# A statistical study of the correlation between interest points and gaze points

Michael Nauge, Mohamed-Chaker Larabi, Christine Fernandez-Maloigne

# A Statistical Study of the Correlation Between Interest Points and Gaze Points

Michael Nauge, Mohamed-Chaker Larabi and Christine Fernandez-Maloigne

XLim Lab, SIC dept., University of Poitiers, France

## ABSTRACT

This study intends to measure the degree of correlation/similarity between the subjective gaze points (obtained by eye tracking experiments) and the objective interest points of several well-known detectors such as Harris and SURF. For each of the latter, we look for the best setting in term of maximization of likeness with the gaze points. For this task, the Earth Mover's Distance (EMD)[1] is used to compare two data-sets with different cardinalities. We also used ANOVA to measure the influence of each parameter involved in the detectors' settings as well as the possible introduced bias. The conclusions of this study are related to the suitability of each detector to estimate the subjective gaze points.

**Keywords:** Subjective Gaze Points, Eyetracking Experiments, Objective Interest Points, Harris, SURF, EMD

## 1. INTRODUCTION

Understanding the human perception and cognition, and modeling the Human Visual System (HVS) is fundamental for the improvement of electronic media systems. Plenty of experiments have been conducted by using eye-trackers for a variety of applications in psychology, human-computer interaction, marketing, cognitive science... Eye tracking experiments allow to determine the salient points/regions on a given image. Understanding and modeling the salient stimulus localization is a hard task due to the complexity and the variety of involved brain and eyes processes. For example the location of the region of interest can depend on the nature of the experiment (free watching or task-oriented) such as demonstrated by Yarbus.[2] However, some studies[3] demonstrated the invariance of the gaze points when text and faces are in the image. The recognition of text and faces are high level processes and not present in all situations. So, in this case other stimuli are attractive. Itti &Koch[4] demonstrated the influence of low level criteria such as intensity, color and direction of the stimulus. Another old and well known criterion is the central biais re-explored by B. W. Tatler[5] who tried to explain why when observers view complex scenes presented on computer monitors, there is a strong trend to look more frequently around the center of the scene than around the corners. The study of the human gaze and eye movement has attracted an important research effort in the last decade.

From another point of view, in the image/video processing field, several algorithms have been constructed following the concept of interest points, like Harris corner detector[6] , Scale-Invariant Feature Transform (SIFT[7]) detector or Speeded Up Robust Features (SURF[8]) detector. Interest points are very important for the characterization of an object or a texture and allow the discrimination of shapes and objects. These algorithms are classically used in motion detection and object recognition. These detectors can be categorized by the kind of the detected features. Harris detector is design to localize the corners, whereas SIFT and SURF are design to detect scale- and rotation-invariant blob-like structures. Lowe[7] explained the interest of the detection of features invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection. He explained that these features share similar properties with neurons in inferior temporal cortex that are used for object recognition in primate vision. H. Bay[8] noted a lot of similarity between his SURF detector and the SIFT detector in term of the kind of blob-like structure detected. But the process to detect theses features is completely different and designed to be faster than SIFT. For this study we focus our study on Harris for the detection of corner, and on SURF to detect the blob-like structure with its low time

Further author information: (Send correspondence to M.N)
M.N.: E-mail: michael.nauge@univ-poitiers.fr
M.C.L.: E-mail: chaker.larabi@univ-poitiers.fr

computation and its robustness. Few papers[9] studied the performance and the robustness of these detectors in image matching task over various distortions.

In this study, we propose to study the performance of these kind of detectors with a totally different point of view, by studying similarity/correlation between the subjective gaze points and the objective interest points. The aim is to determine whether the interest points can be used to predict salient informations on an image like the HVS does. This can help for several applications like quality assessment,[10] simplified saliency maps construction,[11] ... Even though the interest points have not been originally designed to be close to the gaze points, they may have a particular setting that maximizes the similarity between them. For this study a battery of statistical tools is used to test a large range of settings/configurations for two interest point detectors (Harris and SURF).

For our experiments we used the Visual Attention for Image Quality (VAIQ) Database.[12] This database was created after an eye-tracking experiment at the University of Western Sydney in Australia. It contains 53 reference images extracted from the widely used image quality databases (LIVE, MICT, IVC). A comparison with other eye-tracking results on the same images coming from Delft University of Technology (TUD) in Netherlands demonstrated the reliability of the VAIQ database.

The number of interest points extracted by each detector depends of the used parameters and is often different from the number of subjective gaze points. So, for the measurement of similarity between objective and subjective interest points, it is necessary to use a distance/metric able to compare two datasets with different cardinalities. We selected the Earth Mover's Distance (EMD) based on a solution to a special case of the old transportation problem[13] and implemented by Rubner et al.[1] in the context of image retrieval. The EMD Distance is based on the concept of measurement of work needed to move many masses of earth into many holes. Precisely, given two distributions, one can be seen as a mass of earth properly spread in space, the other as a collection of holes in that same space. Then, the EMD measures the least amount of work needed to fill the holes with earth. Here, a unit of work corresponds to transporting a unit of earth by a unit of ground distance. In our case, we can consider than the subjective gaze points are the holes and the interest points are the earth. With this EMD distance we can estimate which detector and setting minimize the cost of the transformation between the subjective and objective values.

We used several statistical tools such as Bartlett, ANOVA, ... to understand the effect and the influence of each parameters for each detector. These studies illustrate that particular parameters can minimize the cost of transformation and predict interest points in accordance with the subjective gaze points. We also proposed a solution to give a scale to facilitate the interpretation of the EMD values by analyzing the mean human behavior. By comparing the best settings for each detector we can also indicate which detector is the most reliable to estimate the subjective gaze points. This study is also a good way to prove that interest points detector share some properties with the HVS.

