Journal of Eye Movement Research
3(2):5, 1-13

A statistical mixture method to reveal bottom-up and top-down factors guiding the eye-movements

Thomas Couronné
Gipsa-Lab, INPG, Grenoble, France

Anne Guérin-Dugué
Gipsa-Lab, INPG, Grenoble

Michel Dubois
LPS, Grenoble II University

Pauline Faye
PSA Peugeot-Citroen, Paris

Christian Marendaz
LPNC, Grenoble II University

When people gaze at real scenes, their visual attention is driven both by a set of bottom-up processes coming from the signal properties of the scene and also from top-down effects such as the task, the affective state, prior knowledge, or the semantic context. The context of this study is an assessment of manufactured objects (here car cab interior). From this dedicated context, this work describes a set of methods to analyze the eye-movements during the visual scene evaluation. But these methods can be adapted to more general contexts. We define a statistical model to explain the eye fixations measured experimentally by eye-tracking even when the ratio signal/noise is bad or lacking of raw data. One of the novelties of the approach is to use complementary experimental data obtained with the "Bubbles" paradigm. The proposed model is an additive mixture of several a priori spatial density distributions of factors guiding visual attention. The "Bubbles" paradigm is adapted here to reveal the semantic density distribution which represents here the cumulative effects of the top-down factors. Then, the contribution of each factor is compared depending on the product and on the task, in order to highlight the properties of the visual attention and the cognitive activity in each situation.

Keywords: Visual attention, Eye movements, "Bubbles" Paradigm, Additive mixture model, Cognitive model, Manufactured Objects

Introduction

In the industrial context of product evaluation, it appears relevant to use non declarative methods to detect which properties of the product attract the customer's attention. Indeed, when a manufactured object is designed, it is necessary to adapt the characteristics of the object –i.e. its design- in order to match both the designer's intentions and the public's expectations.

While some researchers are using ethological methods (manipulations, interactions), sociological ones (interaction between users about the product), or sensorial (physical perception of the product), our contribution to this field is to analyze the visual attention of the customers.

The main idea of this work is to measure the spatial distribution of the visual attention when participants are gazing at a product. This should help to detect what kind of cognitive factors might direct the visual attention and to identify the properties – i.e. overall impressions or attributes- carried by the design which influence the product's assessment. In the present study we want to

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

estimate the contribution of several perceptive and cognitive factors which could potentially guide the visual attention. While the overt visual attention can be inferred from eye-movements (Anderson & Dearbom, 1952) in various situations like driving or playing chess, the links between the visual and the cognitive activities are not obvious during an assessment task. Indeed, the visual information is not systematically processed serially (but in parallel): the eye positions do not reflect the encoding of local visual information (Rayner, 1998). Moreover, in situations with complex task or stimulus, the visual attention is subject to multiple cognitive top-down factors (Yarbus, 1967; Carpenter & Just, 1983; Henderson, 2003) in interaction with bottom-up ones (Chauvin, 2003; Itti & Koch, 2001 ; Itti et al., 2005). Several authors suggest some quantitative models to link eye positions and information encoding during scene processing (Reichle, Rayner & Pollatsek, 2003). Like De Graef (De Graef, 1998), Baccino (Baccino, 2002), or Tatler (Tatler et al., 2006) who use eye-tracking techniques to observe the attention process, we want to extract information from eye-movements about cognitive and affective processes during an assessment task of pictures of car cab interiors. The previous studies that have focused on manufactured objects confirm the complexity of the relationship between the eye positions, the cognitive locus and the assessment task. Hammer (Hammer & Lengyel, 1991) shows that the eyes are directed towards the product areas that support text information. Sharmin (Sharmin, 2004) shows that during a mobile phone evaluation, visual attention may browse the scene depending on at least two strategies: neighboring exploration and holistic information analysis.

Therefore, our approach is not to propose a supplementary model completely dedicated to a precise context to explain the multiplicity of the involved processes, but thanks to this controlled context, we propose a general framework to design a statistical model linking eye movements and visual attention.

Our methodology uses a statistical additive mixture model to estimate the contribution of several a priori distributions to explain the fixation distribution, depending on the visual scene (here the manufactured product) and the task. The first step of the study is to measure the eye-movements of customers during a product evaluation task, and to aggregate them per task and product. Then, we make a hypothesis on the factors which might guide the visual attention. For each of them, a statistical model is defined to design each spatial distribution. In this study, five factors are defined. The first two are independent to the visual scene: the *random effect* and the *centrality bias*. The next two depend on the visual scene: the *visual saliency* predicting the visual attention only from low-level local image features such as colors, edges, contrasts and luminance, and the *information optimization* based on the relative position of the edges. The last one is *semantic*. The model for this last factor consists in extracting the relevant semantic information useful to solve the task. We suggest applying here the experimental "Bubbles" paradigm to evaluate this semantic information. Finally, the contributions of each factor are compared between each experimental situation (task × product).

The first part of this article presents the computational principle of the additive mixture model. We first consider an additive Gaussian model in order to illustrate this model on our experimental eye tracking data and to set the background of the proposed method. In this model, the different modes in the additive model are not necessarily Gaussian but must implement the different a priori guiding factors. Then we detail the design of each a priori mode from empirical distributions, developing the use of the "Bubbles" experimental paradigm to build the semantic distribution. The second part exposes the experimental protocol and the results.

