Can fixation-point and key-point coincide on cultural heritage colour paintings?

Hermine Chatoux, Noël Richard, François Lecellier, Christine Fernandez-Maloigne

Can fixation-point and key-point coincide on cultural heritage colour paintings?

Hermine Chatoux, Noël Richard, François Lecellier and Chrisitne Fernandez-Maloigne

Abstract

This article compares the key-point extracted with a colour key-point detector and the location of fixation point thanks to an eye-tracking experiment. We hypothesize the first fixation points should correspond to key-points with the most significant gradients. The colour detector is based on Harris and Stephens corner detector extended to colour. The eye-tracking experiment was realised on medieval art work. We compare the location of both the detected key-points and the fixation points observed. Between 30% to 50% of the key-points coincide with fixation points. A second analysis display the number of matching key-points over the number of fixation points at a given time. The ratio decreases with the observation time which concur our initial hypothesis. Overall, several of the first fixation points correlate with high response key-points detected with our method.

1 Introduction

Several eye-tracking experiment have been conducted to estimate saliency maps. [Borji and Itti(2012)] offered a wide survey of the different methods used to compare an estimated saliency map. The estimation can focused on top-down analysis (task related) or bottom-up exploration (pre-attentive step). [Borji et al.(2011)Borji, Sihite, and Itti] proposed classification tasks to extract top-down saliency maps. On the contrary [Le Meur(2005)] used low-level features to perform bottom-up saliency maps. These maps are based on the location of fixation points: a point focused by a human eye for sufficiently long time.

It is important to define what represent a key-point: it is a point or an area allowing to characterize the analysed image. Two major family extract key-point, the corner point detection and the *blob* detection. The most used detector is the Difference of Gaussian (DoG) presented by [Lowe(1999)] belongs the the *blob* family. It is based on the second derivative and extract uniform areas. [Bay et al.(2008)Bay, Ess, Tuytelaars, and Van Gool] proposed an adaptation that was computationally faster. The corner detection family can be decomposed in two group: one based on the first derivative and based on pixel comparison. The derivative based group was introduced by [Moravec(1980)] and generalized by [Harris and Stephens(1988)]. They rely on the spatial auto-correlation matrix and study the eigenvalue to determine which pixels are key-points. The second group compares a value set as the centre with values at a given radius from the centrer. It was introduced by [Smith and Brady(1997)]. [Rosten and Drummond(2006)] accelerated the computation with the FAST detector which is often used.

In this article, we wonder about a possible relation between detected key-points and salient fixation point. Key-point detection is based on low-level features. Therefore we expect a better match between key-points and fixation point in the pre-attentive phase. The rest of the observation is linked to the brain analysis reading of the image. So, low-level features are not the main criteria. However, our goal is not to prove both detected key-points and fixation point coincide. But we are interested in estimating what matching we can expect.

The next section will present the two colour key-point detectors used in the comparison. Section 3 presents the eye-tracking experiment: device, images and observers group. The fourth section compares the detected key-points with the fixation points obtained during the experiment. We conclude in the last section.

2 Colour key-point detector

This *vector-key-point* detector is based on the same steps as the [Harris and Stephens(1988)] corner detector. The figure 1 presents these steps which are separated in three phases. First, the gradient needs to be measured, then the corner informations are extracted and finally the key-point decision is made.

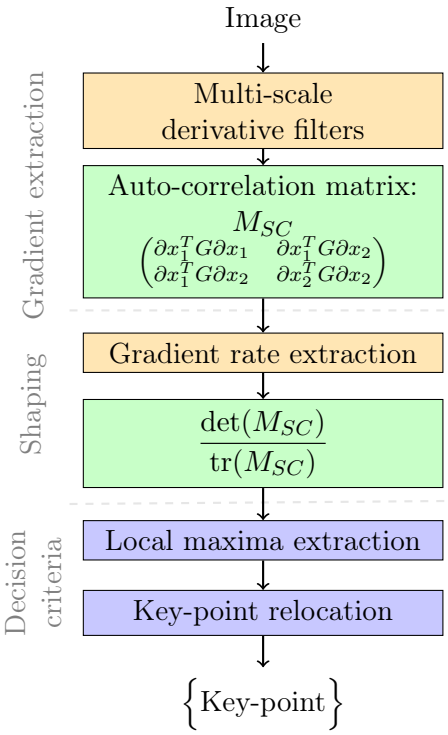


Figure 1: Steps to obtain full-vector key-point

2.1 Gradient extraction

As an image contains corners at different scale, we used multi-scale derivatives filters. [Mikolajczyk and Schmid(2 proposed multi-scale binomial derivative filter. To be more generic, we construct our filter with a Gaussian and its derivative. Equation (1) illustrates these filters:

$$\frac{\partial Filter}{\partial x_i} = \left(\quad \right) \times \left(\quad \right)^T . \tag{1}$$