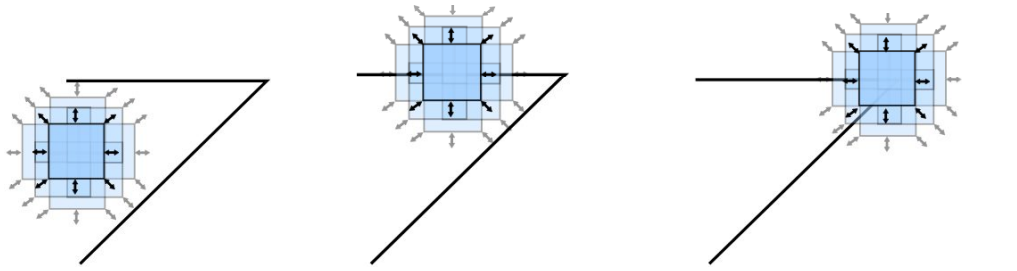
The remainder of this paper is organized as follows : Section 2 gives a description of the interest points detectors uses in this study. Section 3 describes the subjective gaze point database and the retained solution for clustering the multiple gaze points. The experimental study by using different statistical tools is given is section 4. This paper ends with some conclusions and gives ideas of future works.

## 2. OBJECTIVE INTEREST POINTS DESCRIPTION

For this study we used two different types of interest points detectors: the Harris detector for the detections of corners and the SURF detector for the detection of blob-like structures. The following sections give the description of these two retained detectors.

### 2.1 Harris

The Harris corner detector was designed to find points in an image such that, there is only a small number of isolated points detected and the points are reasonably invariant to : rotation, different sampling and quantization, to small changes of scale and small affine transformations. Since this detector satisfies these requirements, it is a standard for matching and tracking task in computer vision.



**flat** — No change in all directions

**edge** — No change in edge direction

**corner** — Significant change in all directions

Figure 1. local changes of the signal with patches shifted in different directions

The Harris corner detector[6] is based on the local auto-correlation function of a signal. The latter measures the local changes of the signal with patches shifted by a small amount in different directions.

Given a shift (u,v) and a point (x,y), the auto-correlation function is defined as:

$$E(u,v) = \sum_{(x,y)\in W} w(x,y)[I(x+u,y+v) - I(x,y)]^2 \tag{1}$$

where $I(.,.)$ denotes the intensity image function. The $w$ windows can be a smooth circular window, like a Gaussian in order to reduce the noise introduced by using a binary and rectangular window. For a small shifts (u,v) we have a bilinear approximation:

$$E(u,v) \approx (u,v)M \begin{bmatrix} u \\ v \end{bmatrix} \tag{2}$$

where $M$ is a $2 \times 2$ matrix computed from image derivatives:

$$M = \sum_{(x,y)\in W} w(x,y) \begin{bmatrix} I_x(x,y)^2 & I_x(x,y)I_y(x,y) \\ I_x(x,y)(I_y(x,y) & I_y(x,y)^2 \end{bmatrix} \tag{3}$$

where M captures the intensity structure of the local neighborhood. Let $\lambda_1$ and $\lambda_2$ be the eigenvalues of matrix M. The eigenvalues form a rotationally invariant description. Three cases of eigenvalues can be considered. In the first case $\lambda_1$ and $\lambda_2$ are small, the windowed image region is of approximately constant intensity, this is a flat/uniform region. In the second case one eigenvalue $\lambda_1$ or $\lambda_2$ is high and the the other is low. This indicates an edge. In the third case $\lambda_1$ and $\lambda_2$ are high and thus indicates a corner. Figure 2 shows the eigenvalues configurations.

A measure of corner response can be applied on each detected corner,

$$R = det(M) - k(trace(M))^2 \tag{4}$$

where

$$det(M) = \lambda_1\lambda_2$$
$$trace(M) = \lambda_1 + \lambda_2$$

k is a constant whose value was determined empirically to give results in the range [0.04, 0.06].

Finally an interest point is a corner with a particular response in respect of these criteria:

$$R > threshold \wedge \forall x,y \in 8 - neighbourhoodf(x,y) \geq f(x',y') \tag{5}$$

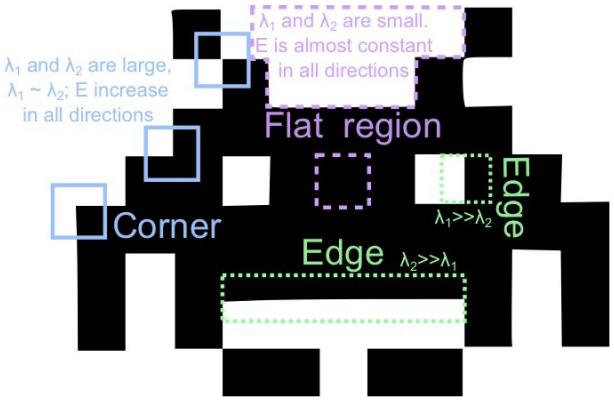



Figure 2. Four configurations of $\lambda_1$ and $\lambda_2$

Table 1. Harris detector function parameters

| Param. | Description |
|---|---|
| qualityLevel (ql) | Characterizes the minimal accepted quality of image corners; the value of the parameter is multiplied by the corner quality measure (r in equation 4). The corners, which quality measure is less than the product, will be rejected. For example, if the best corner has the quality measure = 1500, and the qualityLevel = 0.01 , then all the corners which quality measure is less than 15 will be rejected |
| minDistance (md) | The minimum possible Euclidean distance between the returned corners. |
| blockSize (bs) | Size of the averaging block for computing derivative covariation matrix over each pixel neighborhood. |
| k | weighting parameters used in the formula 4 for the calculation of corner response. |

To conclude, an interest point can be defined as the intersection of two edges. It can also be defined as a point for which there are two dominant and different edge directions in a local neighborhood of the point. The number and the position of detected corners can be tuned by different factors, such as neighborhood size, threshold on the corner response... However, Harris corners are not really scale-invariant. Mikolajczyk and Schmid refined the Lindeberg method[9] in order to create a robust and scale-invariant feature detector with high repeatability named Harris-Laplace.[14]

### 2.1.1 Harris parameters

For this experiment we used the Harris corner detector implementation available by the GoodFeaturesToTrack-Detector function in the openCV Library.[15] Details about parameters are given in table 1.

