# Accurate statistical tests for smooth classification images

**Alan Chauvin** Département de Psychologie, Université de Montréal, Montréal, QC, Canada

**Keith J. Worsley** Department of Mathematics and Statistics, McGill University, QC, Canada

**Philippe G. Schyns** Department of Psychology, University of Glasgow, Glasgow, UK

**Martin Arguin** Département de Psychologie, Université de Montréal, Montréal, QC, Canada

**Frédéric Gosselin** Département de Psychologie, Université de Montréal, Montréal, QC, Canada

Despite an obvious demand for a variety of statistical tests adapted to classification images, few have been proposed. We argue that two statistical tests based on *random field theory* satisfy this need for smooth classification images. We illustrate these tests on classification images representative of the literature from Gosselin and Schyns (2001) and from Sekuler, Gaspar, Gold and Bennett (2004). The necessary computations are performed using the Stat4Ci Matlab toolbox.

Keywords: Classification images, reverse correlation, *Bubbles*, random field theory

## Introduction

In recent years, vision research has witnessed a tremendous growth of interest for regression techniques capable of revealing the use of information (e.g., Ahumada, 1996; Eckstein & Ahumada, 2002; Gosselin & Schyns 2004b). Reverse correlation, one such technique, has been employed in a number of domains ranging from electroretinograms (Sutter & Tran, 1992), visual simple response time (Simpson, Braun, Bargen, & Newman, 2002), single pulse detection (Thomas & Knoblauch, 1998), vernier acuity (Beard & Ahumada, 1998; Barth, Beard & Ahumada, 1999), objects discrimination (Olman & Kersten, 2004), stereopsis (Neri, Parker & Blakemore, 1999; Gosselin, Bacon & Mamassian, 2004), letter discrimination (Watson & Rosenholtz, 1997; Watson, 1998; Gosselin & Schyns 2003), single neuron's receptive field (e.g., Ringach & Shapley, 2004; Marmarelis & Naka, 1972; Ohzawa, DeAngelis, & Freeman, 1990), modal and amodal completion (Gold, Murray, Bennett, & Sekuler, 2000), face representations (Mangini & Biederman, 2004; Gold, Sekuler, & Bennett, 2004; Sekuler, Gaspar, Gold & Bennett, 2004; Konsevich & Tyler, 2004) to temporal processing (Neri & Heeger, 2002). *Bubbles*, a related technique (Gosselin & Schyns 2001; Gosselin & Schyns, 2002; Gosselin & Schyns 2004b; Murray & Gold, 2004), has revealed the use of information for the categorization of face identity, expression, and gender (Adolphs, Gosselin, Buchanan, Tranel, Schyns & Damasio, 2005; Gosselin & Schyns 2001; Schyns, Bonnar & Gosselin, 2002; Smith, Cottrell, Gosselin & Schyns, 2005; Vinette, Gosselin & Schyns, 2004), for the categorization of natural scenes (McCotter, Sowden, Gosselin & Schyns, in press), for the perception of an ambiguous figure (Bonnar, Gosselin & Schyns, 2002) and for the interpretation of EEG signals (Schyns, Jentzsch, Johnson, Sweinberger & Gosselin, 2003; Smith, Gosselin & Schyns, 2004).

Both the *Bubbles* and the reverse correlation techniques produce large volumes of regression coefficients that have to be tested individually. As we will shortly discuss, this raises the issue of false positives: the risk of accepting an event that occurred by chance. Surprisingly, few classification image researchers have taken this into account (for exceptions, see Abbey & Eckstein, 2002; Kontsevich & Tyler, 2004; Mangini & Biederman, 2004). Here, we argue that two statistical tests based on *random field theory* satisfy this need for smooth classification images. The core ideas of random field theory are presented. In particular, the main equations for the tests are given. Finally, the usage of a Matlab toolbox implementing the tests is illustrated on two representative sets of classification images from Gosselin & Schyns, (2001) and Sekuler, Gaspar, Gold & Bennett (2004). But first, in order to identify the critical properties of the proposed statistical tests, we shall discuss some limitations of the two statistical tests that have already been applied to classification images.

## Multiple comparisons

In a typical classification image experiment, an observer has to classify objects partially revealed by additive (reverse correlation) or multiplicative (*Bubbles*) noise fields. The calculation of the classification image amounts quite simply to summing all the noise fields weighted by the observer's responses (Ahumada, 2002; Murray, Bennett, & Sekuler, 2002). By doing this, the researcher is actually performing a multiple regression on the observer's responses and the noise fields (see Appendix: The construction of a classification image for details). A statistical test compares this data against the distribution of a random process with similar characteristics. Classification images can thus be viewed, under the null hypothesis, as expressions of a random N-dimensional process (i.e., a random field). The alternate hypothesis is that a signal – known or unknown – is hidden in the regression coefficients.