The Gaussian function is centred, the spatial filter size (S_F) depends only on the standard deviation σ selected:

$$S_F = (6\sigma + 1) \times (6\sigma + 1).$$

We selected eight following scales form $\sigma = 1$ to 8. Usually, the scales are selected with a constant size ratio. We chose a constant six pixels difference between each scale. It allows a redundancy in certain key-point location. The more scale detected at a key-point, the stronger it is.

These filters allows to measure, marginally, the gradient on every channel. The next step is to combine these gradients.

Inspired form [Di Zenzo(1986), Koschan and Abidi(2005)], we proposed the *full-vector gradient* in [Chatoux et al.(2019a)Chatoux, Richard, Lecellier, and Fernandez-Maloigne] to extract

gradients in the sensor space. We calculate a spatial correlation matrix (M_{SC}) depending on all the hyper-spectral channels and their inter-correlations from the Jacobian:

$$M_{SC} = \begin{pmatrix} \left\| \frac{\partial I(x)}{\partial x_1} \right\|^2 & \left\langle \frac{\partial I(x)}{\partial x_1}, \frac{\partial I(x)}{\partial x_2} \right\rangle \\ \left\langle \frac{\partial I(x)}{\partial x_1}, \frac{\partial I(x)}{\partial x_2} \right\rangle & \left\| \frac{\partial I(x)}{\partial x_2} \right\|^2 \end{pmatrix}, \tag{2}$$

with

$$\left\langle \frac{\partial I(x)}{\partial x_1}, \frac{\partial I(x)}{\partial x_2} \right\rangle = \frac{\partial I(x)}{\partial x_1}^T \cdot G \cdot \frac{\partial I(x)}{\partial x_2}, \tag{3}$$

$$G = \begin{pmatrix} \|s_0\|_2^2 & \langle s_0, s_1 \rangle_2 & \cdots & \langle s_0, s_m \rangle_2 \\ \langle s_1, s_0 \rangle_2 & \|s_1\|_2^2 & \cdots & \langle s_1, s_m \rangle_2 \\ \vdots & \vdots & \ddots & \vdots \\ \langle s_m, s_0 \rangle_2 & \cdots & \langle s_m, s_{m-1} \rangle_2 & \|s_m\|_2^2 \end{pmatrix}. \tag{4}$$

The Gram matrix G uses the scalar products defined for the integrable functions. The functions used are the Spectral Sensitivity Functions (SSF) of each channel:

$$\langle s_i, s_j \rangle_2 = \int_{\mathbb{R}} S_i(\lambda) S_j(\lambda) d\lambda. \tag{5}$$

The correlation matrix will allows us to extract key-point as presented by [Harris and Stephens(1988)].

2.2 Shaping

An intermediate step is necessary to extract the strongest gradients represented by corners. The gradient rate extraction consist in a simple integrative filter. It is often considered as a denoising step in other algorithms ([Harris and Stephens(1988), Mikolajczyk and Schmid(2001)]). With the filters we proposed, the smoothing step is already realized. Yet, an integrative filter adds all the gradients from its window. A corner area contains more edges than an edge one, therefore after the integrative filter, the response from the corner area will be stronger than the edge one. This step allows to increase the response on corner areas.

For the appropriate response function, we based our analysis on the principle given by [Harris and Stephens(1988)]:

- 2 small eigenvalues represents uniform area,
- 1 strong eigenvalue represents edge,
- 2 strong eigenvalues imply a corner area.

The response proposed by [Harris and Stephens(1988)] depends on both matrix invariants: the determinant and the trace. The proposed response function is $R_H(k) = \det(M_{SC}) - k\,\mathrm{tr}(M_{SC})^2$, k being an empirical constant.

To free ourselves from the instability implied by the constant k, we propose a new response:

$$R_{FVKP} = \frac{\det(M_{SC})}{\mathrm{tr}(M_{SC})}. \tag{6}$$

2.3 Decision criteria

Corner areas will give high value on the response function. To select only the corner, we extract the local maxima of high responses. These maxima are labelled corners. Once the key-point have been selected, several informations are attached to it apart from its location: the response value and an angle.

