarded for performing the test. One outlier was excluded by the eye tracking dataset.

#### 3.2 Experimental Eye Tracking Ground Truth

Fixation number (total mean number of fixations for the entire test = 5.79, S.D. = 174.36) and fixation duration (total mean time of fixations for the entire test = 174.37, S.D. = 31) were calculated for all participants and for each type of image (closeup, indoor, outdoor, subjects). We discarded the first fixation from eye tracking data of each image in order to avoid adding irrelevant information from the initial center fixation [33]. For each measure, comparisons by image type were conducted. The one way analysis of variance (ANOVA) found no significant difference for both fixation number ( $F(3,72) = 0.144$ ,  $p > 0.05$ ) and fixation duration ( $F(3,72) = 0.427$ ,  $p > 0.05$ ) among all image types.

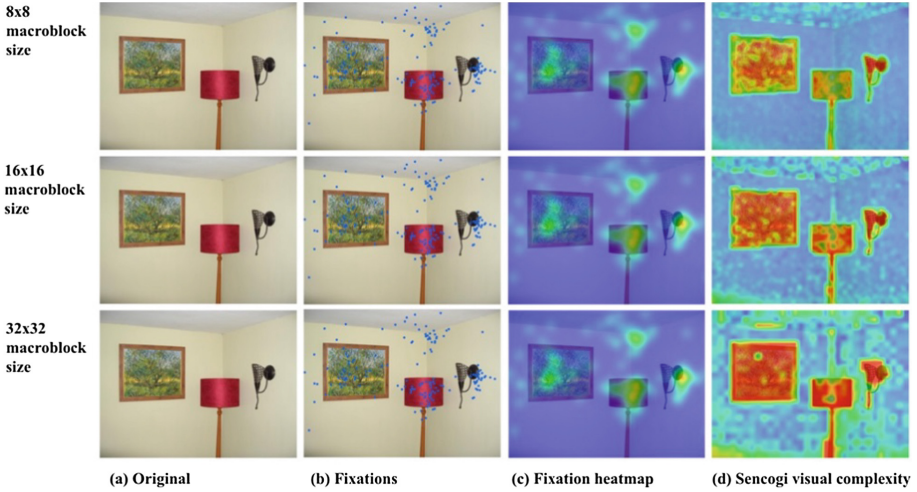
#### 3.3 Performance Assessment of the Sencogi Visual Complexity Model

An analysis was performed to see whether visual complexity could be a predictor of human fixations. An example of fixation and visual complexity maps is given by Fig. 1.

**Performance Assessment Metrics.** To test the hypothesis that visual complexity is a predictor of human gaze, four metrics were performed for comparisons with human fixations: Pearson's Correlation Coefficient (CC), Kullback–Leibler divergence (KL), Shuffled AUC (sAUC), and Similarity or histogram intersection (Sim) (see Sect. 3.1). Table 1 shows the results of metrics for visual complexity maps calculated for  $8 \times 8$ ,  $16 \times 16$  and  $32 \times 32$  macroblocks size. A Chance baseline and a Single Observer baseline were calculated as explained in Sect. 3.1 (Table 1).

**Comparisons Between the Sencogi Model Performance Scores and Chance:** Chance values and visual complexity values were calculated for each metric. Results of paired t-test show a highly significant difference for all the metrics at each macroblock size (Table 2).





**Fig. 1.** Figure shows an example of (a) original picture, (b) fixations of all the participants measured with the eye tracking methodology, (c) fixations heatmap mapped into the same range of values of Sencogi maps, (d) Sencogi visual complexity heatmaps according to the three different macroblock sizes ( $8 \times 8$ ,  $16 \times 16$ ,  $32 \times 32$ ).

**Table 1.** Table shows the results obtained by the Sencogi model, the Single Observer baseline, and the Chance baseline at each macroblock size.

|                                | CC $\uparrow$ | KL $\downarrow$ | sAUC $\uparrow$ | Sim $\uparrow$ |
|--------------------------------|---------------|-----------------|-----------------|----------------|
| Single observer $8 \times 8$   | 0.396         | 8.030           | 0.582           | 0.313          |
| Sencogi $8 \times 8$           | 0.121         | 1.127           | 0.542           | 0.409          |
| Chance $8 \times 8$            | 0.001         | 1.496           | 0.501           | 0.375          |
| Single observer $16 \times 16$ | 0.413         | 7.800           | 0.587           | 0.328          |
| Sencogi $16 \times 16$         | 0.128         | 1.036           | 0.543           | 0.433          |
| Chance $16 \times 16$          | 0.000         | 1.439           | 0.501           | 0.390          |
| Single observer $32 \times 32$ | 0.447         | 7.328           | 0.599           | 0.360          |
| Sencogi $32 \times 32$         | 0.132         | 0.913           | 0.539           | 0.469          |
| Chance $32 \times 32$          | 0.000         | 1.320           | 0.496           | 0.417          |