Methodology

Additive mixture model to explain eye fixation distribution

The most common technique to create density maps is to make convolution between the fixations map and a Parzen kernel (here for example a Gaussian kernel adjusted with the fovea size). It is a non-parametric method which is not useful to extract the clustering structure of the data. So we prefer a parametric modeling using an additive mixture model. Moreover, in the case of noisy data or of lack of robust data, we complete this model with a "random mode" to extract a part of the noisy structure in the data. Therefore, in this section we present the use of an additive mixture Gaussian method to model the spatial distribution of eye-fixations, first without the "random mode", and secondly with the "random noise".

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

This approach is commonly used to estimate the spatial gaze density function. This method is "image-dependant" because its interpretation is directly linked to the objects which compose the scene. The Gaussian additive mixture model is implemented with the "Expectation-Maximisation" algorithm (Dempster, Laird & Rubin, 1977) as a statistical tool for density estimation. The density function *f(x)* of a random uni or multivariate variable *x* is estimated by an additive mixture of *K* Gaussian modes according to the following equation:

$$f(x) = \sum_{k=1}^{K} p_k . \phi(x;\theta_k) ,$$

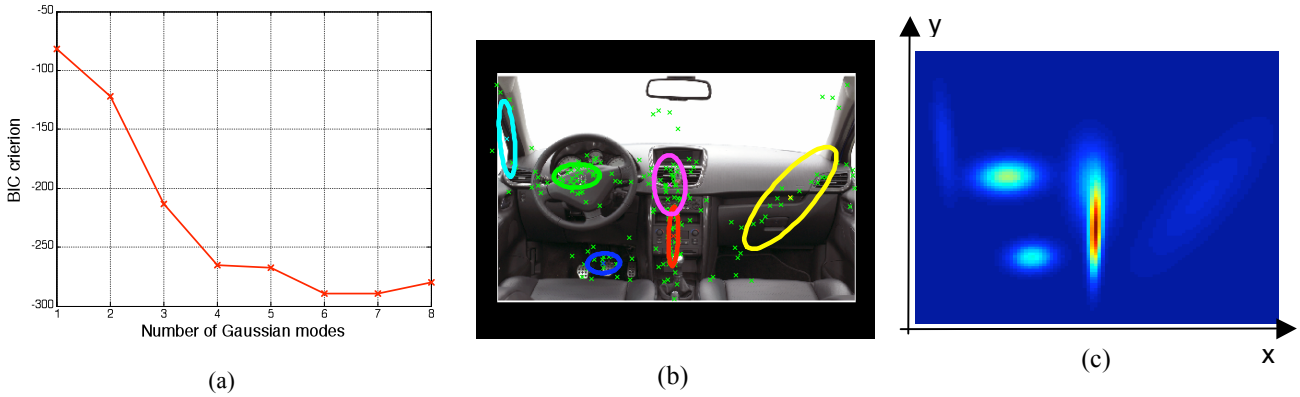
with *K* the a priori number of Gaussian modes, p_k the weight of each mode ($p_1+...p_K=1$), $\phi(x; \theta_k)$ the Gaussian density of the kth mode and θ_k its parameter (mean and covariance matrix).

The number of modes (K) is a priori unknown and must be chosen. The selection model assesses the fitting quality (higher value for K) and the robustness without overtraining (lower value for K). A classical approach is to use an information criterion which balances the likelihood of the model with its complexity (Hastie, Tibshirani & Friedman, 2001). Among the different available criteria, the Bayesian Information Criterion (BIC) (Schwarz 1978) is preferred in a density estimation context (Keribin 2000). A range of possible values of K is chosen depending on the complexity of the visual scene. For each value of K in this interval, the optimal parameters (p_k, θ_k, $k=1...K$) of the mixture are found at the convergence of the Expectation-Maximization algorithm ("EM" algorithm,). From all these sets of parameters, the "best" model is selected: it minimizes the BIC criterion:

$$BIC = -2.L + \nu.\ln(n) , \qquad (1)$$

with *L* the maximum log-likelihood of the estimated model at the "EM" convergence, ν the number of free parameters and *n*, the number of observed data.

Figure 1 illustrates this model on eye fixations dataset when people freely gaze at this visual scene during 8 seconds. The range for the candidate values for *K* is between one and eight (this value is large enough according to the number of objects in the scene)



(a) (b) (c)

Figure 1: (a) Evolution of the BIC for all the K candidate values, (b) Eye fixations during a free viewing of this scene and localization of the Gaussian modes after the model selection, (c) Spatial density f(x, y)of the eye fixations after the model selection. Six Gaussian modes are defined to best fit the experimental data.