The spatial direction θ of the gradient is defined by [Jin et al.(2012)Jin, Liu, Xu, and Song] lifting the imprecision of $\pm\frac{\pi}{2}$ from the initial Di Zenzo expression.

Another step can be added to merge key-points. A corner can be detected at several filter sizes. It is not relevant to keep each size for the same location. Therefore, we propose two different extractions of the key-point.

The first one is based on suppressing overlapping key-points. We keep the key-point with the strongest response function when there is a large overlap between two key-points. It will be called Harris_{S_2} on the rest of the paper.

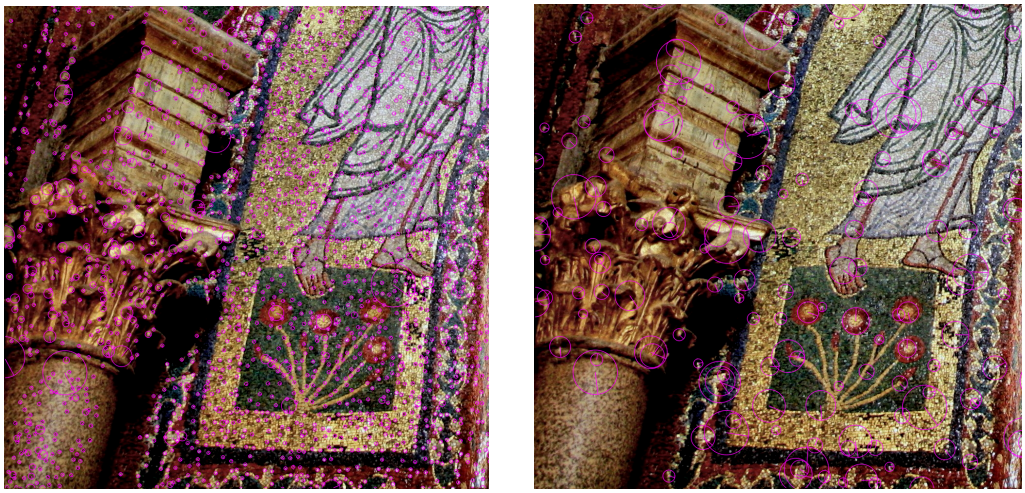
The second detector consider the set of key-points associated to the same location with different scales. We hypothesize that the more scale are detected the stronger the corner is. This detector, Harris_{ME}, will keep a key-point if it has been detected on 3 scale or more and associate the number of detected scales as the response function. The figure 2 presents the results of both detector on the same image. We observe that the detector Harris_{ME} is more selective than the other. The constraint of several scale of detection allows in this image to remove all key-point associated to a mosaic tile.

3 Eye-tracking experiment

To compare the detectors with the human eye, we realized a psycho-visual experiment allowing to eye-track the user during the reading of an image.

The eye-tracker An eye-tracker and its associated software analysed the fixation point of each observer. An eye-tracker is a device that allow to follow the eye movement of a user. The device measured the eye movement observing a screen. It is based on infra-red that will be reflected by pupils only.

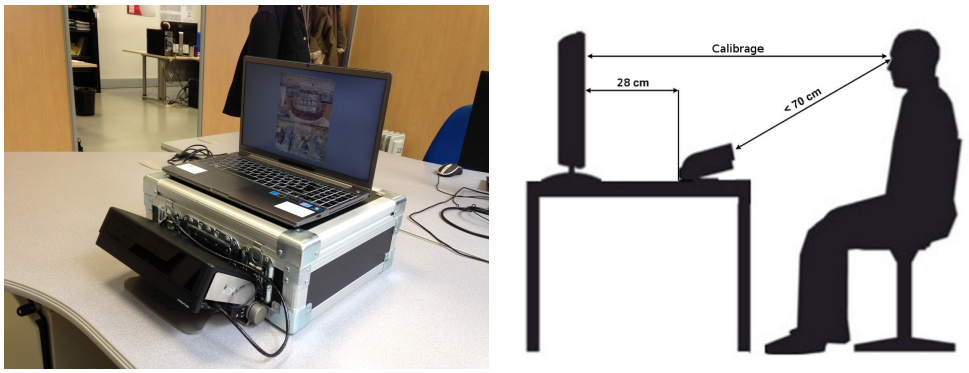
The eye-tracking device used id the *Tobii X-120*. It allows to measure two majors eye-movements. Firstly, there is the saccades and then fixation points. The saccades are high speed movements ($<$50 ms pause between movement). There is no apparent pattern to their trajectory. These saccades allows to create a first image that will allows the brain to extract the area of interest that will be modelled by the fixation points. These points are obtained when the eye is still for a sufficiently long time ($>$300 ms). It allows the brain to analyse the area. The fixation points are the one we want to compare with key-points detected with the detectors.