So far, researchers have used two statistical tests to achieve this end: the Bonferroni correction and Abbey and Eckstein's (2002) Hotelling test. We will argue that these tests are not adapted to some classification images. The former is too conservative when the elements of the classification images are locally correlated; and the latter is not suitable in the absence of *a priori* expectations about the shape of the signal hidden in the classification images

## Bonferroni correction

Consider a one-regression-coefficient Z-transformed classification image (see Appendix: The construction of a classification image and Re-analyzing representative classification images). If this Z-score exceeds a threshold determined by a specified *p*-value, this regression coefficient differs significantly from the null hypothesis. For example, a *p*-value of .05, means that, if we reject the null hypothesis, that is, if the Z-score exceeds a threshold $t_Z$ = 1.64, the probability of a false alarm (or type-I error) is .05. Now consider a classification image comprising 100 regression coefficients: the expected number of false alarms is 100 * 0.05 = 5. With multiple Z-tests, such as in the previous example, the overall *p*-value can be set conservatively using the *Bonferroni correction*: $p_{BON} = p(Z > t_{BON}) * N$, with *N* the number of points in the classification image. Again, consider our hypothetical one-hundred-point classification image. The corrected threshold, $t_{BON}$, associated with $p_{BON}$ = .05, is 3.29. Such high Z-scores are seldom observed in classification images derived from empirical data. In a classification image of 65,536 data points (typical of those found in the literature, like the 256 x 256 classification images from Gosselin & Schyns, 2001, re-analyzed in the last section of this article) $t_{BON}$ becomes a formidable 4.81! For classification images of low (or reduced) dimensionality such as those of Mangini and Biederman (2004) or Kontsevich and Tyler (2004), the Bonferroni correction prescribes thresholds that can be (and have been) attained.

## *A priori* expectations

Two possibilities should be considered: either these classification images really do not contain anything statistically significant (which seems unlikely given the robustness of the results obtained with no Bonferroni correction – e.g., Gosselin & Schyns, 2001; Schyns et al., 2002; Schyns et al., 2003), or the Bonferroni correction is too conservative. Do we have a priori expectations that support the latter and can we use these expectations to our advantage? Abbey and Eckstein (2002), for example, have derived a statistical test far more sensitive than the Bonferroni correction for classification images derived from a two-alternative forced-choice paradigm *when the signal is perfectly known*. Although we often do not have such perfect *a priori* knowledge about the content of classification images, we do expect them to be relatively *smooth*.

The goal of *Bubbles* and reverse correlation is to reveal *clusters* of points that are associated with the measured response: e.g. the mouth or the eyes of a face (Gosselin & Schyns, 2001; Schyns et al., 2002; Mangini & Biederman, 2004; Gold et al., 2004; Sekuler et al., 2004), illusory contours (Gold et al., 2000), and so on. In other words, it is expected that the data points of classification images are correlated, introducing "smoothness" in the solutions. The Bonferroni correction, adequate when data points are independent, becomes far too conservative (not sensitive enough) for classification images with a correlated structure.

In the next section, we present two statistical tests based on *random field theory* that provide accurate thresholds for smooth, high-dimensional classification images.

# Random field theory

Adler (1981) and Worsley (e.g. 1994, 1995a, 1995b, 1996) have shown that the probability of observing a cluster of pixels exceeding a threshold in a smooth Gaussian random field is well approximated by the expected Euler Characteristic. The Euler Characteristic (EC) basically counts the number of clusters above a sufficiently high threshold in a smooth Gaussian random fields. Raising the threshold until only one cluster remains brings the EC value to 1; raising it still further until no cluster exceeds the threshold brings it to 0. Between these two thresholds, the expected EC approximates the probability of observing one cluster. The formal proof of this assertion is the centerpiece of *Random Field Theory* (RFT).

Next we present the main equations of two statistical tests derived from RFT: the so-called *Pixel* and *Cluster* tests, which have already been successfully applied for more than 15 years to brain imaging data. Crucially, these tests take into account the spatial correlation inherent to the data set, making them well suited for classification images.

## Pixel test

Suppose that *Z* is a Z-transformed classification image (see Appendix: The construction of a classification image). In RFT, the subset of *Z* searched for unlikely clusters of regression coefficients – e.g., the face area – is called the search space (S). The probability of observing at least one regression coefficient exceeding *t* is well approximated by

$$P(\max Z > t) \sim \sum_{d=0}^{D} Resels_d(S) \cdot EC_d(t) \qquad (1)$$

where $D$ is the dimensionality of $S$; $EC_d(t)$ is the d-dimensional *Euler characteristic* (EC) density which depends partly on the type of statistic (see Worsley, Marrett, Neelin, Vandal, Friston, & Evans, 1996, and Cao & Worsley, 2001 for EC densities of other random fields); $Resels_d(S)$ is the *d*-dimensional *resels* (resolution elements) which varies with the size and the shape of S. The EC densities of a $D$=2 dimensional Gaussian random field Z are

$$EC_0(t) = \int_t^{\infty} (2\pi)^{-\frac{1}{2}} e^{-\frac{u^2}{2}} du = p(Z > t) \qquad (2)$$

$$EC_1(t) = \frac{\sqrt{4\ln(2)}}{2\pi} \cdot e^{-\frac{t^2}{2}} \qquad (3)$$

$$EC_2(t) = \frac{4\ln(2)}{(2\pi)^{\frac{3}{2}}} \cdot t \cdot e^{-\frac{t^2}{2}} \qquad (4)$$