**Table 2.** Comparisons between the Sencogi model performance scores and the chance baseline for each macroblock size.

|      | $8 \times 8$ |            | $16 \times 16$ |            | $32 \times 32$ |            |
|------|--------------|------------|----------------|------------|----------------|------------|
|      | $t(201)$     | $p$ -value | $t(201)$       | $p$ -value | $t(201)$       | $p$ -value |
| CC   | -11.657      | 0.000      | -11.741        | 0.000      | -12.187        | 0.000      |
| KL   | 30.181       | 0.000      | 33.897         | 0.000      | 31.925         | 0.000      |
| sAUC | -9.011       | 0.000      | -9.217         | 0.000      | -8.373         | 0.000      |
| Sim  | -12.091      | 0.000      | -17.882        | 0.000      | -24.512        | 0.000      |



**Comparisons Among the Sencogi Model Performance and the Single Observer Baseline Performance:** Comparisons at three macroblock sizes show significantly better CC and sAUC scores for the Single Observer baseline than Sencogi model (Table 3).

**Table 3.** Comparisons between the CC and the sAUC scores of Sencogi and the Single Observer baseline for each macroblock size.

|      | 8 × 8          |                 | 16 × 16        |                 | 32 × 32        |                 |
|------|----------------|-----------------|----------------|-----------------|----------------|-----------------|
|      | <i>t</i> (201) | <i>p</i> -value | <i>t</i> (201) | <i>p</i> -value | <i>t</i> (201) | <i>p</i> -value |
| CC   | 15.126         | 0.000           | 15.671         | 0.000           | 17.279         | 0.000           |
| sAUC | 6.733          | 0.000           | 6.733          | 0.000           | 9.109          | 0.000           |

Significantly better KL scores and better Sim scores for Sencogi model than the single observer model were found (Table 4).

**Table 4.** Comparisons between the KL and the Sim scores of Sencogi and the Single Observer baseline.

|     | 8 × 8          |                 | 16 × 16        |                 | 32 × 32        |                 |
|-----|----------------|-----------------|----------------|-----------------|----------------|-----------------|
|     | <i>t</i> (201) | <i>p</i> -value | <i>t</i> (201) | <i>p</i> -value | <i>t</i> (201) | <i>p</i> -value |
| KL  | 29.468         | 0.000           | 29.358         | 0.000           | 28.702         | 0.000           |
| Sim | -9.213         | 0.000           | -9.853         | 0.000           | -10.048        | 0.000           |

**The Effect of Scene Type on the Sencogi Model Performance:** Visual complexity maps were calculated for 8 × 8 macroblock resolution. No significant difference among metric values for image type was found (KL,  $F(3, 36) = 1.213, p > 0.05$ ; sAUC,  $F(3, 36) = 0.860, p > 0.05$ ; Sim,  $(F(3, 36) = 1.367, p > 0.05)$ , except for CC where outdoor images have significantly better CC values than closeup images (CC,  $F(3, 36) = 2.962, p = 0.042$ ). Visual complexity maps were calculated for 16 × 16 macroblock resolution. No significant difference among metric values for image type was found (CC,  $F(3, 36) = 2.704, p > 0.05$ ; KL,  $F(3, 36) = 0.874, p > 0.05$ ; sAUC,  $F(3, 36) = 0.615, p > 0.05$ ; Sim,  $F(3, 36) = 0.746, p > 0.05$ ). Visual complexity maps were calculated for 32 × 32 macroblock resolution. No significant difference among metric values for image type was found (CC,  $F(3, 36) = 1.733, p > 0.05$ ; KL,  $F(3, 36) = 0.520, p > 0.05$ ; sAUC,  $F(3, 36) = 1.068, p > 0.05$ ; Sim,  $F(3, 36) = 0.417, p > 0.05$ ).

### 3.4 Facial Expression Recognition

**The Effect of the Test on Participants’ Affective Valence and Basic Emotions Responses.** The mean time percent of both valence and basic emotions related to test pictures, out of the total time recorded for all the whole test, was calculated (Table 2).