(a) 10522 points extracted by the detector Harris_{S_2}

(b) 687 points extracted by the detector Harris_{ME}

Figure 2: Examples of key-points extracted by both detectors



(a) Experimental settings (b) Experimental diagram

Figure 3: Experimental protocol for eye-tracking on medieval works.

The eye-tracking device needs to be calibrated to precisely localise where the observer is looking. The Figure 3 presents the experimental diagram and a photo of the settings during the experiment.

As the experiment was to be held in two different environments, we chose to display the images on a 17 inches screen. The eye-tracking devices used was created for larger screens. Therefore, observers were watching small images considering the perception distance forbidding a great sensitivity in the details.

The images We chose to set the test duration short (12 to 15 min) in order to keep the observers attention focused during all the experiment. We selected 20 images of medieval works.

Among these, half of it came from the Romane database created by the CESCM (Centre d'Études Supérieur de la Civilisation Médiévale) where come most of our observers. These images are defined as known by the observers.

Ten other images have been extracted on other medievals work databases. These are considered unknown for the "experts".

Most of the images are wall paintings but there are also mosaics, wood paintings... Some of them present a religious theme but other themes have been selected. Images are well preserved other are very degraded. The figure 4 presents some of the used images.

Each images is presented for 30 secondes with 5 secondes of neutral grey in between to rest the eyes. The oberver has to describe the seen image to identify the scene when possible.

Sample group One of the experiment objectives was to verify how different the analysis is done depending on the knowledge of the observer. We have selected two groups to realise the experiment.

The first one is constituted of students from BD to PhD, researchers specialised in the medieval history. 21 observers are in this group.

The second are members of the XLIM laboratory (researchers, administrative or IT personnels). This sample group is unfamiliar with medieval images even if some are image processing users. 16 persons realised the experiment in this group.

Unfortunately some users cannot be kept in the study. Indeed, glasses or moving during the experiment prevented the eye-tracker device to measure sufficiently the eye movement. We kept 31 observers.

4 Comparing detectors key-points and saliency points

In this study, we are looking for a relation between detectors key-points and saliency points. We have extracted key-points with the colour detectors Harris_{S_2} and Harris_{ME} on some of the images

(a) Known image of wall painting

(b) Known image of deteriorated wall painting

(c) Unknown image of mosaic

(d) Unknown image of wall painting

Figure 4: Examples of images used for the experiment.

experiment. To compare as close as possible from the human eye, we have transformed the images in the LMS colour space thanks to the standard matrix CIECAM02 and used the Gram matrix presented in [Chatoux et al.(2019b)Chatoux, Richard, Lecellier, and Fernandez-Maloigne]associated with the LMS curves from [Shrestha(2016)].