## 2.2 SURF

The Speeded Up Robust Features (SURF) aims, like the Harris detector, to detect and describe local features in image for tasks like object recognition and tracking. This method has been presented for the first time in 2006 and revised in 2008[8] by Herbert Bay et al. It focuses on scale and image rotation invariant detectors and descriptors with a good compromise between feature complexity and robustness with a low computation time. This method can be decomposed in three parts: Key Point Detection, Descriptors Extraction and Matching.

The key point detection is based on calculating approximate Hessian response for image points in order to detect blob-like structure. For the scale-space analysis a pyramid of filters (not image) is used to approximate Laplace of Gaussian (LoG), supposedly run faster than SIFT, which uses Difference of Gaussians (DoG) for approximation (Figure 3). The box-filter is used with the integral image in order to have a constant execution time although each filter is more and more large. In order to localise interest points in the image and over

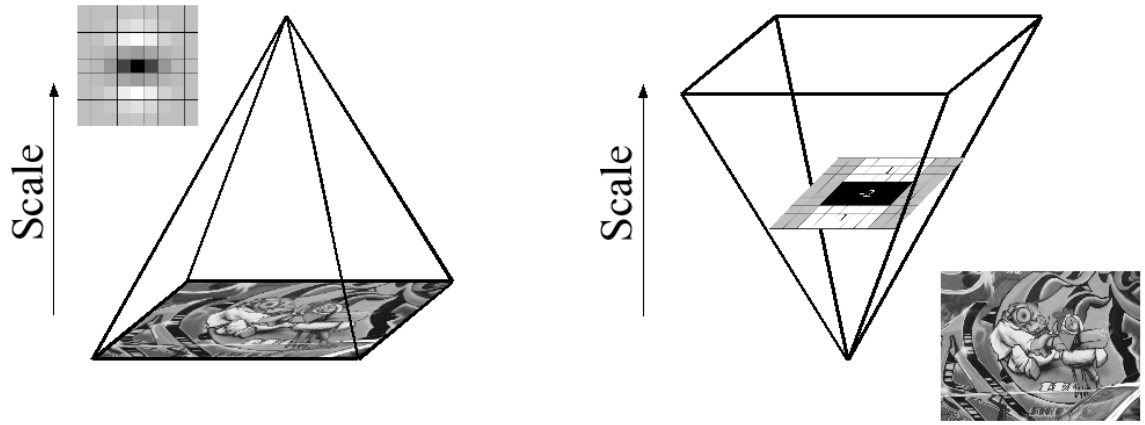



Figure 3. DoG (Left) - LoG (the use of integral images allows the up-scaling of the filter at constant cost (right))

scales, a non-maximum suppression (in a 3×3×3 neighbourhood) is applied. Feature points are maxima of the determinants in the adjacent scale and points like SIFT method.

The descriptor extraction describes the distribution of the intensity content within the interest point neighbourhood (similar to the SIFT gradient informations) and its variants. It is built on the distribution of first order Haar wavelet responses in x and y directions rather than the gradient which exploit integral images for speed, and use only 64 dimensions.Furthermore a new indexing step based on the sign of the Laplacian is introduced, which increases not only the robustness of the descriptor, but also the matching speed.

#### 2.2.1 SURF parameters

The used implementation of SURF is ExtractSURF function in the openCV Library. Details about parameters and tested values are given in table 2.

Table 2. SURF detector function parameters

| param. | Description |
|---|---|
| hessianThreshold | Only features with Hessian larger than that are extracted. good default value is 300-500 (can depend on the average local contrast and sharpness of the image). user can further filter out some features based on their hessian values and other characteristics. |
| nOctaves | The number of octaves to be used for extraction. With each next octave the feature size is doubled (3 by default). |
| nOctaveLayers | The number of layers within each octave (4 by default). |

## 3. SUBJECTIVE GAZE POINTS

### 3.1 eye-tracking database

For our experiments, we used the Visual Attention for Image Quality (VAIQ) Database.[12] This database was created after an eye-tracking experiment at the University of Western Sydney in Australia. It contains 53 reference images extracted from the widely used image quality database (LIVE, MICT, IVC). These images were displayed on a 19" Samsung SyncMaster monitor with a 1280 × 1024 screen resolution. An eye tracker was installed under the screen and the participants were seated at a distance of approximately 60 cm from the screen. This experiment used an EyeTech TM3 eye tracker to record the gaze of the human observers. The accuracy with which the gaze is recorded is approximately 1 deg of visual angle. The eye tracker records gaze points (GP)

at about 40-45 GP/sec. Each image was shown for 12 seconds with a mid-grey screen shown between images for 3 seconds. The number of gaze point by image and by user is between 480-540.

The validity of this database is tested ([16]) with a second experiment at Delft University of Technology (TUD) in Netherlands, with twenty new observers and twenty nine similar images. This second experiment proved a good correlation between these two different experiments in two different laboratories for same images and different observers. So using the VAIQ Database appears reliable.

In addition to the 53 reference images, the gaze points of 15 observers and the generated heatmaps for each images are available. The heatmaps are obtained by applying Gaussian on filtered gaze points. For our experimentation we used the original gaze patterns and not the heatmaps in order to have a set of points to make possible the comparison with the objective points.

#### 3.1.1 Clustering parameters

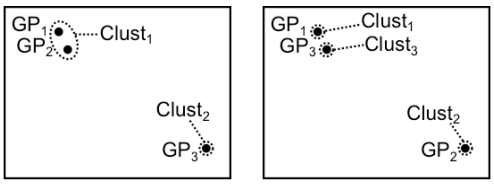
It is not adapted to use the original gaze points because they may be concentrated in a small region while the detectors try to avoid to have a concentration of points. Moreover, the aim of eye-tracking experiments is to produce the average behavior of a human. So we choose to adapt the proposed filtering/clustering[12] in order to create a new set of points than represent this average behavior of human observers. The adapted filtering algorithm is detailed in 3.1.1. The main idea is to create a collection of clusters $C_{collection}$ where each cluster $C_x$ in $C_{collection}$ includes many Gaze Points $GP_x$. The aim is to reduce the number of GP by aggregating all adjacent $GP_x$ where the acceptance distance is fixed by the threshold $Tclus$. Each $C_x$ can be weighted by the number of aggregated $GP_x$. A final step removes all $C_x$ which do not have aggregated enough $GP_x$ where the minimum number of $GP_x$ is defined by $Fmin$. We use $Tclus = 20$ and $Fmin = 4$ like the original algorithm.

