Automatic Detection of the Head of the Hippocampus

Camille Izard
Center for Imaging Science
The Johns Hopkins University

Université Paris XI
Institut Natinal Agronomique Paris-Grignon

September 8, 2004

Abstract

In order to analyze Magnetic Resonance brain Images (MRI), neurobiologists need to define landmarks. These points characterize the anatomy and are identifiable on each subject, so that they can be used as a reference system. Manually detecting these points in 3D structures is tricky even for experienced neuroscientists. This work focuses on the automatic detection of the Head of the Hippocampus from the gray levels of the MRI. The proposed algorithm is composed of two parts. Firstly it deals with a training set in order to learn a template, which is a random image. Secondly it takes in input new images and predict the Head of the Hippocampus location, using the template. The predicted point is the location where the template fits the best in the image. This approach had to face with image normalization and gray level distributions estimation. In addition in order to reduce the computation time, the informative voxels were selected using conditional variance. Finally, experimental results are obtained on images provided by Craig Stark, Department of Psychological and Brain Sciences, the Johns Hopkins University, USA.

Je remercie tout particulièrement mon maître de stage, Bruno Jedynak, pour le temps qu'il m'a consacré au cours des quatres mois de préparation de ce travail, mais aussi Laurent Younès et Don Geman pour leurs précieux conseils.

Je suis également reconnaissante envers tout ceux qui m'ont aidée et soutenue tout au long de cette année de DEA, d'un côté comme de l'autre du grand Atlantique.

Contents

1 Introduction

The central difficulty of the analysis of functional neuroimaging studies is the huge variability of the brain's structures. The alignment of the scans is the first step of any study. However classical methods are not accurate enough to allow a powerful comparison between experiments. Defining some points in the brain, called landmarks, will facilitate the alignment process, even if up to now the landmarking has been a manual step. The challenge is to complete this tough task automatically.

To make the correspondence between two images, point-based registration methods need suitable landmarks. Two different approaches of the automatic detection of landmarks may be distinguished. In the first case a landmark is defined as a point where the surface curvature is strong [1]. They may also but not necessarily corresponds to manually defined landmarks. The main strategy for automatic detection uses 3D filters, to identify high response locations as potential landmarks. Consequently this method has to deal with false positive, but several solutions are proposed.

In the second case, the landmark is previously manually defined as a point which characterizes the anatomy and which is detectable in all scans. Points like the Anterior Commissure (AC) or the Posterior Commissure (PC), used by J. Talairach and P. Tournoux [2] to perform the so-called Talairach transformation on MRI, offer a coordinate system for the whole brain. Although this aligning transformation has been performed on the images, some structures are not clearly located in the brain. Therefore it is necessary to add landmarks which characterize the structure location, independently to the reference system. These points are often extreme voxels of a structure, like the apex of the hippocampal head or the tail of the hippocampus. Once these two landmarks have been detected, the location of the hippocampus is perfectly known, allowing to align accurately the structures.

We will focus on the second class of landmarks, and more particularly on the head of the hippocampus (HoH). There is no literature in the detection of such points in Magnetic Resonance Images. The exceptions are AC and PC [3] [4], but they are singular points, whose detection may be considered as easier than detecting the head of the hippocampus for example. To perfect the alignment seven to eleven points are needed, we chose to start with only one landmark, the apex of the left hippocampal head.

The algorithm we are going to use is composed of two parts. The first one is the off-line algorithm that consists in learning the location of the landmark thanks to the training set's images. It takes in input the gray levels of the image and gives in output a probability map, which is called a template in the rest of this thesis. The second part is the online algorithm, which takes in input new images, and gives in output a prediction of the HoH location, using for the prediction the template created in the first step.

This thesis is structured as follows. We will firstly discuss in section 2 the issue of images' manual landmarking and in section 3 the normalization of the images. Secondly we will introduce in section 4 the model we use to predict the landmark location. Then, once we have obtained the template (section 5), we will consider different ways to select informative variables in this template (section 6). Finally, some results obtained on new images are presented (section 7).

2 The tricky detection of the Head of the Hippocampus

2.1 The Hippocampus in the brain

The brain is described in three parts:

- the forebrain is the largest part of the brain, mainly made of the cerebrum and that contains also the thalamus, the hypothalamus and the limbic system.
- the hindbrain includes the cerebellum, the pons and the medulla oblongata.
- the midbrain is a structure between the forebrain and the hindbrain. This is the region where connections between the spinal cord and the brain are made.

A great and simple introduction to Brain anatomy can be found at [5].
The Hippocampus is a part of the limbic system, a neuronal system composed of disparate anatomical structures, with common functions. Among its different units there are the Hippocampus itself and some immediate surrounding structures, like the amygdala and the subiculum. The hippocampus received its name because of its seahorse shape. It is described with three parts, the body which is sagitally oriented, the head and the tail which are both transversally oriented. It is composed of two interlocking U-shaped laminae, consisting of the cornu Ammonis and the gyrus dentatus. The reader is referred to [6] for additional details about the anatomy. The figure 1 presents the spatial organization of the structures close to the hippocampus in a sagittal slice.

2.2 Working on Magnetic Resonance brain Images

Professor Stark from the Department of Psychological and Brain Sciences, The Johns Hopkins University, USA, provided MRI of the brain, acquired on a Philips Intera 3-Tesla scanner with resolution of $1mm^3$. The images are $161 \times 191 \times 151$ mm and visualized with the software AFNI [7]. They were manually transformed to the Talairach space. The purpose of this step is to surround the entire brain within a grid system, so that all the scans can

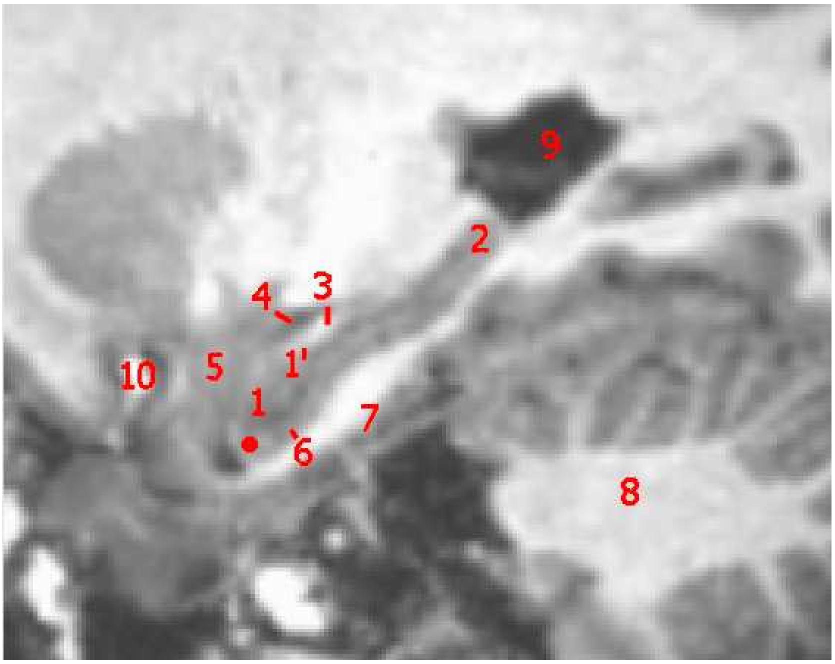

Figure 1: MRI image, Sagittal section of the brain centered on the hippocampus, Bar, 10mm. The dot corresponds to the apex of the head of the hippocampus. 1, hippocampal head cornu Ammonis, 1', hippocampal head gyrus dentatus, 2, hippocampal tail, 3, uncal apex, 4, temporal horn of the lateral ventricle, 5, amygdala, 6, subiculum, 7, parahippocampal gyrus, 8, cerebellum, 9, lateral ventricle, 10, middle cerebral artery

be aligned in the same position. It is a rigid transformation to a coordinate system based on the line between the anterior commissure (AC) and the posterior commissure (PC). Three other lines are used: the vertical AC and the vertical PC which are perpendicular to the AC-PC line, and the midline, which follows the interhemispheric fissure. The volume obtained is limited by a set of landmarks on the cortical periphery, [2]. Since this transformation is already performed on the images we will use as a training set, the proposed method must be used after the Talairach transformation.

