Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Deliverable Responsible	Dipartimento di Informatica Università degli studi di Verona Strada le Grazie, 15 – 37134 Verona (Italy) http://www.di.univr.it/
Project web site	http://simbad-fp7.eu
EC project officer	Teresa De Martino
Document title	WP7: Analysis of brain magnetic resonance (MR) scans for the diagnosis of mental illness final work package report
Deliverable n.	D7.2
Document type	Final Report
Dissemination level	Public
Contractual date of delivery	M 36
Project reference number	213250
Status & version	Definitive version
Work package, Deliverable responsible	WP7, UNIVR
Author(s)	Aydın Ulaş, Umberto Castellani, Manuele Bicego, Vittorio Murino
Additional contributor(s)	Marco Cristani, Alessandro Perina, Dong Seon Cheng, Pasquale Mirtuono

Contents

WP7: Final Deliverable

September 18, 2011

Abstract

This document results from work carried out in the context of the Work package 7 - Analysis of brain magnetic resonance (MR) scans for the diagnosis of mental illness - of the SIMBAD project. This is the final report on results with discussions about successful approaches and issues that remain still open. Overall, this WP addresses the problem of schizophrenia detection by analyzing magnetic resonance imaging (MRI). In general, mental illness like schizophrenia or bipolar disorders are traditionally diagnosed by self-reports and behavioral observations. A new trend in neuroanatomical research consists of using MRI images to find possible connections between cognitive impairments and neuro-physiological abnormalities. Indeed, brain imaging techniques are appealing to provide a non-invasive diagnostic tool for mass analyses and early diagnoses. The problem is challenging due to the heterogeneous behavior of the disease and up to now, although the literature is large in this field, there is not a consolidated framework to deal with it. In this context, advanced pattern recognition and machine learning techniques can be useful to improve the automatization of the involved procedures and the characterization of mental illnesses with specific and detectable brain abnormalities. In this project we have exploited similarity-based pattern recognition techniques to further improve brain classification problem by employing the algorithms developed in other WPs.

1 Introduction

Brain analysis techniques using Magnetic Resonance Imaging (MRI) are playing an increasingly important role in understanding pathological structural alterations of the brain (Giuliani et al., 2005; Shenton et al., 2001). The ultimate goal is to identify structural brain abnormalities by comparing normal subjects with patients affected by a certain disease.

In this project, we focus on schizophrenia. Schizophrenia is a heterogeneous psychiatric disorder characterized by several symptoms such as hallucinations, delusions, cognitive and thought disorders (Bellani and Brambilla, 2008). Although genetic and environmental factors play a role in the disorder its etiology remains unknown and substantial body of research has demonstrated numerous structural and functional brain abnormalities in patients with both chronic and acute forms of the disorder (Shenton et al., 2001; Rujescu and Collier, 2009).

Our main contribution in this WP is to deal with schizophrenia detection as a binary classification problem: we have to distinguish between normal subjects and patients affected

by schizophrenia (Davatzikos, 2004). To this aim we have employed advanced pattern recognition techniques by exploiting the capability of similarity-based methods developed in other WPs to this problem.

We highlight that the problem of schizophrenia detection is very complex since the symptoms of the disease are different and related to different properties of the brain. Thus, although the literature has shown a large amount of promising methodological procedures to address this disease, up to now a consolidate framework is not available.