The resels is given by

$$Resels_d(S) = \frac{V_d(S)}{FWHM^d}, \qquad (5)$$

where $V_0(S)$=1 for a connected search region, $V_1(S)$=half perimeter length of $S$, $V_2(S)$=caliper area of $S$ (a disk of the same area as $S$ gives a good approximation and allows to derive the volumes of the lower dimension – see Cao & Worlsey, 2001). The *FWHM* is the Full Width at Half Maximum of the filter $f$ used to smooth the independent error noise in the image. If the filter is Gaussian with standard deviation $\sigma_b$ then

$$FWHM = \sigma_b \sqrt{8 \ln 2}. \qquad (6)$$

The filter $f$ should be chosen to give the best discrimination, or in other words, to maximize the detection of signal in Z. There is a classic theorem in signal processing, the Matched Filter Theorem, which states that to detect signal added to white noise, the optimum filter should match the shape of the signal. This implies that to optimally detect, say 10 pixel features, we should smooth the data with a 10 pixels *FWHM* filter. But if for instance it was felt that larger contiguous areas of the image were involved in discrimination, then this might be better detected by using a broader filter at the statistical analysis stage (see Worsley, Marrett, Neelin, Vandal, Friston, & Evans, 1996).

This dependency of the *Pixel* test on the choice of an adequate filter has led to a generalization of the test in which an extra dimension, the scale of the filter, is added to the image to create a *scale space* image (Poline & Mazoyer, 1994; Siegmund and Worsley, 1995). The scale space search reduces the uncertainty of choosing a filter *FWHM* but at the cost of higher thresholds.

## Cluster test

The *Pixel* test computes a statistical threshold based on the probability of observing a single pixel above the threshold. This test has been shown to be best suited for detecting focal signals with high Z-scores (Poline, Worsley, Evans & Friston, 1997). But if the region of interest in the search space (the mouth in a face for example) is wide, it has usually a lower Z-score and cannot be detected. We could improve detection by applying more smoothing to the image. The amount of smoothing will depend on the extent of the features we wish to detect (by the Matched Filter Theorem), but we do not know this in advance.

Friston, Worsley, Frackowiak, Mazziotta and Evans (1994) proposed an alternative to the *Pixel* test in order to improve the detection of wide signals with low Z-scores (see Poline et al., 1997, for a review). The idea is to set a low threshold ($t \geq 2.3$ – in the next section, we used $t$ = 2.7) and base the test on the size of clusters of connected pixels above the threshold. The *Cluster* test is based on the probability that, above a threshold $t$, a cluster of size $K$ (or more) pixels has occurred by chance which is calculated in the $D$=2 case as follows (Cao & Worsley 2001; Friston at al., 1994):

$$P(K > k) \sim 1 - \mathrm{e}^{\left(-\mathrm{Resels}_2(S) EC_2(t) p\right)}, \qquad (7)$$

where

$$p = e^{\left(-\left(\sqrt{2\pi} EC_2(t) k\right) / \left(FWHM^2 P(Z > t)\right)\right)} \qquad (8)$$

## Cluster vs. Pixel test

The *Cluster* and the *Pixel* test presented above provide accurate thresholds but for different types of signal. The *Pixel* test is based on the maximum of a random field and therefore is best adapted for focal signal (optimally the size of the FWHM) with high Z-scores (Siegmund and Worsley, 1995; Poline et al., 1997). The *Cluster* test is based on the size of a cluster above a relatively low threshold and therefore is more sensitive for detecting wide regions of contiguous pixels with relatively low Z-scores. The two tests potentially identify different statistically significant regions in smooth classification images. Figure 1 illustrates this point with a one-dimensional classification image comprising 257 pixels convolved with a Gaussian kernel with a *FWHM* equals to 11.8 pixels. For a p-value equals to .05, the *Pixel* test gives a threshold of 3.1 (green line) whereas the *Cluster* test gives a minimum cluster size of 6.9 above a threshold of 2.7 (red line).

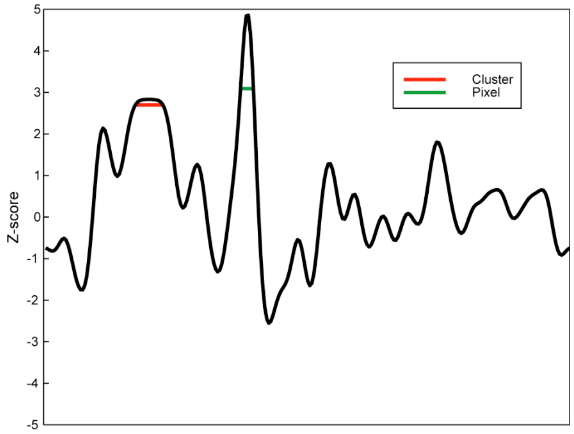


Figure 1. Regions revealed by the *Cluster* (red) vs. the *Pixel* (green) test. See text for details.

Furthermore the interpretation of the results following the application of the *Pixel* and the *Cluster* test differs drastically. On the one hand, the *Cluster* test allows the inference that the clusters of Z-scores larger than the minimum size are significant, not that the individual Z-scores inside these clusters are significant. On the other hand, the *Pixel* test allows the conclusion that each individual Z-score above threshold is significant (Friston, Holmes, Poline, Price, Frith, 1996; Friston et al., 1994; Poline et al., 1997).