**Algorithm 1** adapted gaze points filtering algorithm

```
create empty cluster collection C_collection
{the first GazePoint GP_1 is a special case}
create the first Cluster C_1
add GP_1 in C_1
add C_1 in C_collection
for i = 2 → numberOfGP do
  find the Cluster C_find in C_collection which minimize the euclidean distance D between the coordinate of GP_i
  and the coordinate of C_find
  if D<Tclus then
    add GP_i in C_find
  else
    create new Cluster C_new
    add GP_i in C_new
    add C_new in C_collection
  end if
end for
for j = 1 → numberOfCluster do
  if numberOfGPin Cj<Fmin then
    remove Cj in C_collection
  end if
end for
```

The algorithm proposed in[12] computes only the distance between the current GP and the current Cluster. So the clustering method is dependent of the apparition order of GP (example in Fig 4). In our case, we look for the average behavior of a human, so we search for the cluster which minimizes the distance with the current GP before to check the acceptance of this GP in the founded cluster. So the clustering becomes relatively invariant to the apparition order of GP and allows to aggregate close GPs from different observers.

Figure 5 shows the effect of filtering where the color and the opacity of each circle inform about the importance of each point (the number of aggregated points). The radius of each circle is defined by the far point aggregated




**Clustering A** **Clustering B**

Figure 4. effect of GP order in Clustering method

during the clustering process. A red, opaque and small little circle informs than there are a lot of near gaze points aggregated in this cluster. At the opposite a yellow, transparent and large circle informs than there are few scatter gaze points in this cluster. For the subjective gaze point before clustering (Fig.5-a), the fixation of each viewer has the same importance, so the color is red and we fix the opacity to 10%. In this case, a red opaque region informs that many viewer had looked at the same position. In this example our filtering reduces the number of clusters by a factor close to 50%, reduce the number of circles superposition and emphases the difference between the regions where all human are in accordance and the regions where only few viewers looked at.

(a) subjective points (7328 pts) (b) original filtering (387 pts) (c) used filtering (164 pts)

Figure 5. Subjective Gaze Points from 15 observers before and after filtering/clustering

## 4. STATISTICAL ANALYSIS FOR SIMILARITY MEASUREMENT

Thus study intends to measure the degree of correlation/similarity between the subjective gaze points and the objective interest points of several well-known detectors. For each of the latter, we look for the best setting in term of maximization of likeness with the gaze points. For this task, the Earth Mover's Distance (EMD)[1] is used to compare two datasets with different cardinalities. We also use ANOVA to measure the influence of each parameter involved in the detectors' settings as well as the possible introduced bias.

### 4.1 Earth Mover's Distance (EMD)

The number of interest points extracted by each detector depends on the used parameters and is often different from the number of subjective gaze points. So, for the measurement of similarity between objective and subjective interest points, it is necessary to use a distance able to compare two datasets with different cardinalities. We selected the Earth Mover's Distance (EMD) based on a solution to a special case of the old transportation problem[13] and implemented by Rubner et al.[1] in the context of image retrieval. The EMD Distance is based on the concept of measurement of work needed to move masses of earth into different holes. Precisely, given two distributions, one can be seen as a mass of earth properly spread in space, the other as a collection of holes in that same space. Then, the EMD measures the least amount of work needed to fill the holes with earth. Here, a unit of work corresponds to transporting a unit of earth by a unit of ground distance. In our case, we consider than the subjective gaze points are the holes and the interest points are the earth. Moreover this metric may assign weighting factors on holes and masses of earth. Since the subjective gaze points are clustered (as

explained in section 3.1.1), we can use the number of aggregated points to weight the holes in order to reflect regions of images having the saliency. Furthermore, each objective detector not only indicates the location of interest points but also give additional information. The latter can be the response (R in equation 4) for Harris and the size and the hessian for SURF.

The idea behind using the EMD distance is to be able to estimate which detector and associated settings minimize the cost of the transformation between the subjective points (holes) and objective points (earth masses). The experimental analysis is discussed in the following sections.

## 4.2 Experimentation for the maximization of similarity

To this point, we have already described the used interest points detectors, our method to exploit the gaze points and the metric for similarity measurement. As discussed previously, each of the interest points detectors has several parameters that influence the location and the number of the detected points. In order to find the most suitable parameters, the detectors have been run with thousands of combinations representing their different ranges of values. This operation is applied on the whole set of images from the VAIQ database.[12] Hence, the EMD cost is obtained for each setting applied on each image. Let remind that a low EMD distance means a high similarity between the subjective and objective datasets. However, the same setting will not lead automatically to a minimization of the EMD distance for the whole images. For example, the VAIQ database contains a variety of content including faces, people, animals, close-up shots, wide-angle shots, nature scenes, man-made objects, images with distinct foreground/background configurations, and images without any specific object of interest. In this experiment, there is no specific task to address and no targeted types of images. This is why our focus was mainly on the average distance.

### 4.2.1 Experience 1

There is no previous knowledge neither about the similarity between interest points and gaze points nor on the ranges of parameters' values. Our exploration started without any *a priori* about the setting but with some experience about the usage of each detector. So, the initial ranges of values are listed in tables 3, and 4 respectively for Harris and SURF.

Table 3. Harris parameters

| | |
|---|---|
| ql | $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ |
| md | $\{20, 40, 60, 70, 80, 100, 120\}$ |
| bs | $\{3, 6, 9, 12\}$ |
| k | $\{0.04, 0.08, 0.12, 0.20\}$ |

Table 4. SURF parameters

| | |
|---|---|
| ht | $\{100 : 100 : 6900\}$ |
| no | $\{1, 2, 3, 4, 6, 9, 12, 18\}$ |
| nol | $\{1, 2, 3, 4, 6, 9, 12, 18\}$ |

Since the EMD distance can take in consideration a weighting for each point, we choose to perform this study with and without weightings. When they are not used, each extracted point will be equally weighted. So, in this case only location is taken into account. In the opposite case, additional data such as the strength and the size could be used to study whether the similarity with the human gaze is increased.