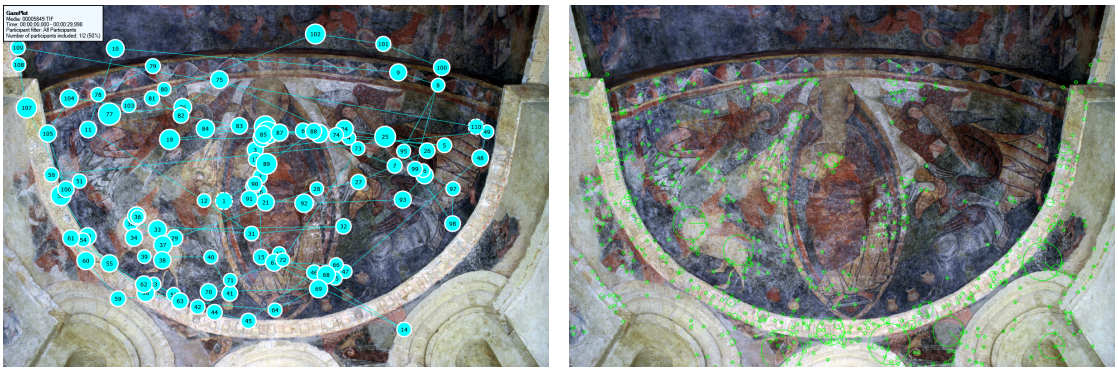
4.1 First results and works hypothesis

The figures 5 presents results on fixation points for an observer compared to the detected key-points for two images. On the first image (fig. 5a), points seems dispersed on the central area (coloured apse). On this image other observers focused less on the frieze inducing larger difference with the detector (fig. 5b). On the second observed image (fig. 5c), fixation points are focused in the remaining painted area avoiding the degraded areas while the detectors (fig. 5d) extracts mostly in the edges between painted and plaster areas as they present strong gradients. This shows a first limits in the analysis of a relation between detected key-points ans salient points. A strong gradient does not necessarily induce high saliency. The number of key-point being superior with the other detector, the percentage of key-point on an edge painting/ plaster is smaller hence the correlation with saliency point is better.

It is worth noting, that the saliency central bias is not considered in this study. The detected key-points are extracted on the whole image while human looks preferably at the centre of the image. Considering the observation duration, the observer should have time to watch the whole image.

The table 1 gives a first results of the relation between detected key-points and fixation points. It presents the percentage of detected key-points located near (less than 10 pixels radius) of a fixation point. Overall 35% of the detected key-points are scrutinize by observers. It gives us a first limits to the expected relation, even if four images are insufficient to draw a conclusion.

The figure 6a presents the number of fixation points over time. These curves corresponds the visual course presents in figures 5a and 5c. With the description task given to the observer they observed the scene with an almost constant frequency of 3 to 4 fixation points per second.



(a) Example of an observer fixation points

(b) 657 points extracted by the detector Harris$_{S_2}$

(c) Example of an observer fixation points

(d) 163 points extracted by the detector Harris$_{ME}$

Figure 5: Examples of extracted key-points for an observer and the key-point detector.

To compare this visual course to the extracted key-points, we hypothesize the observation order can vary from and observer to another but a similar set of fixation point is observed after a certain duration. We need to consider the different level of perception from pre-attentive vision to semantic driven vision. Therefore, only a part of key-point is observed by humans. We expect a theoretical curve as presented in figure 6b.

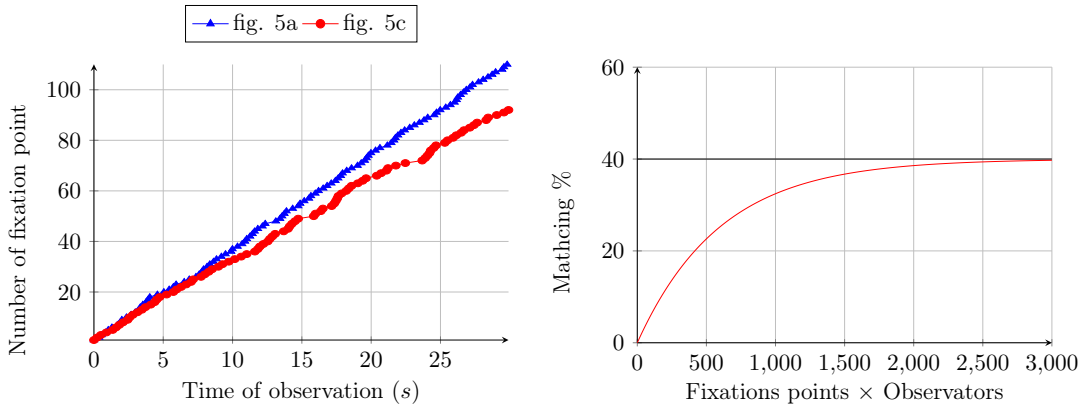
4.2 Correlation between detected key-points and fixation points

The curves from figures 7a and 7b shows the our initial hypothesis (expected results from figure 6b) seems consistent. Nevertheless, these curves does not allow to define the pre-attentive duration. As the detector Harris$_{S_2}$ is less selective, it offers higher correspondence rate between key-points and fixation points. Nonetheless, if some key-points are observed by several observers some are never observed.