## Accuracy

Since the late 1980's, *Random Field Theory* (RFT) has been used to analyze Positron Emission Tomography (PET) images, galaxy density maps, cosmic microwave background data and functional Magnetic Resonance Imaging (fMRI) data. In fact, the RFT is at the heart of two popular fMRI data analysis packages: SPM2 (Frackowiak, Friston, Frith, Dolan, Price, Ashburner, Penny & Zeki, 2003) and fmristat (Worsley, 2003).

Not surprisingly, the accuracy of RFT has been examined extensively. An accurate statistical test must be both sensitive (i.e., high hit rate) and specific (i.e., high correct rejection rate). In particular, RFT has been evaluated in the context of so-called “phantom” simulations (Poline, et al., 1997; Hayasaka & Nichols, 2003, Hayasaka, Luan Phan, Liberzon, Worsley & Nichols, 2004, Nichols & Hayasaka, in press; Worsley submitted). A “phantom” simulation basically consists in generating a lot of smooth random regression coefficients, in hiding a “phantom” in them (i.e., a known signal – usually a disc or a Gaussian), in attempting to detect the “phantom” with various statistical tests, and in deriving, per statistical test, a measure of accuracy such as a d-prime or a ROC area. We singled-out “phantom” simulations for a reason: If we were to compare the accuracy of various statistical tests for the detection of a “phantom” template (e.g., used by a Linear Amplifier Model) in a smooth classification images this is exactly what we would have to do. In other words, these “phantom” simulations inform us just as much about the accuracy of RFT for fMRI data than about its accuracy for classification images.

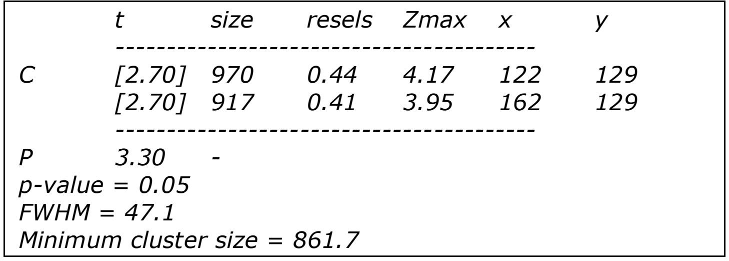
To summarize these accuracy assessments: The *p*-values given by RFT appears to be more accurate than those given by the Bonferroni, the Hochberg, the Holm, the Sidák and the False Discovery Rate provided that the size of the search space is greater than about three times that of the *FWHM* (Hayasaka & Nichols, 2003), that the *FWHM* is greater than about 5 pixels (Taylor, Worsley, & Gosselin, submitted) and that the degree of freedom is greater than about 200. Also, the *Cluster* test is more sensitive and less specific than the *Pixel* test.

# Re-analyzing representative classification images

In the final section of this article, we apply the *Pixel* and *Cluster* tests to classification images representative of the literature from Gosselin and Schyns (2001) and Sekuler, Gaspar, Gold & Bennett (2004). We give sample commands for the Stat4Ci Matlab toolbox throughout.

## Matlab implementation

A mere four pieces of information are required for the computation of the significant regions using the *Pixel* and the *Cluster* tests: a desired *p*-value, a threshold *t* (only used for the *Cluster* test), a search space, and the *FWHM* – or, equivalently, the sigma – of the Gaussian kernel used to smooth the classification image. The main function from the Stat4Ci Matlab toolbox – *StatThresh.m* – inputs this information together with a suitably prepared classification image (i.e., smoothed and Z-transformed), performs all the computations described above, and outputs a threshold for the *Pixel* test as well as the minimum size of a significant cluster for the *Cluster* test. The *StatThresh.m* function makes extensive use of the *stat_threshold.m* function, which was originally written by Keith Worsley for the fmristat toolbox.



```
          t         size     resels   Zmax    x         y
          -----------------------------------------
C         [2.70]    970      0.44     4.17    122       129
          [2.70]    917      0.41     3.95    162       129
          -----------------------------------------
P         3.30      -
p-value = 0.05
FWHM = 47.1
Minimum cluster size = 861.7
```

Figure 2. Sample summary table produced by *DisplayRes.m* (from the re-analysis of classification images from Gosselin and Schyns, 2001; see next section). The numbers between brackets were set by the user. C = Cluster test; P = Pixel test; *t* = threshold; *size* = size of the cluster; *Zmax*, *x* and *y* = maximum Z-score and its coordinates.

Other functions included in the Stat4Ci toolbox perform a variety of related computations: e.g., *ReadCid.m* reads a Classification Image Data (CID) file; *BuildCi.m* constructs classification images from a CID file; *SmoothCi.m* convolves a raw classification image with a Gaussian filter; *ExpectedSCi.m* computes the expected mean and standard deviation of a smooth classification image (see Appendix: The construction of a classification image); *ZTransSCi.m* Z-transforms a smoothed classification image (see Equation 9 and Appendix: The construction of a classification image); *DisplayRes.m* displays the thresholded Z-transformed smooth classification image and ouputs a summary table (see Figure 2). All of these functions include thorough help sections.