As first experiments, the two selected detectors have been run using the combinations of parameters given in tables 3 and 4 with and without weighting. Table 5 (resp. table 6 ) summarizes the results by giving the mean, the minimum, the maximum and the standard deviation of the EMD cost for the best setting without (resp. with) weighting.

Table 5. EMD cost for the best detector's setting without weighting

| Detector | mean | min | max | std |
|---|---|---|---|---|
| Harris | 0,02290 | 0,01547 | 0,04148 | 0,00546 |
| SURF | 0,00132 | 0,00027 | 0,00480 | 0,00092 |

From the results of table 5, one can notice that best settings of the two detectors provide results in two different ranges. SURF appears to be the best with regards to human gaze followed by Harris. This rank is also

Table 6. EMD cost for the best detector's setting with weighting

| Detector | mean | min | max | std |
|---|---|---|---|---|
| Harris (Response) | 72,312 | 12,28 | 175,14 | 39,23 |
| SURF (Hessian) | 44,79 | 6,66 | 130,52 | 28,38 |
| SURF (Scale) | 42,04 | 2,71 | 119,28 | 27,02 |

confirmed on table 6 where some properties are used as weightings for each detector. For the SURF case, two different weightings have been experimented namely $\text{SURF}_{scale}$ and $\text{SURF}_{hessian}$ corresponding respectively to the size of the detected object and its strength. $\text{SURF}_{scale}$ performs better than $\text{SURF}_{hessian}$ and may lead to a first conclusion about the influence of the size. One can also notice than the scores of this two tables are very different. So it is difficult to conclude about the usage of additional information for increasing the similarity with the subjective behavior. This raises other questions about the difference between scores and the best settings obtained with and without weightings. To answer these questions for the case of Harris, table 7 (a) and (b) give respectively the top ten settings (minimizing the mean EMD distance) with and without weightings.

Table 7. Top ten Harris settings

| ql | md | bs | k | mean |
|---|---|---|---|---|
| 1,00E-06 | 20 | 3 | 0,04 | 0,0229 |
| 1,00E-06 | 20 | 3 | 0,08 | 0,0236 |
| 1,00E-06 | 20 | 3 | 0,12 | 0,0244 |
| 1,00E-06 | 20 | 6 | 0,04 | 0,0244 |
| 1,00E-06 | 20 | 6 | 0,08 | 0,0253 |
| 1,00E-06 | 20 | 9 | 0,04 | 0,0256 |
| 1,00E-06 | 20 | 6 | 0,12 | 0,0259 |
| 1,00E-06 | 20 | 3 | 0,2 | 0,0262 |
| 1,00E-06 | 20 | 9 | 0,08 | 0,0263 |
| 1,00E-06 | 20 | 12 | 0,04 | 0,0266 |

(a) Harris without weighting

| ql | md | bs | k | mean |
|---|---|---|---|---|
| 1,00E-06 | 20 | 9 | 0,04 | 72,313 |
| 1,00E-05 | 20 | 9 | 0,04 | 72,314 |
| 1,00E-04 | 20 | 9 | 0,04 | 72,331 |
| 1,00E-06 | 40 | 9 | 0,04 | 72,460 |
| 1,00E-05 | 40 | 9 | 0,04 | 72,461 |
| 1,00E-06 | 40 | 6 | 0,04 | 72,462 |
| 1,00E-05 | 40 | 6 | 0,04 | 72,463 |
| 1,00E-04 | 40 | 9 | 0,04 | 72,468 |
| 1,00E-04 | 40 | 6 | 0,04 | 72,468 |
| 1,00E-03 | 40 | 6 | 0,04 | 72,557 |

(b) Harris with weightings

Several remarks can be made from the previous table. For example, ql=1,00E-06 and md=20 are the values for the top ten experiments without weighting and lead to the conclusion that these values are the most important for Harris. Moreover, they corresponds to the lower bounds of the parameters' ranges. So, the minimization of this parameters appears to be the most important to reduce the cost of the EMD distance. In the Harris definition, *md* tunes the distance between detected points whereas *ql* defines the strength threshold. Low values of these two parameters increase the number of detected points. It can explain the minimization of the EMD distance because of the large availability of points to be matched with human gaze. The behavior of the Harris parameters is also visible on the graphical representations given in figure 6. Each curve corresponds to the variation of the mean EMD distance versus parameter value for *ql*, *md*, *bs* and *k*.

As mentioned already, the EMD distance seems to be sensitive to the number of detected points. For the Harris detector and the associated settings (table 3), the average range of detected points is between [13,02-473,73]. This range is higher for SURF [34,11-5515,30]. We noticed that for a low number of interest points the EMD distance is high and for a large number of interest points the EMD cost tends to zero. This means that the number of points has an influence on EMD and that the two detectors have not been used equally.

To statistically confirm the hypothesis of EMD sensitivity to the number of detected points and to study the influence of each parameter, we used ANOVA. First, there are two essential hypothesis to be checked to use anova: Distributions must be gaussian and variance of distributions must be equal. The second hypothesis can be controlled using a Bartlett test.