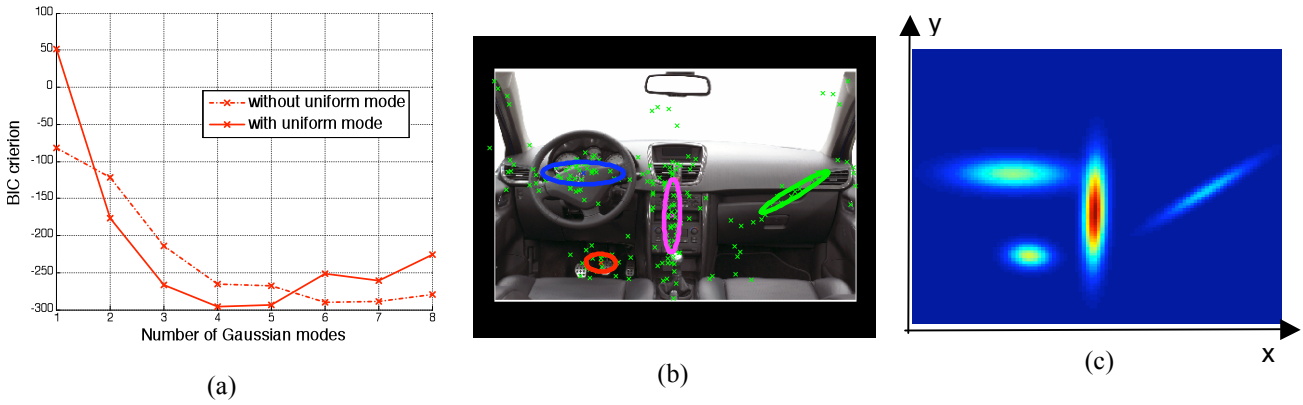
In this example the BIC criterion reaches its optimum for *K=6* when *K* varies from one to eight. The best model has thus six Gaussian modes (figure 1.a). Figure 1.b shows the localization of these modes. Each mode is illustrated by its position (mean) and its spreading at one standard deviation (ellipse). These modes describe the spatial areas which are fixated during the scene exploration, i.e. the probability for each area to be gazed at. Figure 1.c illustrates the resulting spatial density function *f(.)*. Moreover, we can notice that the configuration with seven Gaussian modes can be also a "good" solution: the values for the BIC criterion are very close between these two configurations (*BIC=-289.50* with *K=6*, *BIC=-289.12* with *K=7*). After that the BIC value increases (*BIC = -280.02*).

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

This example illustrates also one common difficulty when using the classical "EM" algorithm faced with noisy data. If the latent data clustering is not very strong, some Gaussian modes can be extracted or not depending on the random initial conditions. That it is the case with the vertical Gaussian mode close to the left side of the image; its extraction depends on the initial conditions. To cope with such situations, we add a supplementary mode in the model: a uniform density. The experimental observations which are not close to latent clusters get contributions to this mode: it is the "noise" mode. The equation of the complete model is then:

$$f(x) = \sum_{k=1}^{K} p_k.\phi(x;\theta_k) + p_u.U(x)\,, \qquad (3)$$

with K the a priori number of Gaussian modes, p_k the weight of each Gaussian mode, p_u the weight of the uniform mode ($p_1 + \ldots p_K + p_u = 1$), $\phi(x; \theta_k)$ the Gaussian density of the k^{th} mode and $U(x)$ the uniform constant density such as $\int_{x \in D} U(x).dx = 1$. The "EM" algorithm is adapted to this model in order to estimate also the contribution p_u. So the model selection concerns the two previous ones, with or without the uniform mode. For the same data, the results are illustrated at figure 2. The minimum value of the BIC criterion is -295.60. This value is reached for the model with four Gaussian modes and the uniform mode.



(a) (b) (c)

Figure 2: (a) Evolution of the BIC for all the K candidate values for the two models, with or without the uniform mode (b) Eye fixations during a free viewing of this scene and localization of the Gaussian modes, after the model selection, (c) Spatial density f(x, y) of the eye fixations, after the model selection. Four Gaussian modes are then defined to best fit the experimental data. The contribution of the uniform mode is not visible on the figure because its density is constant.

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

The contributions of each mode are presented in table 1. We notice the contribution p_u is significant compared to the contributions for the Gaussian modes. This uniform mode explains scattered data. Here there are scattered eye fixations which are not localized on specific areas (around here 17% of the whole fixations), i.e "ambient" eye fixations which are very sensitive to inter-individual variability.

Mode	p_1	p_2	p_3	p_4	p_u
Contribution	0,10	0,08	0,28	0,37	0,17

Table 1: Contribution of each mode. The four first modes are Gaussian, the last one is uniform (see Fig. 2 b,c)

This approach, combining Gaussian mixture and the "EM" algorithm, is common to estimate density functions. Depending on their position and deviation, the modes can be interpreted according to the objects in the scene. Here, the three modes in the left side among the four modes in this scene are related to objects: the steering wheel, the pedals, and the central desk. These objects are very important for the interpretation of the scene. But this approach does not reveal for a given task, if some modes are induced by a similar factor, or if a factor has similar effects on visual scenes which have not similar semantic properties.

Nevertheless, we keep this statistical model as a general framework to set-up the new model in which the modes are not necessarily Gaussian. They must represent the candidate guiding factors across different visual scenes for the same task and not spatial concentrations of eye fixations for a given visual scene.

According to the previous section, the mixture model can be set depending on the density properties of the experimental data. But it can also be done by a priori hypotheses defining the number and the properties of each mode of the mixture. Then the global contribution of the whole fixations to each mode is estimated. This approach is employed by Vincent et al. (Vincent et al., 2009) where the eye positions density is modeled with a mixture of elementary a priori defined densities, each density representing a specific factor which might guide the visual attention. Thus, each mode of the additive mixture is defined by a density modeling one candidate factor which might drive the cognitive analysis of the visual task. The common properties of these two models are the additive mixture of the modes and the "EM" algorithm to define their configuration parameters.

These factors describe both low level and high level processes. Each mode is used to assess the contribution of the associated factor. First, it is necessary to identify these candidate factors, in relation to the visual task and then, their statistical density model. Each of these densities is represented by a spatial density map, either from a specific image processing, or from a manual segmentation and or also from statistical hypotheses, depending on the nature of the attention factors. The "EM" algorithm provides stable results if the a priori distributions are not strongly correlated. Each distribution which codes a guiding factor must provide complementary effect on the studied process, the visual attention. At the convergence of the "EM" algorithm, the contributions of each density are estimated, maximizing the likelihood of the final model which is then completely defined.