Emotions were calculated by iMotions<sup>1</sup> as the mean of 100\* (count frames in which emotion appears/count frames in stimulus) (Table 5).

**Table 5.** Comparisons between the KL and the Sim scores of Sencogi and the Single Observer baseline.

| <i>Affective valence</i> | MTP    | S.D.  |
|--------------------------|--------|-------|
| Neutral valence          | 58.039 | 4.060 |
| Positive valence         | 0.089  | 0.330 |
| Negative valence         | 1.820  | 3.270 |
| <i>Basic emotions</i>    | MTP    | S.D.  |
| Contempt                 | 0.562  | 2.562 |
| Fear                     | 3.280  | 9.298 |
| Surprise                 | 2.239  | 7.799 |
| Joy                      | 0.092  | 0.541 |
| Disgust                  | 0.012  | 0.089 |
| Sadness                  | 0.002  | 0.014 |
| Anger                    | 0.013  | 0.103 |

The one way ANOVA on affective valence means shows significant differences among neutral, positive and negative affective values ( $F(2, 119) = 19840.4, p = 0.000$ ). The post-hoc test (Bonferroni corrected) shows a meantime percent of facial expressions related to neutral affective states that is significantly longer at  $p < 0.01$  than both negative and positive valence values. A multivariate analysis of variance (MANOVA) was performed on affective valence mean time percent, showing no significant difference among image types (subjects, outdoor, indoor, closeup) for neutral, positive and negative affective values (Wilks' Lambda  $F(9, 180) = 0.387, p > 0.05$ ).

Significant time differences among basic emotions were found with respect to each other ( $F(6, 279) = 70.735, p = 0.000$ ). The post-hoc test (Bonferroni corrected) shows a meantime percent of facial expressions related to fear (mean time = 3.281%) and surprise (mean time = 2.240%) that is significantly higher at  $p < 0.01$  than the other measured basic emotions.

A paired t-test comparison among the subsections of the test composed by center-fixation slide, image slide, and image quality rating slide shows significantly higher values of emotions of fear and surprise during the image slides (fear  $t(39) = 4.380, p = 0.000$ ; surprise  $t(39) = 3.177, p = 0.003$ ) and quality rating slides (fear  $t(39) = -4.954, p = 0.000$ ; surprise  $t(39) = -4.737, p = 0.001$ ) compared to center-fixation slides. No significant differences were found for emotions of fear and surprise between the image slides and quality rating slides (fear  $t(39) = -0.187, p > 0.05$ ; surprise  $t(39) = -0.137, p > 0.05$ ).

<sup>1</sup> [www.imotions.com](http://www.imotions.com).

**The Effect of Scene Type on Participants' Basic Emotions.** A MANOVA was performed on emotions mean time percent, showing no significant difference among image types (subjects, outdoor, indoor, closeup) for basic emotions (Wilks' Lambda  $F(21, 201) = 0.565, p > 0.05$ ).

**Eye Tracking Data Compared to Facial Expression Recognition.** No correlation between fixation number, affective valence, and basic emotions was found. A negative correlation was found between fixation duration and fear (Pearson's  $r = -0.383, p = 0.015$ ) and disgust (Pearson's  $r = -0.324, p = 0.041$ ). No other correlation between fixation duration, affective valence, and basic emotions was found.

### 3.5 EEG

**The Effect of the Test on Participants' Approach/Avoidance States.** After a previous automatic decontamination process for artefact removal, the Frontal Alpha Asymmetry (FAA) mean values were calculated (Table 6).

**Table 6.** FAA mean value for each scene type.

| Scene type | FAA    |
|------------|--------|
| Closeup    | 79.915 |
| Indoor     | 82.761 |
| Outdoor    | 81.488 |
| Subjects   | 81.693 |

The one way ANOVA on mean FAA values shows no significant differences among image types ( $F(3, 72) = 0.001, p > 0.05$ ).

**Eye Tracking Data Compared to EEG Frontal Alpha Asymmetry.** No significant correlation between fixations and FAA was found (Pearson's  $r = 0.066, p > 0.05$ ).

## 4 Discussion

The aim of this work was to investigate whether the Sencogi model of visual complexity can accurately predict human foveal vision, and to control for factors that might affect saliency. Three experimental investigations were performed to verify (i) whether Sencogi visual complexity predicts visual fixations, (ii) whether saliency prediction performs better with certain types of visual scenes than others, and (iii) whether the experimental test introduces top-down factors affecting foveal fixations.