## Sekuler, Gaspar, Gold and Bennett

Sekuler, Gaspar, Gold and Bennett (2004) examined the effect of face inversion on the information used by human observers to resolve an identification task. Four classification images extracted using reverse correlation are re-analyzed: one for each combination of two subjects (MAT and CMG) and two conditions (UPRIGHT and INVERTED). Each classification image cumulates the data from 10,000 trials. We will not further describe this experiment. Rather we will limit the presentation to what is required for to application of the *Pixel* and *Cluster* tests.

First, the raw classification images must be convolved with a Gaussian filter, i.e. *smoothed*. The choice of the appropriate Gaussian filter depends essentially on the size of the search space (see Worsley, submitted, for a discussion). We chose a Gaussian filter with a standard deviation of $\sigma_b$ = 4 pixels; its effect are similar to those of the filter used by Sekuler, Gaspar, Gold and Bennett (2004). Second, the smooth classification images must be Z-transformed. This can sometimes be achieved analytically (see Appendix: The construction of a classification image). However, if the number of trials is greater than 200 – as is usually the case with classification images – the Z-transformation can be approximated as follows:

$$ZSCi = \frac{SCi - \overline{SCi}}{\sigma_{SCi}}, \qquad (9)$$

where the mean and standard deviation are estimated directly from the data, preferably from signal-less regions of the classification images (e.g., regions corresponding to a homogeneous background). In the Stat4Ci toolbox, classification image preparation can be done as illustrated in Figure 3.

```
SCi = double(imread('GenderCi.tiff'));
sigma_b = 20; %std of smoothing filter
S_r = double(imread('faceMask.tif'));

vecSCi = SCi(eq(S_r,0));
ZSCi = ZTransSCi(SCi,...
   [mean(vecSCi(:)),std(vecSCi(:))]);
ZSCi = ZSCi .* S_r;

p = .05;       %p-value
tC = 2.7;      %threshold (for Cluster test)
Res = StatThresh(ZSCi,p,sigma_b,tC,S_r);

background = double(imread('w1H.JPG'));
tCi = DiplayRes(Res,background);
```

Figure 3. Sample commands for the Stat4Ci Matlab toolbox (from the re-analysis of classification images from Gosselin and Schyns, 2001).

Once the classification image has been smoothed and Z-transformed, it must be inputted into the *StatThresh.m* function together with the four additional required pieces of information: a *p*-value ($p \le .05$), the sigma of the Gaussian filter used during the smoothing phase ($\sigma_b$ = 4 pixels for this re-analysis), a threshold for the *Cluster* test (equal to 2.7 for this re-analysis) and a search space (i.e., the face region).

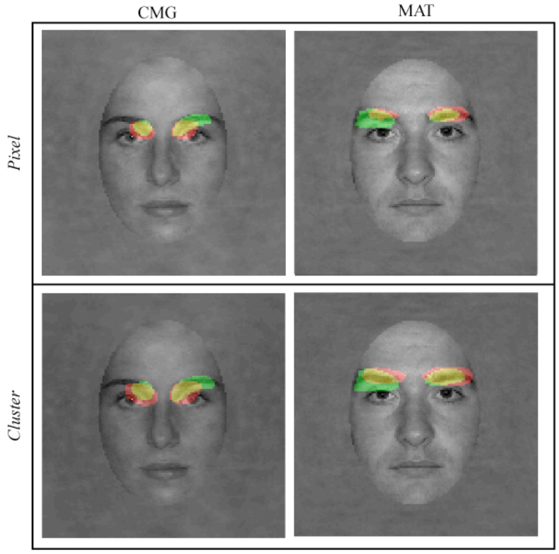


Figure 4. Sekuler, Gaspar, Gold and Bennett's (2004) classification images re-analyzed using the Stat4Ci Matlab toolbox.

The statistical threshold obtained using the *Pixel* test is very low compared with that obtained using the Bonferroni correction (i.e., 3.67 rather than 4.5228; see section The Bonferroni correction). In fact, the *stat_threshold.m* function outputs the minimum between the Bonferroni and the *Pixel* test thresholds. Figure 4 displays the thresholded classification images for both the *Pixel* and *Cluster* tests. For the *Cluster* test, only the clusters larger than the minimum size (i.e., 66.6 pixels) are shown. Red pixels indicate the regions that attained significance in the UPRIGHT condition; green pixels, in the INVERTED condition; and yellow pixels, in both. A face background was overlaid to facilitate interpretation.

## Gosselin and Schyns

Gosselin and Schyns (2001, Experiment 1) examined the information used by human observers to resolve a GENDER and an expressive vs. not expressive (EXNEX) face discrimination task. They employed the *Bubbles* technique to extract two classification images per observer, one per task. Each of the two classification images re-analyzed in this section combines the data from 500 trials executed by subject FG. The classification images can be built either from the opaque masks punctured by Gaussian holes and applied multiplicatively to a face on each trial, or from the center of these Gaussian holes. The former option naturally results in smooth classification images; and the latter option calls for smoothing with a filter, just like the classification images of Sekuler, Gaspar, Gold and Bennett (2004). For this re-analysis, we used a Gaussian filter identical to the one used to sample information during the actual experiment ($\sigma_b$ = 20 pixels). In this case, both options are strictly equivalent. These smooth classification images (see *SCi* in Figure 3) were Z-transformed using Equation 9 with estimations of the expected means and standard deviations based on the signal-less pixels outside the search region (see *S_r* in Figure 3).