To summarize, a noticeable characteristic of this method is that the additive model contains a priori density maps, which are chosen depending on the stimuli, the tasks, and the assumptions to be investigated, and which must be previously characterized.

Setting up the a priori distributions composing the model

Five factors are suggested to explain the observed fixation distribution, each one being modeled by one spatial distribution and being considered as one mode of the additive mixture.

First of all, if the eyes are guided by a random process, the distribution will follow a uniform law: each area of the space has the same probability of being gazed at. In the mixture, this map acts as a "trash" map, capturing fixations which are not explained by other assumptions.

The second factor is a process of central gazing (Tatler, 2007): the "on screen" gazing produces a central bias: the eyes preferentially gaze at the center of the screen and tend to return to it regularly, regardless of the content of the image. This is also the initial gaze position and may also be a rest position. A "centrality map" is thus defined, where the central area has a higher probability to be fixed than peripheral ones. The density is determined by a Gaussian function applied to the center of the image. In the original model proposed by Vincent, the mean and the variance covariance matrix will be adjusted during the algorithm as in a usual "EM" algorithm. Here these parameters are fixed because we want to evaluate the contribution of this factor in the central area (see Figure 3).

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Indeed, if these parameters are set after learning from the "EM" algorithm, this spatial mode can move or not in another place depending on the visual scene.

The third factor comes from the visual bottom-up saliency. One of the basic principles is that the eyes are attracted towards areas of high contrasts combining different low level visual features on textural luminance and chrominance variances. The bottom-up saliency model proposed by Itti (Itti & Koch, 2001) is very popular in this domain. In this work, we use a similar algorithm which is developed in our laboratory. It is based on the same general principles as Itti's, but using a more accurate model at the retina level (Ho, Guyader & Guérin-Dugué, 2009). This map is considered here as merging low-level visual information to predict the relative attractiveness of spatial areas without "top-down" attention factors (see Figure 4).

The fourth factor is based on an information maximization approach (see Figure 5): it comes from experimental observations (Renninger, Coughlan & Vergheese 2005) that the eyes can be spatially positioned in such way to optimize the acquisition of visual information instead of sweep over an area to encode each information. This approach needs to compute from which point of view the details' perception is maximized, limiting the number of saccadic movements. The edges are then extracted by an edge detector (here the "Sobel" detector).Then a spatial clustering algorithm is used to find the position of the edges barycenters. We use the "Mean Shift" algorithm (Fukunaga & Hostetler, 1975; Cheng, 1995). Finally, a Gaussian function, set up with the fovea size, is defined for each cluster and centered on each barycenter. For this factor, the highest probability of gazing at an area is located at the barycenters of the edges and their neighboring areas.

The last factor is a semantic top-down factor: the visual attention is driven by cognitive processes and therefore by the semantic content of the visual scene. Thus, we have to extract the local areas of the scene which contain relevant information regarding the task. To model this factor, we use complementary experiments with the "Bubbles" paradigm using the same visual scenes and the same tasks. Our hypothesis is: the resulting "classification map" obtained by this paradigm gathers the top-down effects in the observed eye fixations density. The constructions of these maps are detailed in the next section (see Figures 6 & 7).

After having estimated the spatial density distribution of each factor and measured the eye-movements, the "EM" algorithm is employed in combination with the additive mixture model to find the best parameters for the model. In our case, these parameters are the relative contribution of each mode (each factor) to the eye fixations density. In the next section, we provide the details of the method to build the semantic map, based on the "Bubbles" paradigm.

Semantic map construction from the "Bubbles" paradigm

Material & Methods

For the top-down factor, we want to extract the relevant parts of the visual scene which should be observed to solve the given task. As far as we know, the "Bubbles" paradigm (Gosselin & Schyns, 2001) has never been employed to compare semantic information and bottom-up effects on visual attention. We apply it here to select, for a given couple image × task, the visual areas encoding relevant information to solve the judgment task. This paradigm was originally designed to identify facial areas associated with facial expressions recognition (Humphreys et al., 2006). It consists in watching the scene through a mask, with only a few parts being maintained visible (the "Bubbles", which are spatially set at random).

Therefore the participants must solve a decision task while gazing the scene through the "Bubbles" mask. This method is relevant when the task resolution requires the capture of local visual information in the scene. The fixation density distribution shows areas which attract the visual attention and the "Bubbles" method identifies the decision areas. Moreover it shows whether the decision is really about local areas or not, and if the judgment task is homogenous or not between participants.

To summarize, this method is efficient when there is a "ground truth" which is related to the correct answers, when the decision is based on local visual areas, and when the decision criteria are stable.

Otherwise, if several "correct answers" exist, if the different visual scenes cannot be discriminated by local areas, or if the decision is subject-dependant, then the algorithm will not be able to extract local areas statistically associated with consensual decisions.

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692



Figure 3: "Centrality map" on the 207 car image.