ANOVA is a statistical test. So like others statistical tests, it compares two hypothesis :
$H_0$ (null hypothesis) means of distributions are equal. In our experiment, this implies that the parameter has

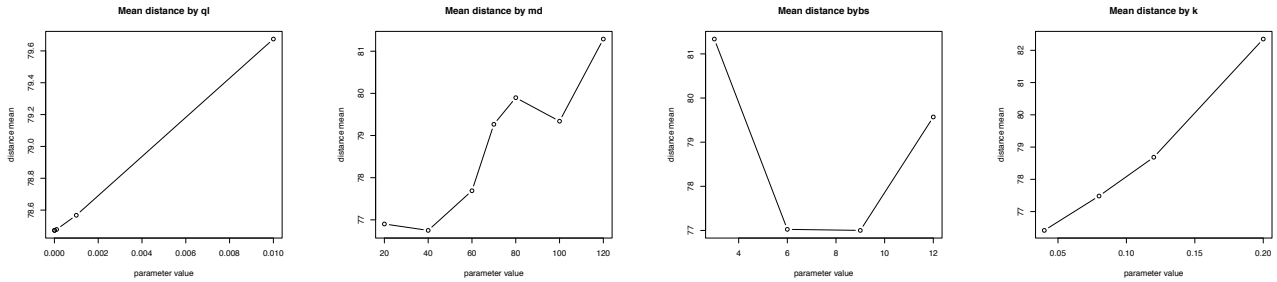



Figure 6. Mean distance for each parameters

no influence.
$H_1$ (alternative hypothesis) : there exists at least one distribution whose mean is different from the others.
ANOVA returns a "p-value" that determines the influence of a parameter. Closer is the value to 0, more influent is the parameter. In practice, when the p-value is less than 0.05, we reject the null hypothesis, and it means that the parameter has an influence.
Bartlett's test of homogeneity of variances confirms that we can use ANOVA in our set for the four parameters explained previously ql,md,bs and k. Results are summarized in Tables 8 and 9. Note that, for column "Influence", an empty box means that the parameter has no influence, and the number of $*$ represents the intensity of the influence (Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1  1).

Table 8. Influence of Harris' parameters on EMD without weighting

| Parameters | p-value | Influence |
|---|---|---|
| ql | $< 2.2e-16$ | $***$ |
| md | $< 2.2e-16$ | $***$ |
| bs | 0.07815 | . |
| k | $2.686e-05$ | $***$ |
| nbPtsMean | $< 2e-16$ | $***$ |

Table 9. Influence of Harris' parameters on EMD with weighting

| Parameters | p-value | Influence |
|---|---|---|
| ql | 0.4632 | |
| md | $6.985e-07$ | $***$ |
| bs | $1.04e-11$ | $***$ |
| k | $< 2.2e-16$ | $***$ |
| nbPtsMean | 1 | |

From the results above, one can notice than the null hypothesis is rejected for the parameters nbPtsMean in table 8 and accepted in table 9. So the mean number of points has an influence on the EMD distance for the measurement without weighting, and no influence with weighting. The plausible explanation for this modification of influence lies in the fact of giving an importance to some points in the case of weighting which minimizes the impact of the remaining points and consequently their number. This is completely different in the case of no weighting because each point has an equal influence and the total number becomes thus influent.

This analysis brings interrogations about the conclusion drawn from table 5 which may be biased. So the best setting determined for each detector is probably not the best one. Moreover, the comparison between detectors is not reliable in this case because the EMD cost is too dependent on the minimum and maximum number of detected points within the tested range. The second experience will take this aspect into account for reducing the observed bias.

### 4.2.2 Experience 2

In order to reduce the bias observed in experience 1, we constrained the number of detected points for the two detectors. This constraint is used to reject all settings producing a mean number of detected points outside the $filtI$ interval. This interval has been defined thanks to an analysis of the subjective data contained in the VAIQ database (cf. section 3.1). In the database, the number of gaze points by image and by user is between [480-540]. Even though each subject observed the same image with the same duration, the number of gaze points is often different. Each subject has its own saccades and fixations. So we choose the $filtI$ interval equal to [420-600] to be as close as possible from the human behavior while allowing several parameters' settings for each detector.

The parameters' ranges for experience 2 changed for Harris because the former ones did not allow to respect the $filtI$ constraint. These parameters are given in tables 10. No additional parameters are tested for SURF due to of its high number of allowed settings (511).

Table 10. Harris parameters

| | |
|---|---|
| ql | $\{10^{-6}, 10^{-7}, 10^{-8}\}$ |
| md | $\{0.1, 1, 10, 20, 30, 40, 50\}$ |
| bs | $\{3, 6, 9\}$ |
| k | $\{0.005, 0.02, 0.04\}$ |

By using the parameters values given in the previous table, the same protocol has been used as in experience 1. The top ten settings allow thus to analyze deeply the influence of each parameter on the minimization of the distance between interest points and gaze points. Hence, the results of this experience are illustrated in table 11 for Harris, table 12 for SURF weighted using the size and finally table 13 for SURF weighted using the hessian.

Table 11. Top ten Harris settings for experience 2

| ql | md | bs | k | mean | min | max | std |
|---|---|---|---|---|---|---|---|
| 1,00E-07 | 20 | 9 | 0,005 | 71,056 | 11,874 | 180,221 | 39,862 |
| 1,00E-08 | 20 | 9 | 0,005 | 71,057 | 11,874 | 180,221 | 39,862 |
| 1,00E-06 | 20 | 9 | 0,005 | 71,061 | 11,874 | 180,222 | 39,860 |
| 1,00E-07 | 20 | 6 | 0,005 | 71,316 | 9,963 | 180,570 | 39,911 |
| 1,00E-08 | 20 | 6 | 0,005 | 71,316 | 9,963 | 180,570 | 39,911 |
| 1,00E-06 | 20 | 6 | 0,005 | 71,317 | 9,963 | 180,570 | 39,911 |
| 1,00E-07 | 20 | 9 | 0,02 | 71,841 | 12,598 | 176,448 | 39,322 |
| 1,00E-08 | 20 | 9 | 0,02 | 71,841 | 12,598 | 176,448 | 39,322 |
| 1,00E-06 | 20 | 9 | 0,02 | 71,842 | 12,598 | 176,449 | 39,322 |
| 1,00E-07 | 20 | 6 | 0,02 | 72,032 | 11,496 | 179,781 | 39,527 |

For the harris detector, the $ql$ parameter receives values between 1,00E-08 and 1,00E-06 in order to keep interest points of medium and high strength. If the ql parameter is too high, the detector cannot detect enough points for our experiment (like in the range of experience 1). $md$=20 gives the best results which confirms the results of experience 1. This parameter defines the minimum distance between two detected points. It seems that $md$ behaves similarly to the parameter $Tclus$ introduced in section 3.1.1 but it needs more investigation to prove it. The parameter $bs$ is equal either to 6 or 9 for the top ten settings. Since it represents the size of observation window, this results means that very small corners cannot be detected. Finally, the values of parameter $k$ has to be very small with regards to the behavior of figure 6 in order to be kept in the top ten. When this parameter is minimized, the detected point can be a horizontal/vertical contour and not only a corner.