Even if it is often manually perform, some methods exist to perform it automatically. We found two papers that describe automatic identification of AC and PC for the Talairach registration. Both papers distinguish two steps, first the midsagittal plane extraction and then the detection of the landmarks. (Vérard et al., 1997) [3] used scene analysis to determine the location of AC and PC once the midsagittal plane has been extracted. (Han and

Park, 2003) [4] find AC and PC by template matching, using template of well conserved surrounding shapes in the image like the corpus callosum's edge. These methods are efficient for the AC and PC automatic detection, however the used approaches are specific to these particular points, considered as easy to detect, since they are isolated in the brain structures.

Once the alignment has been performed, we note remaining variability of approximately 10 mm for the apex of the head of the hippocampus. It shows that even if the transformation has a great accuracy at the whole-brain scale, it ignores the neuroanatomical boundaries or landmarks.

2.3 Manual Detection of the Head of the Hippocampus (HoH)

The apex of the head of the hippocampus is defined as the extreme point of the hippocampus in the sagittal slice that corresponds to the largest extension of the head. This point is really tough to landmarked with precision because of the three dimensional structure of the hippocampus. The angle of the structure with the slices orientation varies from one image to another. Consequently, neuroscientists have to imagine the 3D image to find the best slice and then landmark the extreme point on the head, which is sometimes almost spherical. A white thin line that outlines the boundary with the amygdala helps to locate the apex.

The result obtained with an automatic detector depends largely on the training set precision. Even if it is difficult to evaluate the human performance in HoH landmarking, we ask to both an expert and a graduate student to landmark twice the same images, with several weeks between the two experiments. The mean Euclidian distance between the two landmarks of the expert is $0.93mm$ with standard deviation (σ) 0.96. We consider it as a standard reference for our scans. Student's results are slightly different, leading to a mean distance of $3.92mm$ with a standard deviation of 1.98. To finish, we measured the distance between student's and expert's landmarks. The mean distance is $3.26mm$ ($\sigma = 0.98$). Even if the student may reach the precision of the expert with further training, these results show how challenging it is to landmark the HoH, even manually.

3 Modelling the gray levels to normalize the images

To use the gray levels it is previously necessary to normalize the images. Magnetic Resonance brain Images should ideally be composed of only three intensities corresponding to the three matters that are in the brain. The lowest intensities correspond to the Cerebrospinal Fluid (CSF), the intermediate gray levels to the gray matter (GM) and the highest intensities to the white matter (WM). Of course there are also some intermediate intensities due to the mixed voxels which make the segmentation of MRI a challenging problem. In our case we are looking for a continuous segmentation, to reduce the space of the gray levels to a three dimensional space. One intensity is then characterized by its probability to be in each of the three matters.
Different approaches, that might be complicated, are used in order to solve this problem, we will first consider the easiest, that is to model the gray levels distribution as a mixture of 3-Gaussian distributions.

One image is considered as a set of voxels $\{s \in S\}$, each voxel being associated with a gray level intensity $x_s \in I$. The location of the apex of the hippocampal head is notated y and corresponds to one voxel of the image.

3.1 Mixture of Gaussian distributions

Let's define the random variable Z_s that gives the type of matter $j \in \{CSF, GM, WM\}$ at the voxel s, then the Gaussian mixture can be written, for a given image i:

$$\begin{aligned} P(X_s = x_s) &= \sum_{j=1}^{3} P(X_s = x_s | Z_s = j) P(Z_s = j) = \sum_{j=1}^{3} \alpha_j g_j(x_s) \\ \text{with } g_j(x_s) &= \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp - \frac{|x_s - \mu_j|^2}{2\sigma_j^2} \end{aligned}$$

The eight parameters of the mixture are estimated on each image by an EM algorithm [8]. The algorithm consists in two alternative and iterative steps: Expectation and Maximization. We define a starting point by kmeans and a maximal number of iterations. The log-likelihood of the marginal density is maximized over the parameters θ, the hidden variables α.

$$\max_{\alpha,\theta} \sum_{s\in S} \log \int p_\theta(\alpha, x_s) d\mu(\alpha)$$

The solution is given by the derivative under θ. It is equivalent to say that:

$$E_\theta \left[\frac{d}{d\theta} \log p_\theta(., x_s) | x_s \right] = 0 \tag{1}$$

So θ_{n+1}, given θ_n, is solution of: $E_{\theta_n}[\psi(\theta_{n+1}|x_s)] = 0$.

In case of a Gaussian mixture, one has to maximize over θ and α for the given the image i:

$$\sum_{s\in S} \log \sum_{j=1}^{3} \frac{\alpha_j}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{|x_s - \mu_j|^2}{2\sigma_j^2} \right)$$

Here $\theta = (\alpha_j, \mu_j, \sigma_j^2)$ and the α_j are the hidden variables containing the proportions of the matters in the mixture. Let's define:

$$p_\theta(j, x_s) = \frac{\alpha_j}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{|x_s - \mu_j|^2}{2\sigma_j^2} \right) \tag{2}$$

Given the result of the n^{th} iteration, the posterior probability (2) can be written thanks to the parameters of the previous iteration, this is the Expectation step.

$$p_{\theta_n}(j|x_s)^{(n+1)} = \frac{\alpha_j^{(n)} g_j^{(n)}(x_s)}{\sum_{j'=1}^{3} \alpha_{j'}^{(n)} g_{j'}^{(n)}(x_s)}$$

For the Maximization step, one has to solve the equation (1). In case of a mixture of Gaussian distributions, the partial derivatives of (2) admits an analytic solution.

$$\begin{aligned} \frac{\partial \log p_\theta(j, x_s)}{\partial \mu_j} &= \delta_{jj'}(x_s - \mu_{j'}) \\ \text{so } \mu_j^{(n+1)} &= \frac{\sum_{s\in S} p_{\theta_n}(j|x_s) x_s}{\sum_{s\in S} p_{\theta_n}(j|x_s)} \end{aligned}$$

With N the cardinal of S, the solution for the other parameters is:

$$\alpha_j^{(n+1)} = \frac{\sum_{s\in S} p_{\theta_n}(j|x_s)}{N} \text{ and } \sigma^{2\,(n+1)}_{j'} = \frac{\sum_{s\in S} |x_s - \mu_j^{(n)}|^2 p_{\theta_n}(j|x_s)}{\sum_{s'\in S} p_{\theta_n}(j|x_{s'})}$$

3.2 Results on the images

The normalization is performed on all the images independently, both on the training set and on the testing set. The full histogram of one image is too huge and contains also voxels from the background. Therefore we decided to use a 41-by-41-by-41 voxel box centered on AC (this point is already known since all the images are displayed in the Talairach view). We assume that this box contains the same structures, even if it also contains a small area of the brain that may be variable. The histograms obtained on this box vary from one image to the other, so we can distinguish three types of histograms, figure 2. The first type (2a) is formed by three easy to separate peaks that respectively correspond to the CSF, the GM and the WM. In some other histograms (2b) the gray matter and the white matter do not have a clear boundary, so there is one large peak accumulating the two classes. The third type (2c) is characterized by a really small amount of CSF, so that there are no clear peaks.