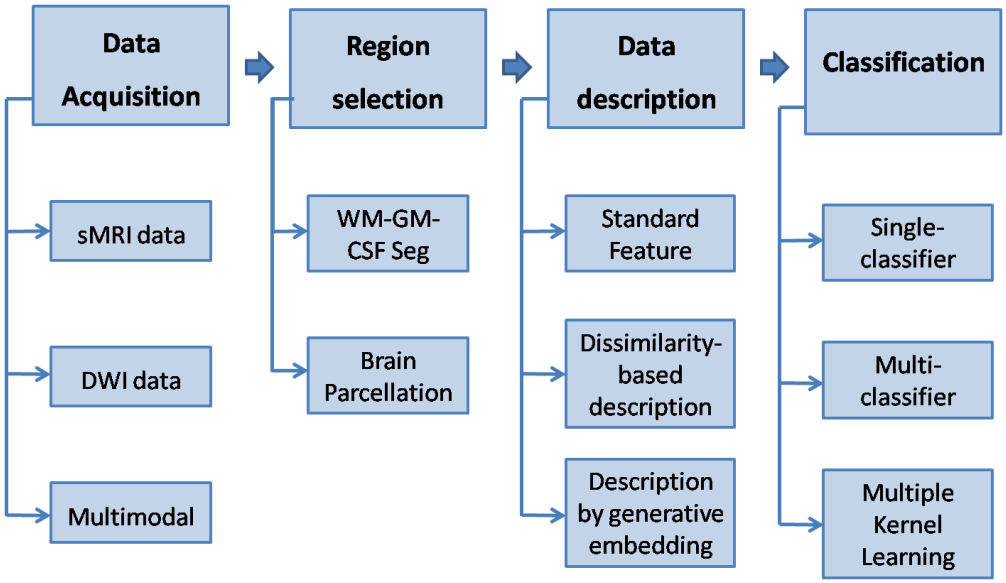

Figure 1: Overall scheme of the proposed working pipeline.

In this project we have exploited different approaches to address Schizophrenia detection. We have defined a general working pipeline composed by four main steps: i) data acquisition, ii) region selection, iii) data description, and iv) classification. Each step may be istantiated in different ways, each one having pros and cons. In this report, for each stage, we summarize the possible choices we adopted in this project. Figure 1 shows the proposed overall scheme of the working pipeline and the involved possibilities. In summary:

- ***Data acquisition*** regards the imaging technique employed to acquire data. Different acquisition modalities are encoding different brain information. In our project we used Structural MRI to deal with morphological properties, and DWI to evaluate functional aspects of the brain. Moreover, in order to integrate different source of information a multimodal approach is exploited.

- ***Region selection*** is necessary to focus the analysis on brain subparts. A common approach is to segment the whole brain among *White Matter* (WM), *Gray Matter* (GM), and *Cerebro-Spinal Fluid* (CSF). Another approach consists of extracting one or more Regions of Interest (ROIs) which are strictly related to the analyzed disease. The brain segmentation in ROIs is in general called *brain parcellation*.

- ***Data Description*** aims at extracting the most useful information for the involved task, in our case brain classification. The standard approach consists of using *features*.

According to the overall aim of this project we exploited the possibility to go beyond features. Indeed, we have investigated two paradigms, derived from research done in other WPs of the project: a dissimilarity-based description (linked to the dissimilarity-based representation scheme developed by Delft Group), and description by generative embedding (linked to our activities in the context of WP2). Note that these two new paradigms represent the core of the SIMBAD project.

- ***Classification*** is the last step of the proposed pipeline. As simplest approach a single classifier has been employed. In order to integrate different source of information at classification stage we exploited two paradigms: multi-classifier approach, and multiple kernel learning.

Roadmap. The report is organized as follows: In Section 2, we present the state of the art in schizophrenia detection. In Sections 3 and 4 we introduce data acquisition and region selection respectively. Then, data description phase is split in: standard features (Section 5), dissimilarity-based description (Section 6), and description by generative embedding (Section 7). We define our approaches of classification using ensembles and Multiple Kernel Learning in Section 8. We explain two case studies which utilize the working pipeline in Sections 9 and 10; and conclude in Section 11.

2 Related Work

Several works have been proposed for human brain classification in the context of schizophrenia research (Shenton et al., 2001). In the following we have organized the state of the art in i) *shape-based* techniques, and ii) *classification-based* techniques.

2.1 Shape-based techniques

Standard approaches are based on detecting morphological differences on certain brain regions, namely Region Of Interests (ROIs). Usually, the aim is the observation of volume variations (Pruessner et al., 2000; Shenton et al., 2001; Baiano et al., 2008). In general, ROI-based techniques require the manual tracing of brain subparts. In order to avoid such expensive procedure, Voxel Based Morphometry (VBM) techniques have been introduced (Ashburner and Friston, 2000; Kawasaki et al., 2007) for which the entire brain is transformed onto a template, namely the stereotaxic space. In this fashion a voxel-by-voxel correspondence is available for comparison purposes. In Kawasaki et al. (2007), a multivariate Voxel-Based Morphometry approach method is proposed to differentiate schizophrenic patients from normal controls. Inferences about the structural relevance of gray matter distribution are carried out on several brain sub-regions. In Voets et al. (2008), cortical changes in adolescent on-set schizophrenic patients are analyzed by combining Voxel-Based with Surface-Based Morphometry (SBM). A different approach consists of encoding the shape by a *global* region descriptor (Timoner et al., 2002; Gerig et al., 2001; Reuter et al., 2009). In Timoner et al. (2002) a new morphological descriptor is introduced by properly encoding both the displacement fields and the distance maps for amygdala and hippocampus. In Gerig