For the SURF detector, $ht$=800 is selected to minimize the EMD distance as shown in figure 8. This parameter is used to reject the very low points defined by their local contrast. The parameter $no$ can take several values (cf. table 12), but figure 8 indicates a stabilization from the value 4. So, we selected a value around 5 and avoided to select higher values to limit the calculation cost. For the parameter $nol$, the value 1 has been selected because when this parameter increases, the number of detected points increases fast and not allows with respect to $filtI$.

By looking to the two tables 12 and 13, we can see that the best settings are the same for the two cases, but the scale weighting have a better cost than the hessian weighting. We can conclude than the location of the points are good but the weighting by a size is better than a weighting by the hessian. The figure 7 shows that the information about size and strength can be complementary. In this example a strong point is detected on the text on the yellow cap by the hessian weighting, but the red cap is highlight by the scale weighting.

With this second experimentation, we can see on table 14 that all the mean scores are different from those obtained in experience 1 (table 6). However, the rank order is still the same ($\text{SURF}_{scale}$, $\text{SURF}_{hessian}$ and Harris).

Table 12. Top ten SURF (weighting Size filtered)

| ht | no | nol | mean | min | max | std |
|---|---|---|---|---|---|---|
| 800 | 9 | 1 | 47,906 | 5,045 | 138,739 | 30,645 |
| 800 | 18 | 1 | 47,906 | 5,045 | 138,739 | 30,645 |
| 800 | 6 | 1 | 47,906 | 5,045 | 138,739 | 30,645 |
| 800 | 12 | 1 | 47,906 | 5,045 | 138,739 | 30,645 |
| 800 | 4 | 1 | 48,702 | 6,408 | 138,739 | 30,824 |
| 900 | 6 | 1 | 48,873 | 7,564 | 140,738 | 31,285 |
| 900 | 12 | 1 | 48,873 | 7,564 | 140,738 | 31,285 |
| 900 | 9 | 1 | 48,873 | 7,564 | 140,738 | 31,285 |
| 900 | 18 | 1 | 48,873 | 7,564 | 140,738 | 31,285 |
| 1000 | 18 | 1 | 48,989 | 4,738 | 142,834 | 32,134 |

Table 13. Top ten SURF (weighting Hessian filtered)

| ht | no | nol | mean | min | max | std |
|---|---|---|---|---|---|---|
| 800 | 9 | 1 | 56,992 | 8,055 | 153,909 | 35,123 |
| 800 | 18 | 1 | 56,992 | 8,055 | 153,909 | 35,123 |
| 800 | 6 | 1 | 56,992 | 8,055 | 153,909 | 35,123 |
| 800 | 12 | 1 | 56,992 | 8,055 | 153,909 | 35,123 |
| 800 | 4 | 1 | 57,138 | 8,448 | 154,162 | 35,134 |
| 900 | 6 | 1 | 57,735 | 8,541 | 155,529 | 35,534 |
| 900 | 12 | 1 | 57,735 | 8,541 | 155,529 | 35,534 |
| 900 | 9 | 1 | 57,735 | 8,541 | 155,529 | 35,534 |
| 900 | 18 | 1 | 57,735 | 8,541 | 155,529 | 35,534 |
| 900 | 4 | 1 | 57,887 | 8,922 | 155,792 | 35,548 |

subjective GP | $\text{SURF}_{hessian}$ IP | $\text{SURF}_{scale}$ IP

Figure 7. Maps of subjective and objective gaze points

Table 14. Comparison of best EMD distances with weighting (experience 2)

| Detector | mean | min | max | std |
|---|---|---|---|---|
| Harris (Response) | 71,056 | 11,874 | 180,221 | 39,862 |
| SURF (Hessian) | 56,992 | 8,055 | 153,909 | 35,123 |
| SURF (Scale) | 47,906 | 5,045 | 138,739 | 30,645 |
| Human (Duration) | 48,153 | 0,984 | 186,294 | 28,959 |

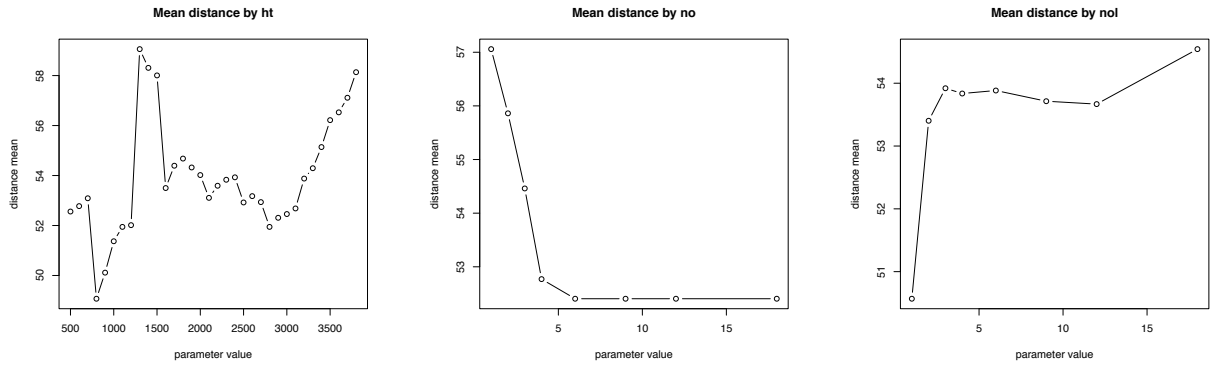



Figure 8. Mean distance for SURF$_{scale}$ parameters

Since Harris and SURF$_{hessian}$ are the worst, we study the parameters influence to determine if an improvement is possible (table 15 and 16). By using the ANOVA results, none of the parameters appears to have an influence on the EMD cost for Harris and SURF$_{hessian}$.