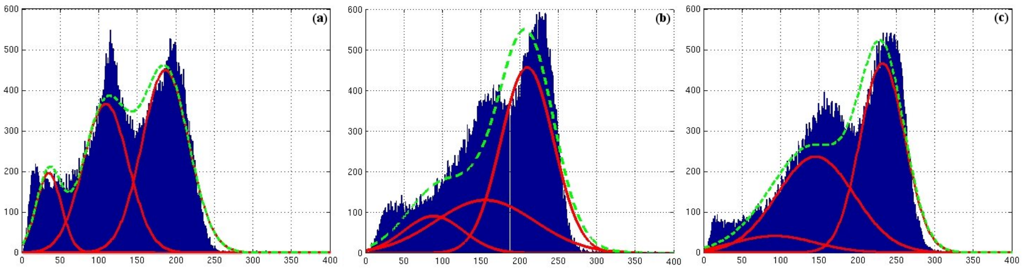

Figure 2: Gaussian mixture parameters estimation on the 41-by-41-by-41 voxel box around AC. In red are the three estimated Gaussian distributions, the dot green line represents the cumulative distribution. (a) corresponds to a histogram with well separated peaks, (b) corresponds to a case with unclear boundary between GM and WM, (c) corresponds to a pathologic case: there are almost no CSF peaks.

The result of the EM algorithm varies a lot from one image to another, depending on the histogram. The (2a) type gives good results and is modeled by three separated Gaussian distributions. The (2b) histograms have

a correct Gaussian for the WM, but the CSF and GM are modeled with large variance distributions, that fit badly to the observations. The (2c) histograms correspond to the worse results. The estimated parameters lead to a higher probability of GM than the probability of CSF even in the lowest intensities. Some common effects are also observed. The peaks of GM and WM, sometimes the CSF's too, are really sharp, that lets think that the Gaussian distribution is not accurate to model the gray levels distribution. We are more concerned about the behavior of the CSF peak. It is shifted to the right because of the distribution truncation in the negative values. The concern is also for the tails of the distributions. Because of its larger variance, in case (2b), the gray matter will be more likely in the very high intensities than the white. We decided not to take care on this aspect first, considering that these voxels are rare enough to be neglected.
For the training set, we kept 14 images whose graylevels distribution is correctly modeled.

3.3 Improving the model of the gray levels

There are different ways to improve the model accuracy. The problem of modelling the gray levels has been studied as a segmentation problem by many researchers. The reader is refereed to a good tutorial on MRI segmentation, [9]
One of the options is to model separately the pixels that contain more than one matter, either GM and CSF or GM and WM.These sets are called the partial volumes. This approach is important when one wants to estimate the volume of a structure with accuracy [10]. The simplest way to these voxels in account is to introduce two additional Gaussian distributions to obtain a five component mixture. This mixture fits better, but there are still some problems in the distribution choice and in the EM results.
If one has an hand-segmentation of the image or of a part of it, then a more complicated model can be used. This time the number of Gaussian components is not fixed at the beginning. Each matter is itself subdivided as a Gaussian distribution mixture without a priori information about the number of components [11]. Even if there are some solutions to the number of components determination problem that are proved to be efficient asymptotically, the problem is still debated in the finite situations.
Adding constraints on the parameters during the estimation, so that the gray for example could not be modeled with a large variance distribution is an-

other way to improve the accuracy of the graylevel model. For example it is possible to introduce a prior distribution on the parameters.
Finally it might also be interesting to try a mixture of different type of distributions. We may lost the possibility to write analytic solutions for the EM but it may allow us to fit better the observed histograms.

4 The translation model

Our approach uses the gray levels of a subset of voxels $A \subset S$ to predict the location of the landmark Y. X_s and Z_s are respectively the gray level intensity and the matter at the voxel s. We would like to express the expectation of the HoH location given the observations at the voxels of A, $E[Y|X_A]$.
To express this expectation we are making some assumptions.

- **1st assumption: the prior distribution**. The observed location of the landmark in the training set is contained in a 7-by-8-by-8 voxel box. Since it would introduce a large error to underestimate the initial variance, we decided to enlarge the initial box. We assumed that the prior is a Uniform distribution $p(y)$ on a 11-by-12-by-12 voxel box V. The problem is then to find one voxel in this set.

$$p(y) = \frac{1}{|V|} \tag{3}$$

- **2nd assumption: conditional independence of the intensities**. We assume that the intensities X_s are independent given the location of the HoH Y. It simplifies a lot the expression of the expectation.

$$P(X_A|Y) = \prod_{s \in A} P(X_s|Y) \tag{4}$$

- **3rd assumption: the translation model**. We assume that the matter j observed in the image i at the voxel s depends only on the translation between s and the HoH location y.

$$P_i(Z_s = j|Y = y) = \pi_{s-y}(j) \tag{5}$$

- **4th assumption: conditional independence of the intensity and the location of the HoH**. We assume that given the matter j observed at the voxel s the intensity x_s and the location of the HoH y are indenpendent.

$$P_i(X_s = x_s|Z_s = j, Y = y) = P_i(X_s = x_s|Z_s = j) \tag{6}$$

- **5th assumption: Gaussian distributions**. We assume that the intensity X_s at the voxel s given the matter Z_s follows a Gaussian distribution. We arrive to the model of Gaussian mixture introduced in the section 3.

$$P_i(X_s = x_s|Z_s = j) = g_{ij}(x_s) \tag{7}$$

The third and the fourth assumptions allow us to distinguish the photometry and the geometry of the image. The photometry is modeled as a mixture of Gaussian distributions and the geometry with the translation model. These two parts of the model can be modified independently. The expression of the expectation is now much easier to write.
For the image i,

$$\begin{aligned} E[Y|X_A] &= \sum_{y \in S} y \frac{P_i(X_A|Y)p(y)}{\sum_{y' \in S} P_i(X_A|Y)p(y')} \\ \text{by (3)} &= \sum_{y \in V} y \frac{P_i(X_A|Y)}{\sum_{y' \in V} P_i(X_A|Y)} \end{aligned}$$

And we can write:

$$\begin{aligned} P_i(X_A|Y) &= \prod_{s \in A} P_i(X_s = x_s|Y = y) \text{ by (4)} \\ \text{by the Bayes formula} &= \prod_{s \in A} \sum_{j=1}^{3} P_i(X_s = x_s|Z_s = j, Y = y) P_i(X_s = x_s|Z_s = j) \\ \text{by (5) and (6)} &= \prod_{s \in A} \sum_{j=1}^{3} P_i(X_s = x_s|Z_s = j) \pi_{s-y}(j) \\ \text{by (7)} &= \prod_{s \in A} \sum_{j=1}^{3} \pi_{s-y}(j) g_{ij}(x_s) \end{aligned}$$

5 The template

The proportions of the matters $\pi_{s-y}(j)$ have to be estimated at each voxel. The resultant 3-channel random image, learned on the training set, is called the template.