et al. (2001) a ROI-based morphometric analysis is introduced by defining spherical harmonics and 3D skeleton as shape descriptors. Improvement of such shape-descriptor-based approach with respect to classical volumetric techniques is shown experimentally. Although results are interesting, the method is not invariant to surface deformations and therefore it requires shapes registration and data resampling. This pre-processing is avoided in (Reuter et al., 2009), where the so called Shape-DNA signature has been introduced by taking the eigenvalues of the Laplace-Beltrami operator as region descriptor for both the external surface and the volume. Although *global* methods can be satisfying for some classification tasks, they do not provide information about the localization of the morphological anomalies. To this aim, *local* methods have been proposed. In (Toews et al., 2009) the so called *feature*-based morphometry (FBM) approach is introduced. Taking inspiration from feature-based techniques proposed in computer vision, FBM identifies a subset of features corresponding to anatomical brain structures that can be used as disease biomarkers.

2.2 Classification-based techniques

In order to improve the capability in distinguishing between healthy and non-healthy subjects, learning by example techniques (Duda et al., 2000) are applied (see for example, (Davatzikos, 2004)). Usually, geometric signatures extracted from the MRI data are used as feature vector for classification purpose (Yoon et al., 2007; Fan et al., 2007; Pohl and Sabuncu, 2009). In (Yoon et al., 2007) a support vector machine (SVM) has been employed to classify cortical thickness which has been measured by calculating the Euclidean distance between linked vertices on the inner and outer cortical surfaces. In (Fan et al., 2007) a new approach has been defined by combining deformation-based morphometry with SVM. In this fashion, multivariate relationships among various anatomical regions have been captured to characterize more effectively the group differences. Finally, in (Pohl and Sabuncu, 2009), a unified framework is proposed to combine advanced probabilistic registration techniques with SVM. The local spatial warps parameters are also used to identify the discriminative warp that best differentiates the two groups. It is worth to note that in most of the mentioned works, the involved classifier was a Support Vector Machine, but more general approaches are also proposed, i.e. Liu et al. (2004). Here, a set of image features which encode both general statistical properties and Law's texture features from the whole brain are analyzed. Such features are concatenated onto a very high dimensional vector which represents the input for a classic learning-by-example classification approach. Several classifiers are then evaluated such as decision trees or decision graphs. In Browne et al. (2008), the authors proposed a neural network to measure the relevance of thalamic subregions implicated in schizophrenia. The study is based on the metabolite N-acetylaspartate (NAA) using *in vivo* proton magnetic resonance spectroscopic imaging. The diffusion of water in the brain characterized by its apparent diffusion coefficient (ADC), which represents the mean diffusivity of water along all directions gives potential information about the size, orientation, and tortuosity of both intracellular and extracellular spaces, providing evidence of disruption when increased (Rovaris et al., 2002). DWI has been shown to be keen in exploring the microstructural organization of white matter therefore providing intriguing information on brain connectivity (Brambilla and Tansella, 2007; Tomasino et al., 2010).

3 Data acquisition

The data set involves a 124 subject database cared by the Research Unit on Brain Imaging and Neuropsychology (RUBIN) at the Department of Medicine and Public Health-Section of Psychiatry and Clinical Psychology of the University of Verona. The data set is composed of MRI brain scans of 64 patients recruited from the area of South Verona (i.e., 100,000 inhabitants) through the South Verona Psychiatric Case Register (Tansella and Burti, 2003). Additionally, 60 individuals without schizophrenia (control subjects) were also recruited. For details on the study population see Appendix B.

3.1 MRI data

MRI scans were acquired with a 1.5 T Magnetom Symphony Maestro Class Syngo MR 2002B (Siemens), and in total, it took about 19 minutes to complete an MR session. A standard head coil was used for radio frequency transmission and reception of the MR signal, and restraining foam pads were used to minimize head motion. T1-weighted images were first obtained to verify the participants head position and image quality (TR = 450 ms, TE = 14 ms, flip angle = 90°, FOV = 230 × 230, 18 slices, slice thickness = 5 mm, matrix size = 384 *times* 512, NEX = 2). Proton density (PD)/T2-weighted images were then acquired (TR = 2500 ms, TE = 24/121 ms, flip angle = 180°, FOV = 230 × 230, 20 slices, slice thickness = 5 mm, matrix size = 410 × 512, NEX = 2) according to an axial plane running parallel to the anterior-posterior (AC–PC) commissures to exclude focal lesions. Subsequently, a coronal 3-dimensional magnetization prepared rapid gradient echo (MP-RAGE) sequence was acquired (TR = 2060 ms, TE = 3.9 ms, flip angle = 15°, FOV = 176 × 235, slice thickness = 1.25 mm, matrix size = 270 × 512, inversion time = 1100) to obtain 144 images covering the entire brain. In Figure 2 we can see a slice of a subject acquired by using MRI.