The figures 7c and 7d present the ratio of observed key-points over all the observers fixation points depending on the time ordered fixation points. The curves decreases with time. The decrease vary in speed and tends toward zero as the number of fixation points increases indefinitely with time while the number of key-points is finite. It reinforces our hypothesis that mainly the

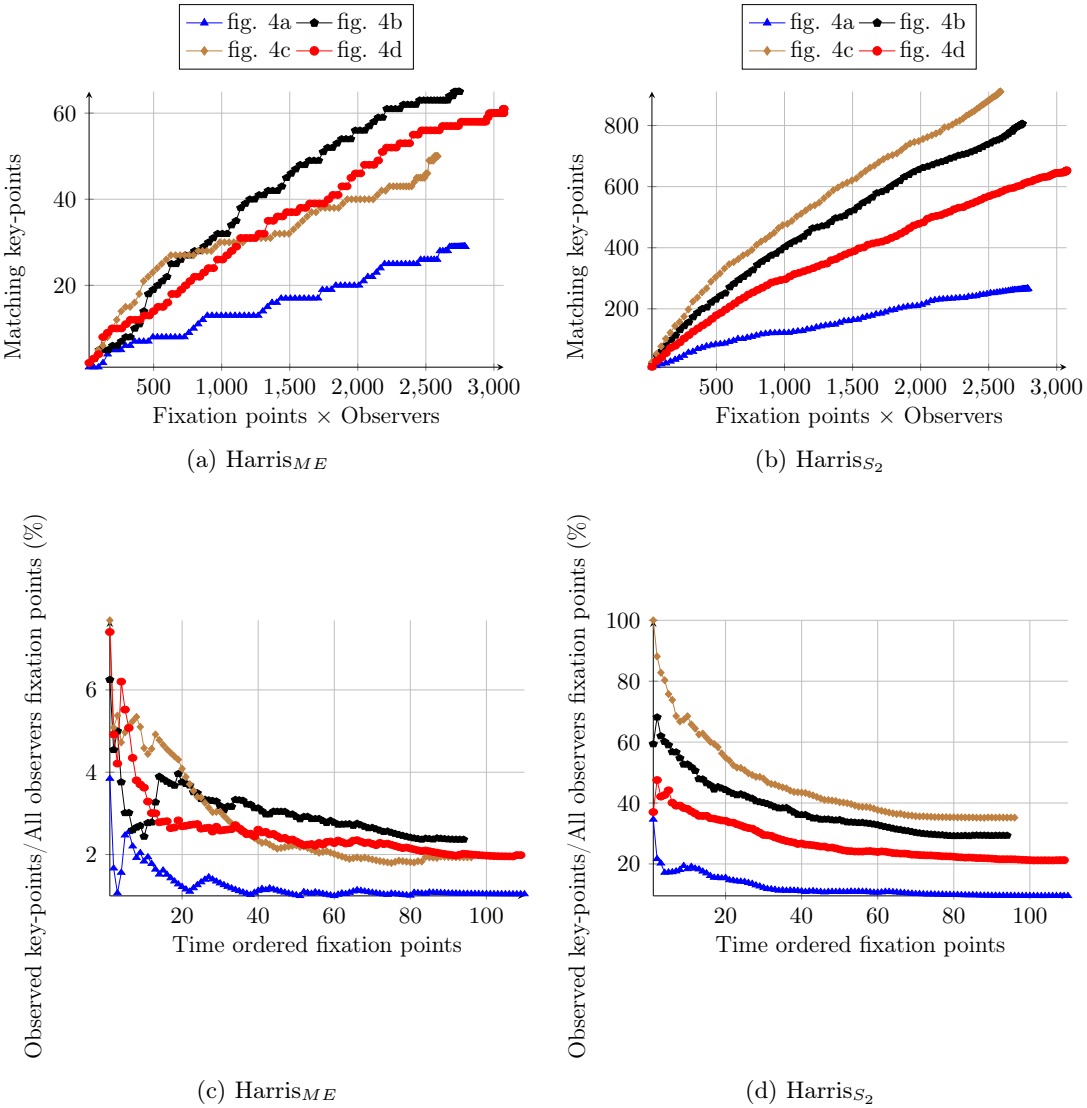
Table 1: Numbers of key-points detected by our proposition. The matching percentage is the number of key-point looked by an observer divided by the total key-point number.