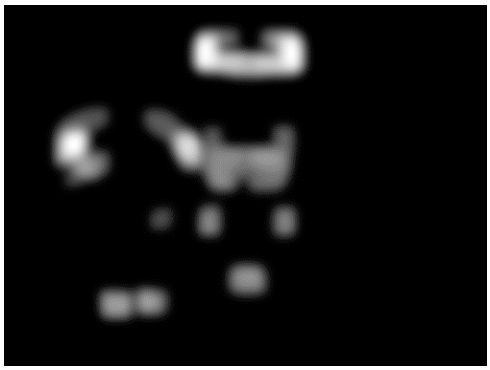
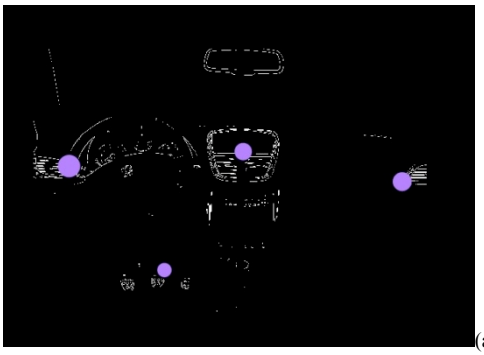
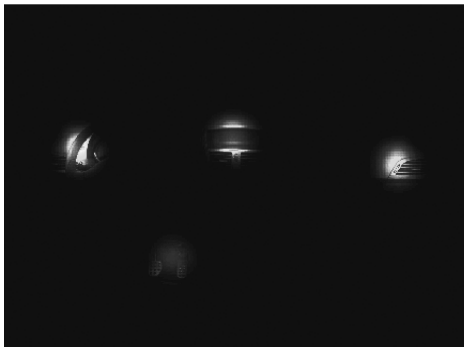


Figure 4: Saliency map for 207-sport car image.



(a)

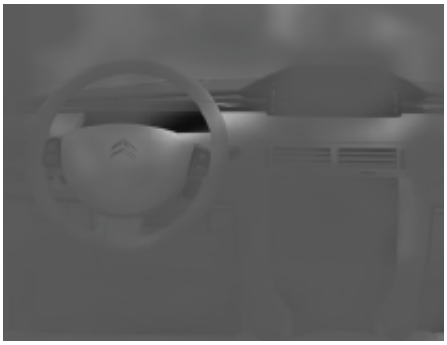
Figure 5: Set up the « Maximization » map for the 207 image from (a) a Edge filter and clustering -here 4 clusters-



(b) to obtain the final map after a Gaussian smoothing.



Figure 6: (a) Classification image for 207, assessment of the sport character of the 207.



(b) Classification image for C6, assessment of the luxury property.

DOI 10.16910/jemr.3.2.5

This article is licensed under a
Creative Commons Attribution 4.0 International license.

ISSN 1995-8692

The "Bubbles" paradigm proposed by Gosselin (Gosselin & Schyns, 2001) statistically links the subjects' answers with the areas of the visual scene gazed during a decision task. These local areas are called the "diagnostic areas".

The stimuli we employ are visually and semantically complex, and the decision activates high-level processes. Moreover, a consensus between participants is necessary in order to extract some stable diagnostic areas: one "right" answer and a "false" one must exist, and this alternative must be homogenous to all the participants. Therefore, we adapt this paradigm to a paired-wise comparison task, in order to assess the stimuli with a reference (Humphreys et al., 2006). The "Bubbles" are set at random for one image of the pair and the same localizations are set for the second image. See Figure 7. The decision is taken after the visual inspection of both images of the pair, having a similar masking (left and right sides of the screen): in the pair, one image has the required property, the second, does not. The description of the different properties studied of the car's cab interiors are described in the next section.

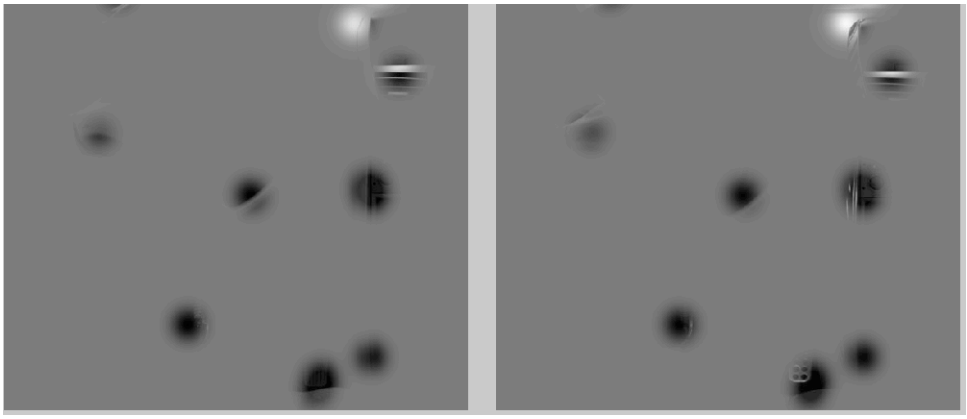


Figure 7: A "Bubbles" paired-wise comparison for 207 car cabs.

The algorithm adjusts automatically the number of "Bubbles", to be adapted to the performance of the participants (from 70 up to 80% of correct answers) during the trial and will move towards a setup threshold of correct answers. The surface of each bubble is set such as its radius is one angular degree (fovea). On average (depending on the complexity of the scene), between 10% and 15% of the picture is visible. Finally the algorithm estimates the correlations between the correct answers and the position of the visible areas, and provides a probabilities distribution. This is the probability for a spatial area to be associated to a right answer. In other words, the "Bubbles" paradigm provides a spatial map of right decision making: the classification map.