The first step of its computation is to superimpose all the images of the training set, centered on the landmark location y_0. So there are as many points as images in the training set to estimate by an EM algorithm the proportions in each location. With j the matter, i the image of the training set and y the location of the HoH, the log-likelihood to maximize over the $\pi_{s-y}(j)$ can be written:

$$\sum_i \log \sum_{j=1}^{3} \frac{\pi_{s-y}(j)}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{|x_s^{(i)} - \mu_{ij}|^2}{2\sigma_{ij}^2}\right)$$

In the template, figure 3, each point is defined as a translation to origin y_0, the head of the hippocampus. The different channels of the RGB image represent the probability to observe at a given translation from y_0 the three matters. There are three types of points in the template. Some are mixed corresponding to a very variable part from an image to another, the rest has pure color and is either located in a large volume of same color or on an edge. When the color is pure it means that in most of the image this point is in the corresponding matter.

It is noticeable that the closer to y_0 a structure is, the sharper and more precise its edges are. However, small structures around the hippocampus disappear. For example the white line that is used by neuroscientists as a boundary between the amygdala and the hippocampus, is particularly thin, so the variability between the images erases it.

When more images are used for the computation, some structures are blurred, and the template is smoothed even if there are still really sharp boundaries. This computation gives an image which is as large as the original, but obviously many voxels have either no links with the location of the head of the hippocampus or bring redundant information. Therefore, it is necessary to select carefully the relevant locations in order to reduce the computing time without loosing in accuracy.

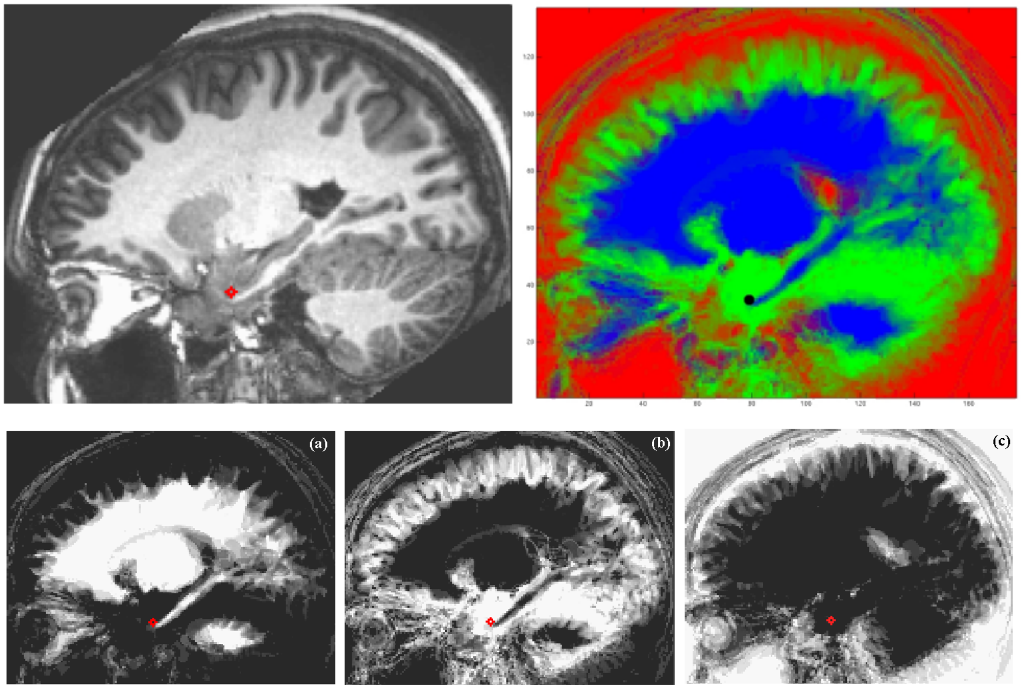

Figure 3: The template, sagittal slice containing the landmark, the red or black dot. The top left image is an example of raw data, the top right is the RGB representation of the template (Red = $\pi_{s-y_0}(CSF)$, Green = $\pi_{s-y_0}(GM)$, Blue = $\pi_{s-y_0}(WM)$). The second line (a, b, c) represents the separated channels (from left to right: CSF, GM, WM), the high intensity corresponds to a high probability of observing the matter at this location.

6 Selection of the informative voxels

The aim of the selection step is to drastically reduce the number of voxels necessary to compute the expectation. The challenge is to discrimine the informative voxels among a lot of not informative ones.

6.1 Ranking the voxels

Sorting the voxels by their amount of information about the HoH location is a natural idea. In a first method, which we called the ranking method (R), the voxels which bring more information than a fixed threshold are picked. We need to define a measure of the information.

6.1.1 Expectation of the Conditional Variance

The problem allow us to use the Euclidian distance and the variance. We choose to rank the voxels by the expectation of the conditional variance.

$$E_{Y,X_s}\|Y - E_Y(Y|X_s)\|_2^2 \tag{8}$$

(8) measures the expected variance reduction of the head location Y, due to the knowledge of the intensity observed at the voxel s. However the computation of this value is really heavy, therefore the expectation of the conditional variance, given the matter at each voxel s is preferred.

$$E_{Y,Z_s}\|Y - E_Y(Y|Z_s)\|_2^2 \tag{9}$$

and one has

$$E_{Y,Z_s}\|Y - E_Y(Y|Z_s)\|_2^2 \leq E_Y\|Y - E_Y(Y)\|_2^2$$

It is noticeable that (9) is equivalent to (8), if for each image i, all the matters have distinct supports. With the usual notations,

$$\begin{aligned}
&\text{If } I = \bigcup_{j=1}^{3} I_j \quad \text{with} \quad \forall \{j, j'\} \in \{1,2,3\}^2, I_j \cap I_{j'} = \emptyset, \\
&\text{then } \forall x_{s_{j'}} \in I_{j'}, \\
P(Y = y|X_s = x_{s_{j'}}) &= \sum_{j=1}^{3} P(Y = y|Z_s = j)P(Z_s = j|X_s = x_{s_{j'}}) \\
&= \delta_{jj'}P(Y = y|Z_s = j) \\
\text{also } E_Y[Y|X_s = x_{s_j}] &= E_Y[Y|Z_s = j].
\end{aligned}$$

Even if we are not in this case, we can assume that (9) is close to (8).
The lower the conditional variance is, the more informative the voxel is. The

expectation (9) can be developed,

$$
\begin{aligned}
&Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \text{ with } y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \text{ and } \forall k \in \{1, 2, 3\}, y_k \in V_k \\
&E_{Y,Z_s} \|Y - E_Y(Y|Z_s = j)\|_2^2 \\
&= \sum_{k=1}^{3} E_{Y_k,Z_s} \|Y_k - E_{Y_k}(Y_k|Z_s = j)\|_2^2 \\
&= \sum_{k=1}^{3} \sum_{j=1}^{3} \sum_{y_k \in V_k} \left[(y - E_{Y_k}(Y_k|Z_s = j))^2 P(Y_k|Z_s = j) P(Z_s = j) \right] \\
&= \sum_{k=1}^{3} \sum_{j=1}^{3} \sum_{y_k \in V_k} \left[\left(y_k - \sum_{y'_k \in V_k} y'_k \frac{\pi_{s-y'_k}(j)}{\sum_{y''_k \in V_k} \pi_{s-y''_k}(j)} \right)^2 \frac{\pi_{s-y_k}(j)}{\sum_{y''_k \in V_k} \pi_{s-y''_k}(j)} \right] \\
&= \sum_{k=1}^{3} \left[\sum_{j=1}^{3} \sum_{y_k \in V_k} y_k^2 \frac{\pi_{s-y_k}(j)}{\sum_{y''_k \in V_k} \pi_{s-y''_k}(j)} - \sum_{j=1}^{3} \left(\sum_{y_k \in V_k} y_k \frac{\pi_{s-y_k}(j)}{\sum_{y''_k \in V_k} \pi_{s-y''_k}(j)} \right)^2 \right]
\end{aligned}
$$

The last expression can be computed by convolutions so that the total computation of the variance takes around thirty minutes.