Next, the Z-transformed smooth classification image (see *ZSCi* in Figure 3) is inputted into the *StatThresh.m* function with the four additional required pieces of information: a *p*-value ($p \leq .05$), the sigma of the Gaussian filter used during the smoothing phase ($\sigma_b$ = 20 pixels for this re-analysis), a threshold for the *Cluster* test (equal to 2.7 for this re-analysis – see *t*C in Figure 3) and a search space (see *S_r* in Figure 3). See Figure 3 for all the relevant Stat4Ci toolbox commands.

Again, the statistical threshold obtained using the Pixel test is extremely low compared with that obtained using the Bonferroni correction: 3.30 rather than 4.808 (see section The Bonferroni correction). Figure 5 displays the thresholded classification images for both the *Pixel* and *Cluster* tests. For the *Cluster* test, only the clusters larger than the minimum size (i.e. 861.7 pixels) are shown. Red pixels indicate the regions that attained statistical significance. A face background (see *background* in Figure 3) was overlaid to facilitate interpretation.

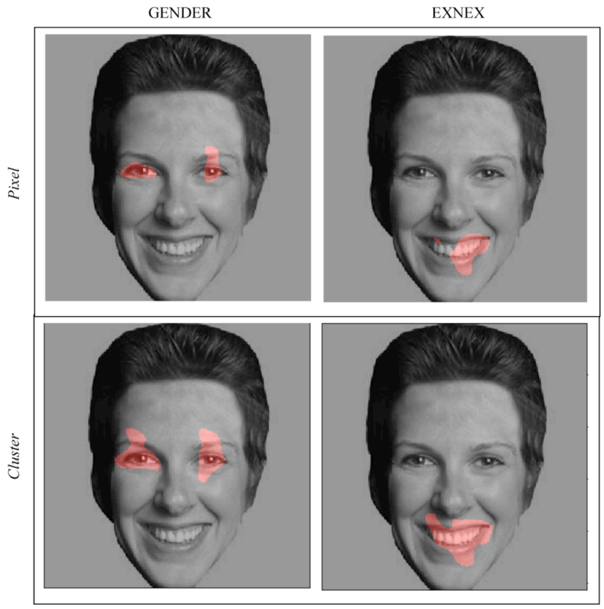


Figure 5. Two of Gosselin and Schyns' (2001) classification images re-analyzed using the Stat4Ci Matlab toolbox.

# Take-home message

We have presented two statistical tests suitable for smooth, high-dimensional classification images, in the absence of *a priori* expectations about the shape of the signal. The *Pixel* and the *Cluster* tests, based on *Random Field Theory*, are accurate within known boundaries discussed in the article. These tests require only four pieces of information and their computation can be performed easily using the Stat4Ci Matlab toolbox. We expect these tests to be most useful for researchers applying *Bubbles* or reverse correlation to complex stimuli.

# Acknowledgments

This research was supported by a NSERC (249974) and a NATEQ (03NC082052) grant awarded to Frédéric Gosselin; by a NATEQ (84180) grant awarded to Martin Arguin and Frédéric Gosselin; and by a ESRC grant R000237901 to Philippe G. Schyns. We thank Allison Sekuler, Carl Gaspar, Jason Gold and Patrick Bennett for having kindly given us access to their classification images.

Commercial relationships: None.

Corresponding author: Frédéric Gosselin

Address: Département de psychologie, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada, H3C 3J7.

Email: frederic.gosselin@umontreal.ca.

## Appendix: The construction of a classification image

In a reverse correlation or *Bubbles* experiment, an observer is presented with a noise field on trial $i$ ($i$ = 1, ..., $n$) and produces the response $Y_i$.

At a particular pixel $v$, we suppose that some feature of the noise field, $X_i(v)$, is correlated with the response. In a reverse correlation experiment, $X_i(v)$ might be the added noise at pixel $v$; in a *Bubbles* experiment, $X_i(v)$ might be the actual bubble mask at pixel $v$. We aim to detect those pixels where the response is highly correlated with the image feature of interest. The sample correlation at pixel $v$ is

$$\mathrm{C}(\mathrm{v}) = \sum_{\mathrm{i}} \frac{\left(X_i(v) - \bar{X}(v)\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_i \left(X_i(v) - \bar{X}(v)\right)^2}\sqrt{\sum_i \left(Y_i - \bar{Y}\right)^2}}, \qquad (10)$$

where bar indicates averaging over all $n$ trials. It is straightforward to show that if there is no correlation between image features and response, then

$$\mathrm{Z}(\mathrm{v}) = \frac{\sqrt{\mathrm{n}\text{-}2}\,C(v)}{\sqrt{1 - C(v)^2}} \sim \sqrt{n}\,C(v) \qquad (11)$$

has a Student-t distribution with $n$-2 degrees of freedom provided $X_i(v)$ is Gaussian, but in any case $n$ is usually very large so the standard normal distribution will be a very good approximation, by the Central Limit Theorem.