					
Harris$_{ME}$	Key-point Nb	62	164	152	163
	% matching	46.7%	39.6%	32.9%	37.4%
Harris$_{S_2}$	Key-point Nb	657	1961	2276	1423
	% matching	40.5%	42.0%	40.0%	45.9%



(a) Fixation points depending on duration (b) Expected relation

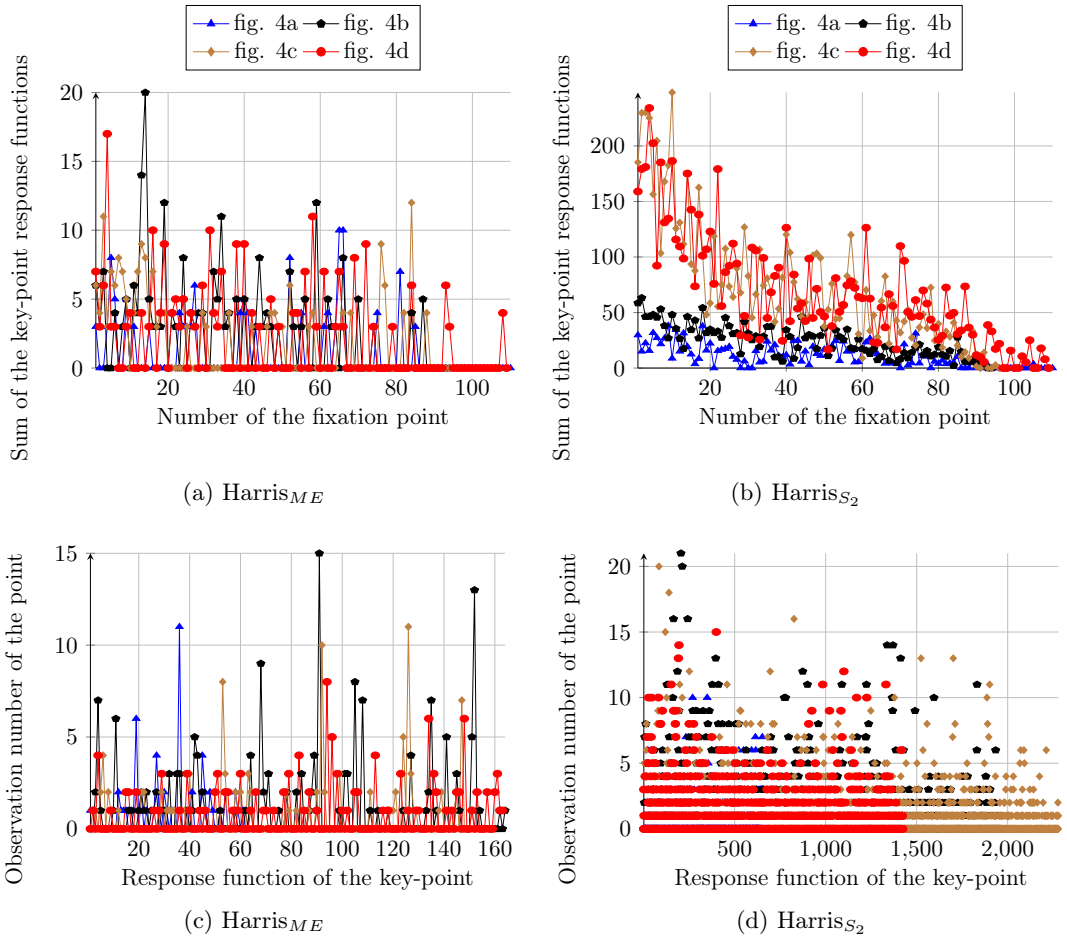
Figure 6: Preliminary results. The figure 6a presents the relation between the number of fixation point and time. The second one (fig. 6b) presents the expected relation between key-points and fixation point.



(a) Harris$_{ME}$ (b) Harris$_{S_2}$

(c) Harris$_{ME}$ (d) Harris$_{S_2}$

Figure 7: Match between key-point and saliency depending on the detector. The figures 7a and 7b present the number of key-point observed depending on the time (fixation points number × observers). Figures 7c and 7d present the number of key-point observed over the number of fixation point.



(a) Harris$_{ME}$ (b) Harris$_{S_2}$

(c) Harris$_{ME}$ (d) Harris$_{S_2}$

Figure 8: Harmony between key-point and saliency points depending on the intensity of the point. The figures 8a and 8b present the sum of the response function depending on the chronologically ordered fixation point. Figures 8c and 8d present the number of observation ordered by decreasing response function.

pre-attentive phase will match with detected key-points.

The previous figures showed the first fixation points have a higher match rate than the later ones. We can ask ourselves, if the matching key-points corresponds to the stronger ones. The figures 8a and 8b presents the sum of the response functions of the matching key-points depending on the time ordered fixation points. If a relation is difficult to extract form the detector Harris$_{ME}$, it is clear for Harris$_{S_2}$: stronger key-points are observed on the pre-attentive phase.