The conditional expectation measures the information which is brought by a voxel. The histogram, figure 4, shows that there is only few information in the knowledge of the matter observed at one given voxel. The best voxels give a conditional standard deviation around 5.10 when the initial value is 5.84. In addition most of the voxels have really little information about the landmark location, it confirms the importance of the selection.

6.1.2 Simulation

In order to understand where the informative voxels are located in the template, simulated images were generated using the same gray levels parameters and hippocampus location in the image than in the training images. Each image is composed of a large sphere centered on a virtual hippocampal head, and a second smaller sphere. The intensity of the voxels contained in the background, the large sphere and the smaller sphere was simulated using the parameters of respectively the CSF, the Gray Matter or the White Matter. We used the usual algorithm to compute the template and the information

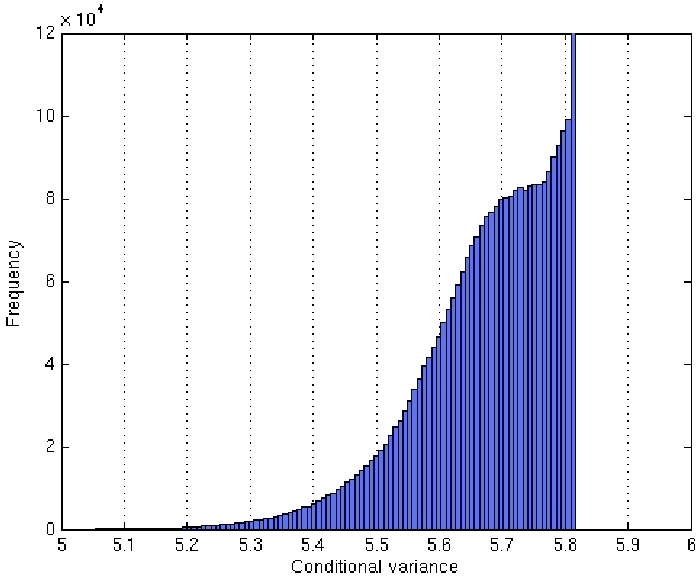

Figure 4: Histogram of the information contained in the template. We represent here the squared root of the expected variance, i.e. the conditional standard deviation.

map (map of the expected variance reduction). The experiment was repeated three times, changing the relation between the spheres and the location of the landmark in the simulated images, figure 5.

The points with lowest variance or standard deviation, which are represented in blue in the information map, are the most informative. The first experiment shows that the points contained in a large constant color volume are useless for the prediction. The most informative points correspond to a symmetric tube along the template edges which are connected to the head of hippocampus location, i.e. the sharp edges of the template. When an edge is not linked to the HoH location, like the small sphere in the second simulation or the large sphere in the last simulation, the corresponding template edge is blurred and neglected by the edge detector we built previously. This simulation mainly shows that the expected variance reduction is acting as a sharp edge detector on the template.

6.1.3 Applying the edge detector on MR Images

Applying the edge detector to MR Images produces an information map, figure (6), whose correspondence with the anatomical structures are difficult

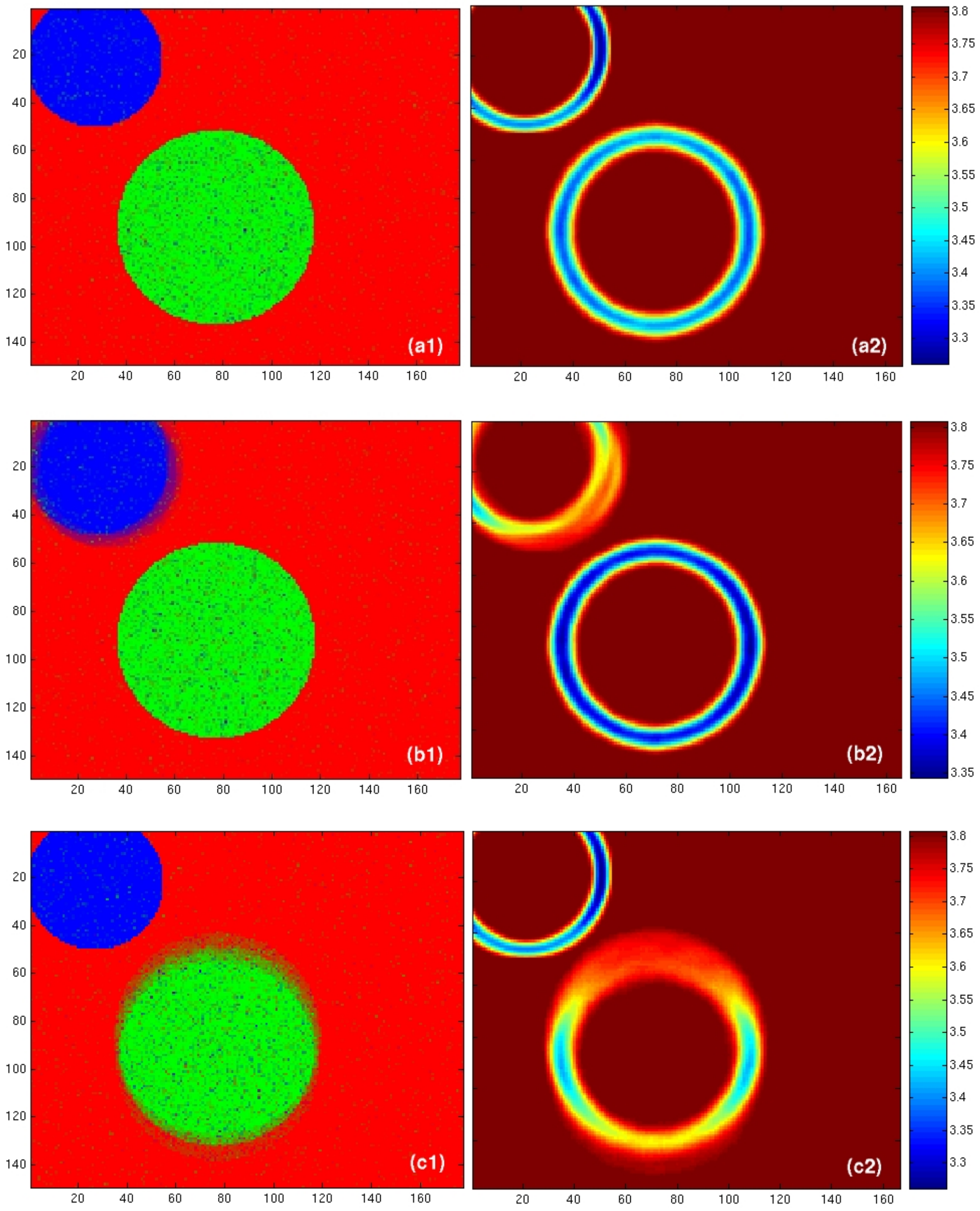

Figure 5: Simulated images genrerated with the parameters of the training set images and the location of HoH in the training set. The left column represents the template obtained in the 3 experiments and the right column contains the resultant information map. The bar represents the conditional standard deviation scale. In the experiment (a), both the small and the large sphere are rigidly linked to the HoH location. In the experiment (b), the smaller sphere is totally independent. In the experiment (c), the large sphere is linked to the location of the HoH, but with a little noise.