The experiment was designed with the Stat4Ci[1] Matlab Toolbox provided by F. Gosselin. The classification task is carried out with 10 participants per condition and 320 decisions per subjects. At least 900 tries are performed on each pair and each task. For each try, the pictures are partially masked by the bubbles. Consequently, the participant has to make a decision on partial information. If the answer is false, we can consider that the visible information through the "bubbles" isn't related to the assessment task. Otherwise, if the answer is right, the visual information is sufficient: the visible parts through the "Bubbles" are significant regarding to the task. By cumulating the answers of all the participants, the correct answers are statistically correlated with the visible areas.

The participants are 64% male and 36% female; average age is 31.4 years old. They are not working in the automotive sector (design, marketing or communication).

Two car cab interiors are chosen to be judged by participants: Peugeot 207 (207) and Citroën C6 (C6), designed in two versions: sports versus standard for 207, white versus black for C6. We therefore obtain four visual stimuli.

1 http://www.mapageweb.umontreal.ca/gosselif/labogo/Stat4Ci.html

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Two tasks are chosen: for the 207, the participants assess the "sport" car's design or the "quality level" of the interior cabs. For C6, instructions are to evaluate the "quality level" or the" luxury quality" of the interior cabs.

Results

It must be noted that the quality of the results depends on the experimental conditions and the number of trials. Figure 6.a presents the classification results for the 207 (sport character), and this map actually represents decision areas. Moreover, for the task on the "high level" assessment, the decision areas appear less locally accurate than for the task on "sport" type.

For C6, the decision areas are not spatially localized. Actually, for the "quality level" assessment, the choice is not homogeneous among the subjects: the "Bubbles" algorithm cannot extract converging areas. For the "luxury" assessment task (see Figure 6b), the choice effectively points to the white modality ("correct answer"), but the only decision criterion is the color.

Oculometry measures: assessment of the factors driving the eyes

Material & method

We have defined in the preceding section eight experimental conditions (4 pictures * 2 assessments). In this experiment, we add a control task: four groups observe each stimulus without instructions (4 conditions).

Twelve groups are thus formed, having ten participants per group, and for the control (free viewing task), four groups are formed, one per interior compartment. The participants are selected in such way to have homogeneous groups in terms of age and gender; they all bought a medium-range car in the last two years; they live in France and do not work in the automotive sector. The participants' ages are homogeneously distributed from 20 to 60 years old (median: 35 y.o.).

The cars pictures are exposed at scale 0.80 of their real size on a screen of 160cm × 125cm at 2m from the participant (42° visual angle). The eye-tracker employed is the FACELAB® 4.1, used in precision mode (1.5°) with a sampling frequency of 60Hz. Two warm up tries are done before the first judgment trial. Each participant realizes two assessment tasks, one on a 207 and one on a C6. This order (207 or C6 first) is counterbalanced. After a calibration step, the sequence begins with a black slide, and the participant is asked to stare at a dot located at the center of the screen. Then, the image is exposed for 8 seconds. Finally the participant gives his answer, and the sequence is repeated.

Twenty eye-movements sequences are recorded per product and task. After eliminating wrong measures or failed trials (around 20%), we obtain around sixty sequences of 8 seconds per condition.

Results

In order to explain the spatial fixations density for each experimental condition, we have then defined the statistical model based on an additive mixture of the five distributions previously described which might compete to guide the visual attention in this context. At the convergence of the "EM" algorithm, we obtain the contribution of each of these five factors. These contributions are considered as the relative effect of each factor guiding the eye-movements (the sum of the contributions value is equal to one). Table 2 shows these contributions for each experimental condition (stimulus × task). Table 3 shows the results for the free viewing task.

0bject	Task	Random	Centrality	Saliency	Info-Max	Semantic
207-sport	sport assessment	0.00	0.41	0.00	0.15	0.44
207-sport	quality	0.00	0.50	0.00	0.15	0.34
207-standard	sport assessment	0.00	0.24	0.00	0.24	0.53
207-standard	quality	0.00	0.63	0.00	0.19	0.18
C6-white	luxury	0.00	0.61	0.12	0.19	0.08
C6-white	quality	0.00	0.56	0.06	0.26	0.12
C6-black	luxury	0.00	0.52	0.12	0.31	0.05
C6-black	quality	0.00	0.73	0.00	0.22	0.05

Table 2: Contribution of each factor for each experimental condition (judgment tasks).

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Object	Random	Centrality	Saliency	Infomax	Semantic
207-standard	0.00	0.62	0.00	0.15	0.18
207-sport	0.00	0.45	0.00	0.24	0.31
C6-white	0.00	0.67	0.04	0.19	0.08
C6-black	0.00	0.77	0.00	0.31	0.04

Table 3: Contribution of each factor for each experimental condition (free viewing task).

First of all, the "Random" factor does not contribute to the model: the fixations are not randomly distributed and they can be explained by the other factors. Secondly, the distinction between 207 and C6 conditions is highlighted. For 207, the semantic map explains the eye-movements better than the low-level maps ("Saliency" and "InfoMax") which have weak contributions. For C6, the situation is reversed; the low-level maps explain the experimental data better than the high-level semantic map. Moreover, the centrality bias is stronger for C6.