Table 15. Harris_response (range 2)

| Param. | p-value | Influence |
|---|---|---|
| ql | 1 | |
| md | | |
| bs | 0.3778 | |
| k | 0.860 | |
| nbPts | 1 | |

Table 16. SURF$_{hessian}$

| Param | p-value | Influence |
|---|---|---|
| ht | 0.757 | |
| no | 0.6201 | |
| nol | 0.0098 | ** |
| nbPts | 1 | |

Table 17. SURF$_{scale}$

| Param | p-value | Influence |
|---|---|---|
| ht | 0.04747 | * |
| no | $2.044e-08$ | *** |
| nol | 0.03675 | * |
| nbPts | 1 | |

Even though this study allowed to determine an order among the three tested configurations, there still some questions related to the interpretation of the EMD cost. For example, is EMD=47 a low and acceptable cost to reveal a similarity with the human behavior? To find an answer to this question, we decided to study the observers' data available in the VAIQ database. For each observer and each image, we calculate the EMD cost between computed between its gaze points and the average gaze points of all other observers. The latter is obtained by applying the clustering process described in section 3.1.1. In order to exploit the weightings in the EMD distance, we used the duration of the gaze points fixation to weight the observers' points.

Table 14 gives the values of EMD costs weighted by the duration. The mean EMD cost is around 48 representing the mean cost for an observer in comparison to the average of all the others. We can notice than SURF$_{scale}$ has an average cost under this value. So, it is possible to conclude that this detector behaves in a comparable way to a human. We can also notice a high maximum distance for the human case in comparison with all other detectors.

To conclude, all detectors can find interest points more or less in accordance with the human eyes. We can notice than all detectors take in consideration the local contrast to define the strength of the corner/edge/blob structure. The human attention is also guided by this low level property (local contrast of luminance and chrominance). We also noticed that the SURF detector takes also in consideration the size of the detected points as well as the human observer. Independently of the targeted task, the human observer is also attracted by the text area and in this case Harris performs better than the others.

## 5. CONCLUSION AND FUTURE WORKS

In this study, we have measured the correlation/similarity between the subjective gaze points (obtained by an eye-tracking experiment on a variety on pictures) and the objective interest points detected by Harris and SURF detectors. An important effort has been put in this work to minimize the measurement bias. The first possible bias was related the selected image database. As we wanted images representing a large variety of content,


Harris_response | $SURF_{hessian}$ | $SURF_{hessian}$

Figure 9. maps of subjective and Harris points

we used databases with the majority of the state-of-the-art images with pictures of faces, animals, landscapes, manufactured objects . . . The second possible bias concerns the quality and the reproducibility of the eye tracking experiment. For the selected eye-tracking data, it has been mentioned previously that another work has validated them.

In this work, we proposed an adaptation of the gaze points clustering process to have more comprehensive data with a better representation of the average behavior of observers. To measure the similarity between subjective and objective gaze points, the EMD distance has been used. This distance is interesting because it allows comparing two datasets with different cardinalities and by assigning a weighting factor to each point. Moreover, in comparison to other distances, EMD is a real metric which respect the triangle inequality.

By this study, We found for each detector the settings which maximize the similarity with the human gaze points by a large range of tests and analysis of the setting's parameters. We have analyzed the subjective data in order to compare and rank the objective detectors. We found that the EMD cost to transform the SURF points to the subjective clustered gaze points is under the cost to transform an observer to the whole observers data. As a conclusion, all detectors can detect interest points in accordance with the human visual system function of the selected settings. We noticed than all detectors take in consideration the local contrast as the HVS does.

In this experiment, we considered the size and the strength of the SURF interest points separately. It will be interesting to combine these two informations in order to increase the similarity with the human gaze points. Moreover, since the detection of blob-like structure appears suitable to mimik the HVS, it is interesting to compare SURF with other blob detectors such as SIFT detector.

Finally, it seems that with suitable settings, the interest points detectors can generate a kind of simplified saliency map that can be used for automatic applications like quality assessment or image coding. They can

also be combined with additional informations like the central bias, the face detection, ... in order to increase the efficiency of the saliency map construction.

## ACKNOWLEDGMENTS

This work has been supported by the project CAIMAN funded by the French Research Agency.

## REFERENCES

[1] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *IEEE International Conference on Computer Vision*, pp. 59–66, Jan 1998.

[2] A. L. Yarbus, "Eye movements and vision," in *Plenum Press*, 1967.

[3] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," in *Journal of Vision 9(12):10*, pp. 1–15, 2009.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* **20**(11), 1998.

[5] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision* **7**(14), pp. 1–17, 2007.

[6] C. Harris and M. Stephens, "A combined corner and edge detector," in *4th Alvey Vision Conference*, (Manchester), 1988.

[7] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision*, **2**, pp. 1150–1157, 1999.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)* **110**(3), pp. 346–359, 2008.

[9] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis & Machine Intelligence* **27**(10), pp. 1615–1630, 2005.

[10] M. Nauge, M.-C. Larabi, and C. Fernandez, "A reduced-reference metric based on the interest points in color images," in *PCS2010*, pp. 610–613, (Nagoya, Japan), 2010.

[11] M. Nauge, M.-C. Larabi, and C. Fernandez, "A hierarchical saliency map generation based on the human visual system properties," in *WPCIP'2010*, (Nagoya, Japan), 2010.

[12] U. Engelke, A. J. Maeder, and H.-J. Zepernick, "Visual attention modeling for subjective image quality databases," in *IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, Oct 2009.

[13] G. B. Dantzig, "Application of the simplexmethod to a transportation problem," in *Activity Analysis of Production and Allocation, JohnWiley and Sons*, pp. 359–373, 1951.

[14] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *ICCV*, pp. 525–531, 2001.

[15] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools* , 2000.

[16] U. Engelke, H.-J. Z. H. Liu, I. Heynderickx, and A. Maeder, "Comparing two eye tracking databases: The effect of experimental setup and image presentation time on the creation of saliency maps," in *IEEE Picture Coding Symposium (PCS)*, Dec 2010.