to find. The informative parts of the brain are represented in blue, since the

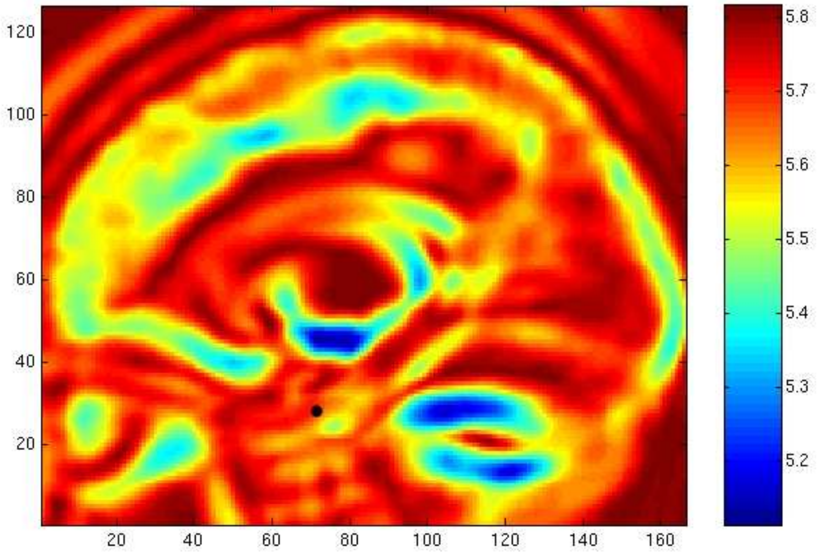

Figure 6: Sagittal slice of the information map. the black dot represents the HoH location and the color bar the conditional standard deviation scale. Blue corresponds to low values and red to high values.

conditional variance is low. The front and back edges of the cerebrum seem to be interesting, whereas the top part is less informative, probably because of the circumvolutions variability. The most informative part is close to the landmark (the black dot) and corresponds to the boundary of white matter and gray matter. In addition, the cerebellum edges response to the edge detector, even if there is no easy anatomical explanations.
We have some informative points, but it remains the question of the number of points to use. In the easiest method we threshold the ranking and keep all the points below the limit. Even if the size reduction is already huge, most of the selected voxels are redundant, so that we believe it is still possible to remove lots of them. However using redundant variables may have some advantages like protecting from overfitting, [12].

6.2 Avoiding redundancy thanks to spatial constraints

We believe that close voxels contain similar information, and that it is possible to select only few points and to reach a comparable accuracy. Spatial constraints are added to the selection process, which is then called (R+SC). The points are now chosen if and only if they are at least $5mm$ away from all the previously selected voxels. To compare the selected points in the two methods we ran this second algorithm on the sets of selected points, see the table 6.2. The effective of points drop off, because most of the them were spatially related. On the one side, spatial constraints partially suppress the redundancy but on the other side, the set of selected points becomes sensitive to overfitting. This selection process has a complexity of $O(N)$, with N the initial number of points.

Method	Nb of selected points selected					
R	186	2230	5846	13,902	30,348	62,178
R + SC	6	40	91	204	402	724

Table 1: Number of voxels selected with the Ranking method (R) or with the Spatial Constraints method (R+SC). The set of R corresponds exactly to all the voxels below the threshold, and lot of voxels are removed from this set thanks to the spatial constraints. The different thresholds are 5.10, 5.20, 5.25, 5.30, 5.35, 5.40.

6.3 Voxel Selection by Conditional Variance

6.3.1 Principle

Another way to pick the best voxels would be to select the set $A = \{t(1), ..., t(K)\}$ of voxels that minimizes $E_{Y,Z_s}\|Y - E_Y(Y|Z_{t(1)}, ..., Z_{t(K)})\|_2^2$. But this expression required to estimate 2^{K+1} probabilities with a reduced size of sample. In (Fleuret, 2003) [13], the author proposes a new approach based on conditional mutual information to solve an equivalent selection problem. He deals with the trade-off between the information given by a feature and the independence to the already picked features. We adapted the method to the conditional variance and called it R+CV. The selection follows the iterative scheme:

$$t(1) = \arg\min_s E\|Y - E(Y|Z_s)\|_2^2 \tag{10}$$

$$\forall k, 2 \leq k < K, t(k+1) = \arg\min_s \left\{ \max_{l \leq k} E\|Y - E(Y|Z_s, Z_{t(l)})\|_2^2 \right\} \tag{11}$$

To interpret this minmax criterion, one can write it:
Z_t an already selected voxel, Z_s a remaining voxel, A_k the subset of voxels selected after k iterations:

$$\begin{gathered}\forall s \in S, \forall t \in A_k, E\|Y - E(Y|Z_s, Z_t)\|_2^2 \\ = E\|Y - E(Y|Z_s)\|_2^2 - E\|E(Y|Z_s) - E(Y|Z_s, Z_t)\|_2^2\end{gathered}$$

Because

$$\begin{aligned} E_Y E_{Z_s,Z_t} & \left[(Y - E(Y|Z_s, Z_t)) \cdot \underbrace{(E(Y|Z_s) - E(Y|Z_s, Z_t))}_{\phi(Z_s,Z_t)} |Z_s, Z_t \right] \\ & = E_Y \left[\phi(Z_s, Z_t) \underbrace{E_{Z_s,Z_t}[(Y - E(Y|Z_s, Z_t))|Z_s, Z_t]}_{0} \right] = 0 \end{aligned}$$

Finally (10) can be written as

$$\min_s \left\{ E\|Y - E(Y|Z_s)\|_2^2 - \min_t E\|E(Y|Z_s) - E(Y|Z_s, Z_t)\|_2^2 \right\} \tag{12}$$

The first part of (12) favors informative voxels, whereas the second favors voxels different from the previous picked ones. For two informative voxels s and s', if s is equivalent to an already picked voxel, then the second term is zero. If s' is different then the second part add a negative term, that makes the total smaller; so s' is chosen.

6.3.2 Implementation

The implementation of this selection is more complicated. We obtained by personal communication from François Fleuret, an efficient algorithm to compute it. Whereas the naive implementation has a complexity of $O(K \times S \times T)$ with K the number of features to select, S the total number of features, T the number of already picked features at a given step, he introduces a huge reduction in variance calculation. During the iteration the quantity whose minimum is looking for, can only increase when updated. Consequently it is useless to update during this iteration all the scores that are already above the temporary selected one. Adapted to the variance, it gives the algorithm 1.