(i) **Does Sencogi visual complexity predict visual fixations?** Sencogi visual complexity maps were compared to human visual fixation maps using four of the most suitable, commonly used assessment metrics (CC, KL, sAUC, and Sim).

Statistical comparisons of the results with a Chance baseline were performed. The Chance baseline assigns saliency values to random pixels (chance). The results show that Sencogi has significantly better scores than chance (i.e., they are closer to human fixations).

Sencogi performs better than the Single Observer baseline at predicting overall visual saliency in two assessment metrics, i.e., Kullback–Leibler divergence and Similarity, likely because they both penalize sparser sets of predictions such as the Single Observer baseline, which might miss fixations that emulate average fixation behavior [26]. Comparisons were made between the Sencogi model performance and the Single Observer baseline performance in predicting visual fixations. The Single Observer baseline obtained significantly higher scores for the Pearson’s Correlation Coefficient (CC) metric and for the shuffled AUC metric. These results can be explained by the nature of the metrics used for assessing the two models. As the CC metric does not assume that saliency models are probabilistic [26], it will produce a worse score for models such as Sencogi that don’t include any possible systematic bias of observers. The other metric in which Sencogi received significantly lower values than the Single Observer baseline is sAUC. sAUC is not invariant to center bias. A model with high central predictions is ranked lower than a model with more peripheral predictions [26].

Sencogi received significantly lower sAUC scores than the Single Observer baseline, meaning that the Single Observer baseline could have sets of predictions that are too sparse to emulate average viewing behavior in the central part. In this perspective, the Single Observer baseline is not penalized by the sAUC metric because it does not account for the central predictions of all the observers’ fixations as Sencogi does. Therefore, for most measures that are not biased, Sencogi correctly predicts visual fixations.

*(ii) Is Sencogi visual saliency prediction sensitive to scene type?* The experimental images shown during the test were divided into four groups according to the type of scene depicted, i.e., images showing closeup, indoor, outdoor, and subject scenes in order to investigate whether image type influences the Sencogi performance scores. We found no significant difference for  $16 \times 16$  and  $32 \times 32$  visual complexity maps. Only maps with  $8 \times 8$  pixels macroblocks had significantly higher linear correlation scores (CC) for images showing outdoor scenes compared to the other scene types, in particular closeup scenes. This result might be related to the nature of outdoor scenes, in which subjects are usually smaller than other scene types. However, this result is related to a barely detectable statistically significant difference that happens only for one of the four metrics (CC). The other assessment metrics (KL, sAUC, and Sim) do not seem to be sensitive to any difference among scene types. A new in-depth investigation on the effects of scene type on Sencogi performance is needed.

*(iii) Does the experimental test introduce top-down factors affecting eye foveal fixations?* Eye fixation duration reflects how much the fixated information is processed by cognitive processes [35]. The more a visual stimulus requires the allocation of the analysis process, the more fixation time is prolonged [36] because of the increase in mental load. In this work, the average eye fixation lasts about 180 ms per image, which is about 20% of the presentation time, meaning that the nature and the duration of the task did not cause any significant increase in cognitive workload to participants. No

differences among visual scene type were found for fixation numbers and fixation duration, meaning that in this study, scene type did not affect eye fixations in a significant way.

During the presentation of experimental images, the overall neutral emotional state in participants was significantly higher than positive and negative emotional states. Both emotional valence and basic emotions were not affected by scene type, confirming that the type of content in the images used for this study did not affect the emotional quality of the interaction. However, if we analyze only the longitudinal differences among basic emotions, participants' facial expressions related to fear and surprise were significantly higher than the others (joy, sadness, anger, contempt and disgust). This result suggests that the emotions of fear and surprise might be due to the nature of the task, which is a time-based trial (participants have five seconds to examine the quality of each image) that does not provide any feedback to let the participant know whether they are correctly performing the assigned task. This assumption is supported by significantly higher scores of fear and surprise during both images and scoring slides presentation, compared to neutral areas where participants are not asked to perform a judgment task, such as the center fixations period. Emotional states do not correlate with the number of fixations, but the duration of eye fixations negatively correlates with emotions of fear and disgust, confirming that the more observers feel negative emotions, the less likely they are to keep gazing at a certain fixation point [37].

Findings on EEG measures showed no differences in FAA mean values among image types, nor the number and the duration of fixations, meaning that participants' approach/withdraw psychophysiological processes did not significantly change during the test. These findings highlight that no intervening top-down factor related to test or to the material affected eye foveal fixations in a significant way.