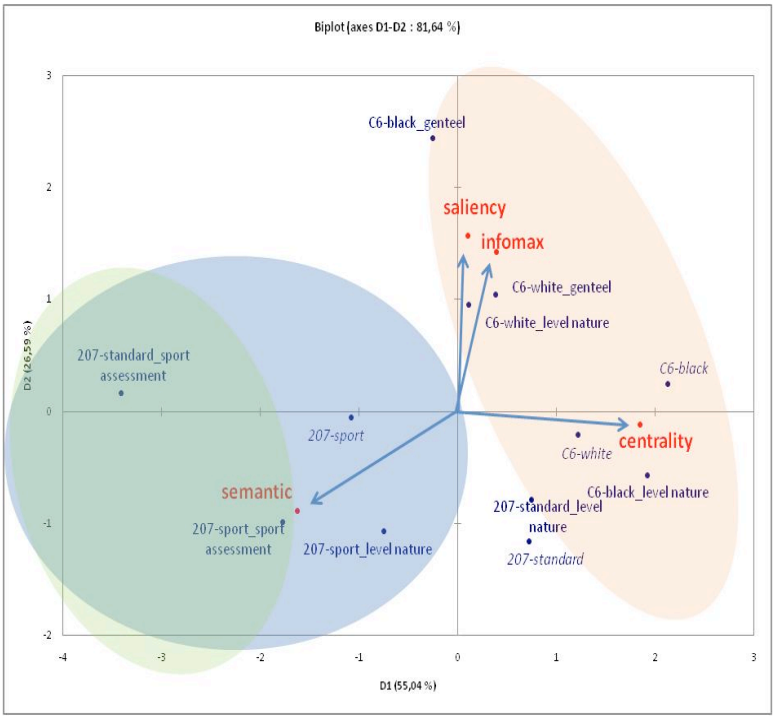


Figure 8: Projection of each experimental condition on the first two principal components (81.64% of the observed inertia).

A Principal Components Analysis appears here as a very useful way to compare the various eye movement sets in order to highlight the similarities and differences between the different models. For this, a dataset is created with the eight experimental conditions for the assessment task, merged with the four experimental conditions for the free viewing task. This provides twelve situations described by the four factors with a non null contribution corresponding to three degrees of freedom. There is one constraint: the contribution sum is set to one. The resulting biplot of the projection on the first two principal components (see Figure 8) shows mainly two

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

trends: the C6 products versus the 207 products, depending on the semantic factor contribution. For the C6 products, the visual attention seems to be guided by bottom-up factors (visual saliency, maximization). For the 207 products, two subgroups are discriminated by the contribution of the "centrality" factor: the assessment of the sporting property induces attention more focused on the decision areas than the other tasks (highest cognitive level).

Therefore, the eye positions depend of course on the task and may in some cases overlap with the visual decision criteria. When the central areas are strongly gazed at, it might be because the attention is less directed to local areas, or because the central local information is watched.

But, when participants gaze freely at the 207 products (without a judgment task), the fixation distribution is well explained by the semantic map (which is built by the "Bubbles" experiment on the basis of judgment process). Thus, even without instructions, the decision areas are particularly observed for the 207. The interpretation of the link between visual attention and decision is therefore complex. For the C6, we notice that the free viewing attention is well explained by the centrality hypothesis, but not by saliency or maximization, as might have been foreseen.

Discussion

This approach seems very efficient to highlight the relative weights of different factors which can guide the distribution of the visual attention. This method allows comparison to several experimental conditions, and to identify specific features of the attention processes for each situation.

About the results, several points must be discussed. First, the respective effects of the factors are quite similar for the 207 (either sport or standard) with and without instruction: it does not mean that the participants gaze these areas because they have information useful for the task or if there are some physical or cognitive distinctiveness areas. Even if the task requires local visual information, these areas might be gazed at because of the judgment processes, but also because of attractiveness, visual complexity, or object identification processes. Second, if the random effect appears null, the centrality one is rather highly weighted. It is convergent with the Tatlers' observations (Tatler, 2007), showing that either in a free viewing task or in a search task, the eyes tend to stand in the center. Our results therefore confirm that even in a judgment task, this bias is strong, but with our experimental protocol, we can't decide if the bias comes from the initial position of the eyes or because it is an optimal location to gaze at the scene. Third, the saliency appears to be null for the 207 cars while the semantic ones is high, and in all the cases the saliency weights are lower than the information-maximization hypothesis. The saliency map is considered to model the locations where the overt attention goes, but it does not seem to model the areas where the participants often gaze at. We can suggest that in such kind of judgment task, the saliency effects are counterbalanced, not by the task (cf. free viewing results) but by the knowledge of the object, and therefore by the semantic content of the scene. While the Information-maximization seems to contribute well, it can be explained by the fact that this map models the optimal locations to gaze at the objects of the scene, which is linked with the a priori knowledge of the spatial structure of the objects.

About the materials employed, we use pictures of car cabin interiors; therefore some information is missing between the real object and its representation (three dimensional depth, ecological immersion…). Moreover, these objects are very well known (pre-existing cognitive structures) and are consistent (the precepts are organized in a coherent manner), the representation is dense (a lot of objects can be gazed at simultaneously), and they may be interpreted at multiple levels, from colors and textures, to the presence / lack of functionalities, the relative position of the object, the overall attractiveness,...These observations converge to confirm our methodology in various experimental contexts, i.e. to measure the attention processes in real scenes, and to test other tasks and several kinds of objects.