Algorithm 1: Selection algorithm using the conditional variance

```
initialization;
for $s = 1..S$ do
    $ps(s) \leftarrow E\|Y - E(Y|Z_s)\|_2^2$;
    $m(s) \leftarrow 0$;
end
$m(1) \leftarrow \arg\min_s(ps)$;
for $s = 1..S$ do
    $ps(s) \leftarrow E\|Y - E(Y|Z_s, Z_{m(1)})\|_2^2$;
    $m(s) = 1$;
end
main iterations;
for $t$=$2$:$K$ do
    $s* = E\|Y - E(Y)\|_2^2$;
    for $s = 1 : S$ do
        while $ps(s) < s*$ $and$ $m(s) < t - 1$ do
            $m(s) = m(s) + 1$;
            $ps(s) = max\left(ps(n), E\|Y - E(Y|Z_s, Z_{nu(m(s))})\|_2^2\right)$;
        end
        if $ps(s) < s*$ then
            $s* = ps(s)$;
            $nu(t) = s$;
        end
    end
end
```

There are 2 initialization loops on all the features because of the properties of the conditional variance. The shortcut described above is applying from iteration 3 up. Indeed the second loop is computing the conditional variance, but this variance is always inferior to the initialization, so that to find the minimum we have to calculate all the scores.

$$\forall s \in S, E\|Y - E(Y|Z_s, Z_{m(1)})\|_2^2 \leq E\|Y - E(Y|Z_s)\|_2^2 \qquad (13)$$

It takes around 60 minutes to select 700 points among 60,000, whereas the time-consuming computation of (12).

7 Results

The algorithm is tested on the training set and then on new images. For numerical reasons we are not able to compute the expectation $E[Y|X_A]$, the product of probabilities is rounded to zero when the cardinal of A is too large. For the following results, the prediction of the HoH location is given by the point of V that maximizes the probability $P(X_A|Y)$. All the selection methods are tested with the online algorithm. The test is computed for the R method with 186, 2230 and 6000 points; it corresponds to all the points that have an expected standard deviation below respectively 5.1, 5.20 and 5.25. The second test is performed using the R+SC method with up to 750 points. Finally the R+CV method is tested using the same effective of points than in the second experiment.
The prediction error is the Euclidian distance between the expert's landmark and the predicted location, figure 7 and figure 8.

7.1 Testing on the training set

Whichever selection method is chosen, the mean error decreases when the number of voxels increases. However, the amount of points necessary to obtain a stable accuracy depends on the selection method.
For the ranking method, it is already stabilized with 2230 voxels. When spatial constraints or constraints on the conditional variance are added, this threshold drop off to 750 voxels.
We compare the results of the different methods. the best accuracy achieved is $1.98mm$ ($\sigma = 0.82$). With only 750 voxels the second and the third methods achieve an accuracy of respectively $1.53mm$ ($\sigma = 1.37$) and $2.96mm$

($\sigma = 1.95$). We obtain the smallest variance with the first selection and the best mean error with the second. As for the computation time, it takes several minutes to predict a location thanks to the first method and only several seconds with the second and third one. We believe that the ranking selection is a lot more robust to overfitting; it could explain its low mean error and its low variance. The less robust is probably the Conditional Variance selection where redundancy is almost totally removed. The results seem to confirm this lack of robustness since its variance and mean error are the highest. Because of the computation time and the best achieved accuracy, we prefer the selection of voxels with spatial constraints.

If we compare the results with the human performance, see subsection 2.3, it is clear that the algorithm is learning the location of the HoH, since it almost reaches the expert's accuracy ($1.53mm$, $\sigma = 1.37$; $0.93mm$, $\sigma = 0.96$). However we have to keep in mind that the final error mean is to add to the initial error of the expert in the training set.

Surprisingly, the algorithm is unable to predict the location of the HoH for one of the images of the training set. Looking at the singular image, we notice that it is an image with almost pathological gray levels behavior. We will verify in the subsection 7.2 that this gray level modelling may have an influence on the prediction quality.

7.2 Testing on a new set of images

We firstly use 5 images whose gray levels are correctly modeled. As in case of the training set, the mean error decreases when voxels are added. The thresholds of stability seem to be the same even if the final results are not as good as the results previously reached. The best accuracy achieved is a mean error of $3.91mm$ with $\sigma = 2.08$ with the ranking method. Adding spatial constraints modifies a little the results with a mean error of $4.14mm$ and $\sigma = 1.75$. The results are worse with the third selection algorithm, achieving a mean error of $5.32mm$ ($\sigma = 2.55$).

The more voxels are used, the best the results are. The method with constraints on the conditional variance is less robust than the other: both the resultant mean error and variance are high. Even with the ranking selection that is using up to 6000 points to do the prediction, the results seem to be quite not good, even if they are comparable to student's results. This high generalization error could be explained by the small size of the training set. At each location 2 parameters are estimated, plus the selection of only few

voxels. It may explain the overfitting of the training set. Using more images may help to learn more variability and to decrease the generalization error.
Secondly we made the same prediction on four images whose graylevels are badly modeled. They are from the third group described in the figure 2. Because the probability of the CSF is always low, the distinction between the CSF and the GM is not good. Therefore we thought that the model would not be able to predict correctly the location of the HoH. The test on this set of images do not confirm this intuition. Indeed the results are similar to those obtained on the other testing images. (R: $3.50mm, \sigma = 1.93$; R+SC: $4.14mm, \sigma = 2.09$). To explain that, we made the hypothesis that the algorithm is using more voxels on boundaries between GM and WM than on boundaries between CSF and GM. We will need to look more carefully at the location of the selected voxels, to validate or invalidate this hypothesis.

8 Conclusion and Future work

The method we proposed is applicable at any location of the brain and predicts a location in only few seconds. The first improvements will be to use more images for the template estimation and to overcome the numerical limitations to obtain the expectation and the covariance matrix. We will then be able to know how the candidates voxels are distributed. We also hope that this complementary information about the prediction error will help us to understand why sometimes the algorithm is unable to predict the good location.

Since in the model the photometry is separated from the geometry of the image, we can imagine improvements independently for these two questions. For the photometry, the priority is to improve the model of the gray levels. It could be helpful to add constraints in the EM and/or to modify the distributions in the mixture to fit the peaks of the histogram. We could also consider a new volume, built by neuroscientists that would decrease the noise in the normalization box.

As for the geometry, the model deals only with translation between the landmark and the observed voxels. We will try to take in account other transformations or deformations in the model. Up to now we are only using the gray levels of the image. We will use other features combined with the gray levels.

Since several landmarks have to be detected, we can try to work pairwise or even with more landmarks at the same time. Indeed we believe that detecting simultaneously the head and the tail of the Hippocampus for example may be more efficient than detecting them independently.

References

[1] Thomas Hartkens, Karl Rohr, and H. Siegfried Stiehl. Evaluation of 3d operators for the detection of anatomical point landmarks in mr and ct images. *Computer Vision and Image Understanding*, (86):118–136, 2002.

[2] J.Talairach and P. Tournoux. *Co-planar stereotaxic Atlas of the Human Brain.* Thieme Medical Publishers, 1988.

[3] Laurent Vérard, Pascal Allain, Jean-Claude Baron Jean-Marcel Travère, and Daniel Bloyet. Fully automatic identification of ac and pc landmarks on brain using scene analysis. *IEEE Transactios on medical imaging*, 16(5), October 1997.

[4] Yejt Han and Hyun Wook Park. Automatic brain mr image registration based on talairach reference system. In *Proceedings 2003 International Conference on Image Processing*, volume I, pages 1097–1100, 2003.

[5] Lundbeck Institute. The brain explorer. *www.brainexplorer.org.*

[6] Henri M. Duvernoy. *The Human Hippocampus: Functional Anatomy, Vascularization and Serial Sections with MRI.* Springer, 2nd edition, 1998.