From another point of view, figures 8c and 8d present the number of observation of a key-point depending on the key-point ordered by decreasing response functions. For the detector Harris$_{ME}$, the number of observation is overall constant whatever the response function. On the contrary, for the detector Harris$_{S_2}$, the higher response functions key-point have a slightly higher probability of being observed by humans. We cannot extract a definite tendency with these two figures. It corroborates our initial caution when comparing key-points and fixation points, After the pre-attentive phase, the analysis is brain driven to extract the semantic meaning that was pronounced during the experiment.

5 Conclusion

Overall, several of the first fixation points correlate with high response key-points detected with our method. This comparison supports our conjecture based on strong correlation between the

first fixation points corresponding to the pre-brain analysis and the corner key-points extracted with our detector.

The comparison should be more developed to better understand the link between fixation points and key-points detection. In fine, this could guide the future key-point detector to a better harmony with the visual perception and brain analysis.

The fixation points analysed were obtained by looking the images for a long time. Moreover, the task was to identify the scene when possible which is a top-down approach. As said in the introduction key-points are low-level features not immediately related to these approaches. Comparing key-points and fixation points associated to another task such as looking for details could induce a better matching rate.

Another limit of this analysis is that the key-points are selected to be corners while fixation points can be a corner or the centre of a uniform area. Therefore, this study should be completed with a one comparing a blob detector and the fixation point. This can be a response on the battle corner/blob detection: both are of interest in regard to human vision!

References

[Bay et al.(2008)Bay, Ess, Tuytelaars, and Van Gool] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008. ISSN 10773142. doi: 10.1016/j.cviu.2007.09.014.

[Borji and Itti(2012)] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012.

[Borji et al.(2011)Borji, Sihite, and Itti] Ali Borji, Dicky N Sihite, and Laurent Itti. Computational modeling of top-down visual attention in interactive environments. In *BMVC*, volume 85, pages 1–12, 2011.

[Chatoux et al.(2019a)Chatoux, Richard, Lecellier, and Fernandez-Maloigne] H. Chatoux, N. Richard, F. Lecellier, and C. Fernandez-Maloigne. Full-vector gradient for multi-spectral or multivariate images. *IEEE Transactions on Image Processing*, 28(5):2228–2241, May 2019a. ISSN 1057-7149. doi: 10.1109/TIP.2018.2883794.

[Chatoux et al.(2019b)Chatoux, Richard, Lecellier, and Fernandez-Maloigne] Hermine Chatoux, Noël Richard, François Lecellier, and Christine Fernandez-Maloigne. Gradient in spectral and color images: from the di zenzo initial construction to a generic proposition. *JOSA A*, 36(11):C154–C165, 2019b.

[Di Zenzo(1986)] Silvano Di Zenzo. A note on the gradient of a multi-image. *Computer vision, graphics, and image processing*, 33(1):116–125, 1986.

[Harris and Stephens(1988)] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Citeseer, 1988.

[Jin et al.(2012)Jin, Liu, Xu, and Song] Lianghai Jin, Hong Liu, Xiangyang Xu, and Enmin Song. Improved direction estimation for di zenzo's multichannel image gradient operator. *Pattern Recognition*, 45(12):4300–4311, 2012.

[Koschan and Abidi(2005)] Andreas Koschan and Mongi Abidi. Detection and classification of edges in color images. *IEEE Signal Processing Magazine*, 22(1):64–73, 2005.

[Le Meur(2005)] Olivier Le Meur. *Attention sélective en visualisation d'images fixes et animées affichées sur écran : modèles et évaluation de performances - application*. PhD thesis, 2005. URL `http://www.theses.fr/2005NANT2063`.

[Lowe(1999)] David G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[Mikolajczyk and Schmid(2001)] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 525–531. IEEE, 2001.

[Moravec(1980)] Hans P Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report, STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1980.

[Rosten and Drummond(2006)] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Computer Vision ECCV 2006*, pages 430–443. Springer, 2006.

[Shrestha(2016)] Raju Shrestha. Simulating colour vision deficiency from a spectral image. *Studies in health technology and informatics*, 229:392–401, 2016.

[Smith and Brady(1997)] Stephen M. Smith and J. Michael Brady. SUSAN: a new approach to low level image processing. *International journal of computer vision*, 23(1):45–78, 1997.