Provided that $\sum X_i=0$ and $\sum Y_i=0$, the *ZTransSCi.m* function from the Stat4Ci toolbox implements Equation 10. In this case, the numerator is simply the sum of all the noise fields weighted by the observer's responses. The remaining term in C($v$) can be approximated if $X_i(v)$ is white noise $W_i(v)$ (i.e. independent and identically distributed noise at each pixel) convolved with a filter $f(v)$, that is if

$$X_i(v) = \sum_u f(v-u)\, W_i(u)\,. \qquad (12)$$

In the case of reverse correlation, $W_i(v)$ is usually a Gaussian random variable and often there is no filtering, so $f(v)$ is zero except for $f(0)$=1. In the case of *Bubbles*, $W_i(v)$ is a binary random variable taking the value 1 if there is a bubble centered at v, and 0 otherwise. The

$$\sum_{\mathrm{i}} \left(X_i(v) - \bar{X}(v)\right)^2 \sim n\sigma^2 \sum_u f(v)^2\,, \qquad (13)$$

where $\sigma^2$ is the variance of the white noise. For reverse correlation, $\sigma^2$ is the variance of the Gaussian white noise. For *Bubbles* it is the Binomial variance

$$\sigma^2 = \frac{N_b}{N}\left(1 - \frac{N_b}{N}\right) \sim \frac{N_b}{N}\,, \qquad (14)$$

where $N_b$ is the number of bubbles and $N$ is the number of pixels.

The Central Limit Theorem ensures that, at the limit, Z(v) is a Gaussian random field with an effective FWHM equals to the FWHM of the filter f(v). The rate of convergence toward Gaussianity depends partly on the predictive variable and partly on the total number of bubbles per Resels. Worsley (in preparation) has examined the exactness of the p-values given by the Gaussian procedures presented in this article in function of these two factors: At 10,000 bubbles per Resels, the p-values given by the Gaussian procedures depart from the true p-values by less than +/- 0.04 logarithmic unit; at 500 bubbles per Resels, a figure more often encountered in practice (e.g., Gosselin & Schyns, 2001), the discrepancy can be as much as +/- 0.3 logarithmic unit. If the predictive variable has a positively skewed distribution, the Gaussian procedure is liberal; and if it has a negatively skewed distribution, as is usually the case in practice (e.g., Gosselin & Schyns, 2001), the Gaussian procedure is conservative.

## References

Adler, R. J. (1981). *The Geometry of Random Fields*. New York: Wiley.

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P. G. & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature, 433*, 68-72 [PubMed] [Article]

Ahumada, A. J., Jr. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception*, 26, 18. [Link]

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, 2(1), 121-131, [PubMed] [Article]

Abbey, C. K., & Eckstein, M. P. (2002). Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. *Journal of Vision*, 2(1), 66-78. [PubMed] [Article]

Barth, E., Beard, B. L., & Ahumada, A. J. (1999). Nonlinear features in vernier acuity. In B. E. Rogowitz and T. N. Pappas (Eds.), *Human Vision and Electronic Imaging IV, SPIE Proceedings, 3644*, paper 8.

Beard, B. L. & Ahumada, A. J. (1998). A technique to extract the relevant features for visual tasks. In B. E. Rogowitz and T. N. Pappas (Eds.), *Human Vision and Electronic Imaging III, SPIE Proceedings, 3299*, 79-85.

Bonnar, L., Gosselin, F., & Schyns, P. G. (2002). Understanding Dali's Slave Market with the Disappearing Bust of Voltaire: A case study in the scale information driving perception. *Perception, 31*, 683-691. [PubMed][Article]

Cao, J., & Worsley, K. J. (2001). Applications of random fields in human brain mapping. In M. Moore (Ed.) *Spatial Statistics: Methodological Aspects and Applications, Springer Lecture Notes in Statistics, 159*, 169-182.

Eckstein, M. P., & Ahumada, A. J. (Ed.). (2002). Classification images: A tool to analyze visual strategies [Special issue]. *Journal of Vision, 2(1)*, i-i, http://journalofvision.org/2/1/i/, doi:10.1167/2.1.i. [PubMed] [Article]

Frackowiak, R., Friston, K. J., Frith, C., Dolan, R., Price, C., Ashburner, J., Penny, W., & Zeki, S. (2003) *Human Brain Function*, Academic Press; 2 edition

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta J. C., & Evans A. C. (1994). Assessing the Significance of Focal Activations Using Their Spatial Extent. *Human Brain Mapping, 1*, 214-220.

Friston KJ, Holmes A, Poline JB, Price CJ, Frith CD. (1996) Detecting activations in PET and fMRI: levels of inference and power. *Neuroimage. 4(3)*, 223-35. [PubMed] [Article]

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology, 10*, 663-666. [PubMed] [Article]

Gold, J. M., Sekuler, A. B., & Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cognitive Science, 28*, 167-207. [Article]

Gosselin, F., Bacon, B. A., & Mamassian, P. (2004). Internal surface representations approximated by reverse correlation. *Vision Research, 44*, 2515-2520. [Article]

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition. *Vision Research, 41*, 2261-2271. [PubMed] [Article]

Gosselin, F., & Schyns, P. G. (2002). RAP: a new framework for visual categorization. *Trends in Cognitive Science, 6*, 70-77. [Article]

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of memory representations. *Psychological Science*, 14, 505-509. [PubMed] [Article]

Gosselin, F., & Schyns, P. G. (2004a). No troubles with Bubbles: A reply to Murray and Gold. *Vision Research*, 44(5), 471-477. [PubMed] [Article]