## 5 Conclusion

In this paper, we validated the performance of the Sencogi visual complexity model for predicting visual fixations of static frames. Visual complexity maps were compared with human fixations derived from a judgment task. Findings show that Sencogi significantly surpasses Chance and performs better than the model of a single observer for metrics that penalize models that are too sparse to emulate average viewing behavior (in particular, the KL metric and the Sim metric). We found no noteworthy differences in predicting visual saliency among the four scene types of the images (closeup, indoor, outdoor, and subjects).

Investigations of psychophysiological measures during the test excluded the possibility that, in this test, cognitive and affective psychological factors such as approach/withdraw processes, emotional valence or basic emotions relate to fixation number and/or duration. This lack of emotional correlation suggests that psychological factors did not affect the participants' performance. Observers' mental load and affective valence (overall neutral) were not influenced by task-related factors such as scene type. The only basic emotions seen during the test (fear and surprise) were not significantly present over the whole test. Fear and surprise are most likely related to the demand of assigning a quality score rather than to the image content. Future work will

focus on comparisons between the performance in predicting saliency of Sencogi and traditional state of the art saliency methods. This study has shown that Sencogi can produce visual complexity maps that predict visual fixations for static images, but Sencogi can also produce spatio-temporal visual complexity maps, so future studies on video databases will be conducted in order to validate the temporal components of the Sencogi model.

## References

1. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. *Cogn. Psychol.* **12**(1), 97–136 (1980). [https://doi.org/10.1016/0010-0285\(80\)90005-5](https://doi.org/10.1016/0010-0285(80)90005-5)
2. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Vaina, L.M. (ed.) *Matters of Intelligence*. Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science), vol. 188, pp. 115–141. Springer, Dordrecht (1987). [https://doi.org/10.1007/978-94-009-3833-5\\_5](https://doi.org/10.1007/978-94-009-3833-5_5)
3. Kummerer, M., Wallis, T.S., Gatys, L.A., Bethge, M.: Understanding low-and high-level contributions to fixation prediction. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4789–4798 (2017). <https://doi.org/10.1109/iccv.2017.513>
4. Itti, L., Koch, C.: Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**(3), 194 (2001). <https://doi.org/10.1038/35058500>
5. Zhang, Y.Y., Zhang, S., Zhang, P., Zhang, X.: Saliency detection via background and foreground null space learning. *Sig. Process. Image Commun.* **70**, 271–281 (2019). <https://doi.org/10.1016/j.image.2018.10.005>
6. Shen, J., Itti, L.: Top-down influences on visual attention during listening are modulated by observer sex. *Vision. Res.* **65**, 62–76 (2012). <https://doi.org/10.1016/j.visres.2012.06.001>
7. Borji, A., Sihite, D.N., Itti, L.: What stands out in a scene? A study of human explicit saliency judgment. *Vision. Res.* **91**, 62–77 (2013). <https://doi.org/10.1016/j.visres.2013.07.016>
8. Koehler, K., Guo, F., Zhang, S., Eckstein, M.P.: What do saliency models predict? *J. Vis.* **14**(3), 14 (2014). <https://doi.org/10.1167/14.3.14>
9. Zhang, W., Borji, A., Wang, Z., Le Callet, P., Liu, H.: The application of visual saliency models in objective image quality assessment: a statistical evaluation. *IEEE Trans. Neural Netw. Learn. Syst.* **27**(6), 1266–1278 (2016). <https://doi.org/10.1109/TNNLS.2015.2461603>
10. Muddamsetty, S.M., Sidibe, D., Tremeau, A., Meriaudeau, F.: Spatio-temporal saliency detection in dynamic scenes using local binary patterns. In: *2014 22nd International Conference on Pattern Recognition* (2014). <https://doi.org/10.1109/icpr.2014.408>
11. He, X. (ed.): *IScIDE 2015*. LNCS, vol. 9243. Springer, Cham (2015). <https://doi.org/10.1007/978-3-319-23862-3>
12. Yubing, T., Cheikh, F.A., Guraya, F.F.E., Konik, H., Trémeau, A.: A spatiotemporal saliency model for video surveillance. *Cogn. Comput.* **3**, 241–263 (2011). <https://doi.org/10.1007/s12559-010-9094-8>
13. Mu, N., Xu, X., Zhang, X.: A spatial-frequency-temporal domain based saliency model for low contrast video sequences. *J. Vis. Commun. Image Represent.* **58**, 79–88 (2019). <https://doi.org/10.1016/j.jvcir.2018.11.012>
14. Rapantzikos, K., Avrithis, Y., Kollias, S.: Dense saliency-based spatiotemporal feature points for action recognition (2009). <https://doi.org/10.1109/cvpr.2009.5206525>

15. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(1), 171–177 (2010). [https://doi.org/10.1007/978-3-642-37431-9\\_41](https://doi.org/10.1007/978-3-642-37431-9_41)
16. Oikonomopoulos, A., Patras, I., Pantic, M.: Spatiotemporal salient points for visual recognition of human actions. *IEEE Trans. Syst. Man Cybern. Part B (Cybernetics)* **36**(3), 710–719 (2005). <https://doi.org/10.1109/tsmcb.2005.861864>
17. Dhiman, C., Vishwakarma, D.K.: A review of state-of-the-art techniques for abnormal human activity recognition. *Eng. Appl. Artif. Intell.* **77**, 21–45 (2019). <https://doi.org/10.1016/j.engappai.2018.08.014>
18. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 1597–1604. IEEE, June 2009. <https://doi.org/10.1109/cvpr.2009.5206596>
19. Mele, M.L., Millar, D., Rijnders, C.E.: The web-based subjective quality assessment of an adaptive image compression plug-in. In: *1st International Conference on Human Computer Interaction Theory and Applications, HUCAPP, Porto, Portugal (2017)*. <https://doi.org/10.5220/0006226401330137>
20. Mele, M.L., Millar, D., Rijnders, C.E.: Using spatio-temporal saliency to predict subjective video quality: a new high-speed objective assessment metric. In: Kurosu, M. (ed.) *HCI 2017. LNCS*, vol. 10271, pp. 353–368. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-58071-5\\_27](https://doi.org/10.1007/978-3-319-58071-5_27)
21. Mele, M.L., Millar, D., Rijnders, C.E.: Sencogi spatio-temporal saliency: a new metric for predicting subjective video quality on mobile devices. In: Kurosu, M. (ed.) *HCI 2018. LNCS*, vol. 10902, pp. 552–564. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91244-8\\_43](https://doi.org/10.1007/978-3-319-91244-8_43)
22. Schneirla, T.C.: An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal (1959). <https://doi.org/10.1177/000306517001800210>
23. Ekman, P., Friesen, W.V.: *Manual for the Facial Action Coding System*. Consulting Psychologists Press (1978). <https://doi.org/10.4135/9781483381411>
24. Redi, J., Liu, H., Zunino, R., Heynderickx, I.: Interactions of visual attention and quality perception. In: *Human Vision and Electronic Imaging XVI (2011)*. <https://doi.org/10.1117/12.876712>
25. Rijnders, C.E.: U.S. Patent Application No. 15/899,331 (2018)
26. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.* (2018). <https://doi.org/10.1109/tpami.2018.2815601>
27. Jasper, H.H.: The ten-twenty electrode system of the International Federation. *Electroencephalogr. Clin. Neurophysiol.* **10**, 370–375 (1958). <https://doi.org/10.1080/00029238.1961.11080571>
28. John, E.R.: *Neurometrics: clinical applications of quantitative electrophysiology*, vol. 2. Wiley (1977). <https://doi.org/10.1177/002076407902500222>
29. Ekman, P.: *Facial action coding system (FACS). A human face (2002)*
30. Le Meur, O., Baccino, T.: Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behav. Res. Methods* **45**(1), 251–266 (2013). <https://doi.org/10.3758/s13428-012-0226-9>
31. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations (2012)
32. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F.: Where should saliency models look next? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9909, pp. 809–824. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46454-1\\_49](https://doi.org/10.1007/978-3-319-46454-1_49)



33. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113. IEEE, September 2009. <https://doi.org/10.1109/iccv.2009.5459462>
34. BT.500: Methodology for the subjective assessment of the quality of television pictures (n. d.). <http://www.itu.int/rec/R-REC-BT.500-7-199510-S/en>. Accessed 9 Oct 2017
35. Yarbus, A.L.: Eye Movements and Vision, New York (1967). <https://doi.org/10.1007/978-1-4899-5379-7>
36. Jacob, R.J.: The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Trans. Inf. Syst. (TOIS)* **9**(2), 152–169 (1991). <https://doi.org/10.1145/123078.128728>
37. Mele, M.L., Federici, S., Dennis, J.L.: Believing is seeing: fixation duration predicts implicit negative attitudes. *PLoS ONE* **9**(8), e105106 (2014). <https://doi.org/10.1371/journal.pone.0105106>