The analysis of eye-movement data is one of the most striking points of such kind of behavioral experiments. If the experimenters have some hypothesis about the areas which will be gazed at, the definition of areas of interest independent to the measures is interesting. The transition occurrences between areas, the temporal patterns of fixation, or the characteristics of the eye-movements per area (fixation frequencies, delays before the first fixation, duration, saccade amplitude) can be studied. If there is no hypothesis about the areas which will be observed, or if

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

we need to compare the similarities between several tries, the analysis using the density estimation is relevant. After having established the density of the fixations, they can be either compared directly, or being employed to compare several tries regarding to a specific density reference (Tatler et al., 2006) data employed to estimate the density will have an effect on the number of local maxima, their highest values, and their topology.

Regarding the statistical method proposed by Vincent, we slightly modify it. First we use one generic map instead of several dedicated maps to model the bottom-up factors on the basis of a visual saliency process (chrominance, edges, and luminance) regardless of the task. Secondly, we add semantic information based on the "Bubbles" paradigm (regardless of physical characteristics of the scene) which seems well adapted to build a top-down map, to be defined for each task and each picture. This is our main contribution to this method. At last, the relevance of the additive mixture model comes from the assumption that each driven factor is complementary to the others. As a consequence of these properties, the weaknesses of this method are the following: First, the fusion model of the different factors is simple (additive mixture). If there are complex interactions, they are not taken into account. And then, if the factors are strongly correlated, the "EM" algorithm will be unstable. A second limitation comes from the "Bubbles" method: if there is a weak consensus among participants, or if the decision areas are not local, some common diagnostic areas do not appear. Finally, concerning the global model, we can note that the factors estimations are relatively stable across different trials and initial conditions. Nevertheless, the confidence estimates might have been computed using a bootstrap resampling to confirm this robustness.

In this context we have in one hand, a complex process of visual attention depending on factors which interact with each other, and in the other hand, the experimental data reflects the great variability of the subjects' behavior. A statistical approach performed in a sufficient number of tries is relevant. Even if the assumptions of the statistical model are simple as it is the case for the additive mixture, this approach remains relevant while the aim is to capture the very main effects. So the additive mixture model appears especially well adapted to such kinds of paradigms.

References

Anderson, H. I., & Dearbom, W.F. (1952). *The Psychology of Teaching Reading*, Ronald Press Co.

Baccino,T. (2002). Oculométrie Cognitive. in Tiberghien G. (Ed.), *Dictionnaire des Sciences Cognitives*, Paris :Armand Colin, 100-101.

Carpenter, P. A., & Just, M. A. (1983). What your eyes do while your mind is reading. In Rayner K. (Ed.), *Eye movements in reading: Perceptual and language processes*, New York: Academic Press, 275-307.

Chauvin, A. (2003). Perception des scènes naturelles: Etude et simulation du rôle de l'amplitude, de la phase et de la saillance dans la catégorisation et l'exploration des scènes naturelles. *Phd Thesis*, University Pierre Mendes-France, Grenoble, http://www.lis.inpg.fr/stages_dea_theses/theses/these/Th_Chauvin.html

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(8), 790-799.

De Graef, P. (1998). Prefixational object perception in scenes: objects popping out of schemas. In Underwood G. (Ed.), *Eye guidance in reading and scene perception*, Oxford, UK : Elsevier, 315-338.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society*, series B, 39(1), 1-38.

Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1), 32–40.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, 41(17), 2261-2271.

Hammer, N., & Lengyel, S. (1991). Identifying Semantic Markers in Design Products: The Use of eye movement recordings in industrial design. In Schmid, R., & Zambarbierei, D. (Eds.), *Oculomotor control and cognitive processes*, Elsevier Science, Amsterdam, 445-455.

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences*, 7(11), 498-504.

Ho-Phuoc, T., Guérin-Dugué, A., & Guyader, N. (2009). A computational saliency model integrating saccade programming, *Int. Conf. of Bio Inspired Systems and Signal Processing*, BIOSIGNALS2009, jan. 2009, Porto, Portugal.

Humphreys, K., Gosselin, F., Schyns, P. G., & Johnson, M. H. (2006). Using "Bubbles" with babies: A new technique for investigating the informational basis of infant perception. *Infant Behavior and Development*, 29(3), 471-475.

Itti, L., & Koch, C. (2001). Computational Modelling of Visual Attention. *Nature Reviews Neuroscience*, 2(3), 194-203.

Itti., L., Rees, G., & Tsotsos, J. K. (Eds) (2005). *Neurobiology of Attention*. Elsevier Academic Press, 1-744.

Keribin, C. (2000). Consistent estimation of the order of mixture. *The Indian Journal of Statistics (Sankhya)*, Series A, 62, 49-66.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372-422.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26(4), 445-526.

Renninger, L. W., Coughlan, J., & Vergheese, P. (2005). *An information maximization model of eye movements*. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), Advances in neural information processing systems (Vol. 17). Cambridge, MA: MIT Press.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2), 461-464.

Sharmin, S. (2004). Studies of Human Perception on Design Products. *Technical report*, University of Tampeere: http://www.cs.uta.fi/research/thesis/masters/Sharmin_Selina.pdf

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, 46(12), 1857-1862.

Tatler, B. W. (2007). 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):4, 1-17.

Vincent, B.T., Baddeley, R.J., Correani, A., Troscianko, T., & Leonards, U. (2009). Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Visual Cognition*, 17(6-7): 856-879.

Yarbus, A.L. (1967). Eye movements during perception of complex objects. in Riggs L.A. (Ed.), *Eye Movements and Vision*, Plenum Press, New York, chapter VII, 171-196.

DOI 10.16910/jemr.3.2.5
This article is licensed under a
Creative Commons Attribution 4.0 International license.
ISSN 1995-8692