Gosselin, F., & Schyns, P. G. (Ed.). (2004b). A picture is worth thousands of trials: Rendering the use of visual information from spiking neurons to recognition [Special issue]. *Cognitive Science, 28*, 141-146. [Article]

Hayasaka, S., & Nichols, T. (2003) Validating cluster size inference: random field and permutation methods. *NeuroImage, 20*, 2343-2356. [PubMed][Article]

Hayasaka, S. Luan Phan, K. Liberzon, I. Worsley, K. J., & Nichols, T. (2004) Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage, 22*, 676-687. [PubMed][Article]

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: estimating the information employed for face classifications. *Cognitive Science, 28*, 209-226. [Article]

Kontsevich, L. L., & Tyler, C. W. (2004) What makes Mona Lisa smile? *Vision Research, 44*, 493-1613. [PubMed] [Article]

Marmarelis, P. Z., & Naka, K. I. (1972). White-noise analysis of a neuron chain: An application of the Wiener theory. *Science, 175*, 1276-1278. [PubMed]

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. G. (in press). The use of visual information in natural scenes. *Visual Cognition.*

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision, 2(1)*, 79-104, [PubMed] [Article]

Murray, R. F., & Gold, J. M. (2004) Troubles with bubbles, Vision Research, 44(5), Pages 461-470. [PubMed] [Article]

Neri, P., & Heeger, D. (2002) Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nature Neuroscience, 5*, 812-816. [PubMed] [Article]

Neri, P., Parker, A. J., & Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. *Nature, 401*, 695-698. [PubMed] [Article]

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science, 249*, 1037-1041. [PubMed]

Olman, C., & Kersten, D. (2004). Classification objects, ideal observers & generative models. *Cognitive Science, 28*, 141-146. [Article]

Poline, J-B., & Mazoyer, B.M. (1994) Analysis of individual brain activation maps using hierarchical description and multiscale detection. *IEEE Trans. Med. Imag.*, *13*(*4*), 702–710. [Article]

Poline, J-B., Worsley, K. J., Evans, A. C., & Friston, K. J (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. *Neuroimage, 5*, 83-96. [PubMed] [Article]

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science, 28*, 247-166. [Article]

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science, 13*, 402-409. [PubMed] [Article]

Schyns, P. G., Jentzsch, I., Johnson, M., Schweinberger, S. R., & Gosselin, F. (2003). A principled method for determining the functionality of ERP components. *Neuroreport, 14*, 1665-1669. [PubMed] [Article]

Sekuler, A. B., Gaspar, C. M., Gold, J. M., & Bennett, P. J. (2004). Inversion leads to quantitative changes in faces processing. *Current Biology, 14*, 391-396. [PubMed] [Article]

Siegmund, D.O. and Worsley, K.J. (1995). Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Annals of Statistics*, *23*, 608-639.

Simpson, W. A., Braun, J., Bargen, C., & Newman, A. (2000). Identification of the eye-brain-hand system with point processes: A new approach to simple reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1675-1690. [PubMed] [Article]

Smith, M. L., Gosselin, F. & Schyns, P. G., (2004). Receptive fields for flexible face categorizations. *Psychological Science*, 15, 753-761. [PubMed] [Article]

Smith, M. L., Cottrell, G., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions of emotions. *Psychological Science, 16*, 184-189 [Article]

Sutter, E. E., & Tran, D. (1992). The field topography of ERG components in man–I: The photopic luminance response. *Vision Research, 32*, 433-446. [PubMed] [Article]

Taylor, J. E., Worsley, K. J. & Gosselin, F. (submitted). Maxima of discretely sampled random fields, with an application to 'bubbles'.

Thomas, J. P., & Knoblauch, K. (1998). What do viewers look for when detecting a luminance pulse? *Investigative Opthalmology and Visual Science, 39*, S404.

Vinette, C., Gosselin, F., & Schyns, P. G. (2004). Spatiotemporal dynamics of face recognition in a flash: It's in th eyes! *Cognitive Science, 28*, 289-301. [Article]

Watson, A. B. (1998). Multi-category classification: template models and classification images. *Investigative Opthalmology and Visual Science, 39*, S912.

Watson, A. B., & Rosenholtz, R. (1997). A Rorschach test for visual classification strategies. *Investigative Opthalmology and Visual Science, 38*, S1.

Worsley, K.J. (1994). Local maxima and the expected Euler characteristic of excursion sets of chi^2, F and t fields. *Advances in Applied Probability*, *26*, 13-42.

Worsley, K. J. (1995a). Boundary corrections for the expected Euler characteristic of excursion sets of random fields, with an application to astrophysics. *Advances in Applied Probability*, *27*, 943-959.

Worsley, K. J. (1995b). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Annals of Statistics*, *23*, 640-669.

Worsley, K. J. (1996). The geometry of random images. *Chance*, *9*, 27-40.

Worsley, K. J. (2003). FMRISTAT: A general statistical analysis for fMRI data. Available: http://www.math.mcgill.ca/keith/fmristat/

Worsley, K. J. (submitted). An improved theoretical P-value for SPMs based on discrete local maxima. *NeuroImage*.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping, 4*, 58-73. [Article]