[7] R. Cox. Afni software package. *www.afni.nimh.nih.gov*, 2004.

[8] Paulus Petrus Bernardus Eggermont and Vincent N. LaRiccia. volume I: Density Estimation, chapter 3.

[9] Koen Van Leemput. Unifying statistical classification and geometrical models for powerful automated segmentation: Techniques, applications and validations. In *Tutorial to Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2003.

[10] J.T. Ratnanather, K.N. Botteron, T. Nishino, A.B. Massie, R.M. Lal, S.G. Patel, S.Peddi, R.D. Todd, and M.I. Miller. Validationg cortical surface analysis of medial prefrontal cortex. *NeuroImage*, 14:1058–1069, 2001.

[11] Carey E. Priebe, Michael I. Miller, and J. Tilak Ratnanather. Segmenting magnetic resonance images via hierarchical mixture modelling. *submitted*, 2003.

[12] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, (3):1157–1182, 2003.

[13] François Fleuret. Binary feature selection with conditional mutual information. *Rapport Technique INRIA*, (4941), 2003.

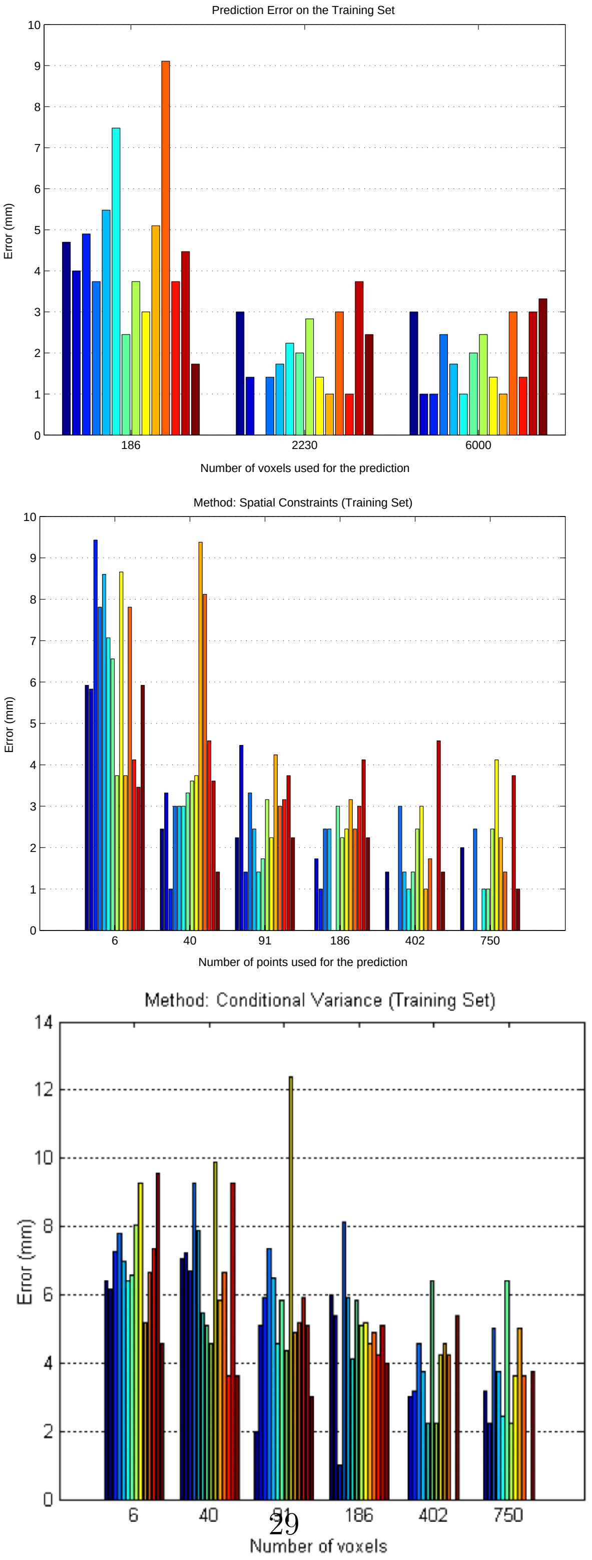

Figure 7: From top to bottom the graphics represent the prediction error depending on the number of voxels used in the online algorithm with respectively the R, R+SC and R+CV method. A bar corresponds to a single image of the training set.

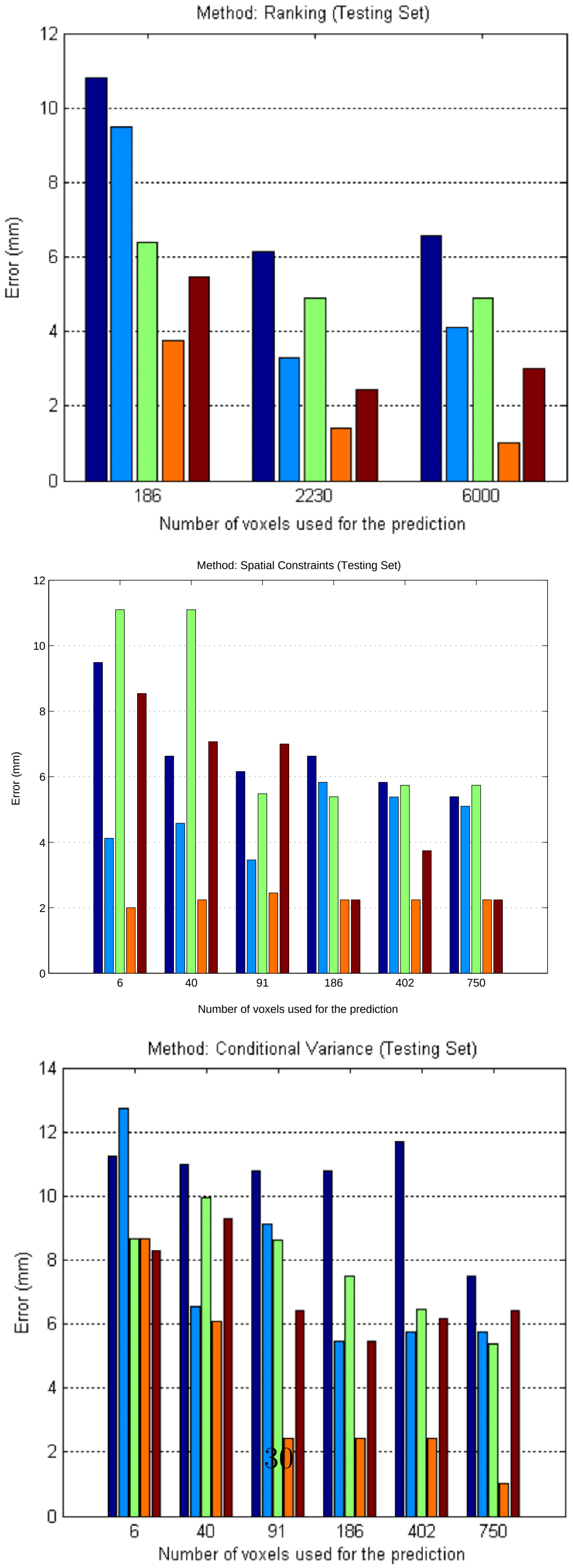

Figure 8: From top to bottom the graphics represent the prediction error depending on the number of voxels used in the online algorithm with respectively the R, R+SC and R+CV method. A bar corresponds to a single image of the testing set.