3.2 DWI data

Diffusion-weighted imaging (DWI) investigates molecular water mobility within the local tissue environment, providing information on tissue microstructural integrity. The diffusion of water in the brain is characterized by its apparent diffusion coefficient (ADC), which represents the mean diffusivity of water along all directions (Taylor et al., 2004). Thus, ADC gives potential information about the size, orientation, and tortuosity of both intracellular and extracellular spaces, providing evidence of disruption when increased (Rovaris et al., 2002). ADC has also been used to explore regional grey matter microstructure, being higher in the case of potential neuron density alterations or volume deficit (Ray et al., 2006).

Diffusion weighted echoplanar images in the axial plane parallel to the AC-PC line (TR = 3200 ms, TE = 94 ms, FOV = 230 × 230, 20 slices, slice thickness = 5 mm with 1.5-mm gap, matrix size = 128 × 128, echo-train length = 5; these parameters were the same for $b = 0$, $b = 1000$, and the ADC maps). Specifically, three gradients were acquired in three orthogonal directions. ADC maps (denoted by D_{ADC}) were obtained from the diffusion images with $b = 1000$, according to the following equation:

$$-bD_{\text{ADC}} = \ln[A(b)/A(0)]$$

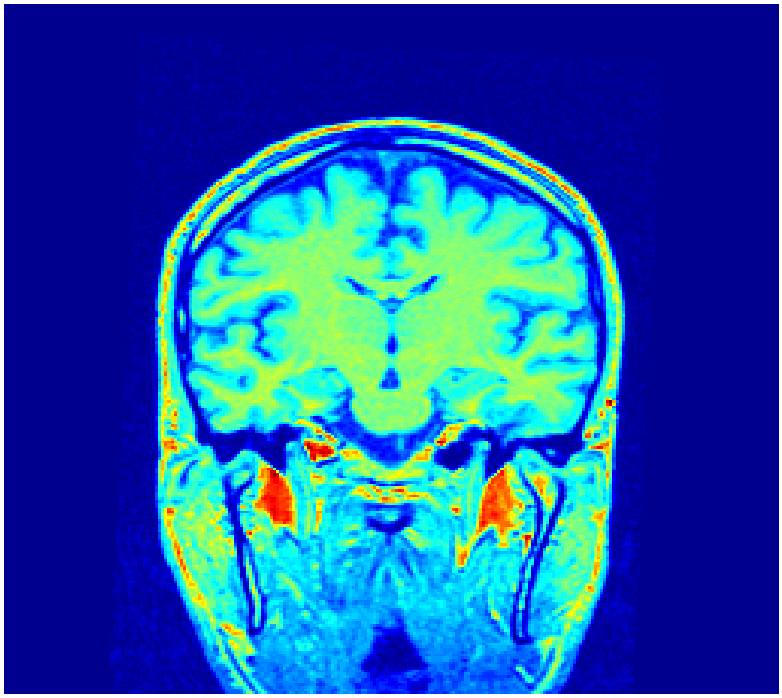

Figure 2: A slice acquired by 3D Morphological technique.

where $A(b)$ is the measured echo magnitude, b is the measure of diffusion weighting, and $A(0)$ is the echo magnitude without diffusion gradient applied. In Figure 3 we can see a slice of a subject acquired by using DWI.

3.3 Multimodal approach

Multimodal approach can be employed when different kinds of acquisition procedures are used for the same subject. As can be seen in Figures 2 and 3, while MRI images are more reliable, DWI resolution is very low and it's hard to segment ROIs from these DWI images. In order to integrate such data, a *co-registration* procedure is necessary.

The co-registration consists of matching high-resolution (also known as T1-w) and DWI images defined in different coordinate systems. Open source libraries of National Library of Medicine *Insight Segmentation and Registration Toolkit* are adopted for the co-registration procedure, while Tcl/Tk code and VTK open source libraries are chosen for the graphic interface. Digital Imaging and Communications in Medicine format (DICOM) tag parameters necessary for the co-registration are: Image Origin, Image Spacing, Patient Image Orientation, and Frame of Reference.

Assuming the same anatomy topology for different studies, a Mutual Information technique based on Mattes algorithm is applied. An in-house software for multimodal registration

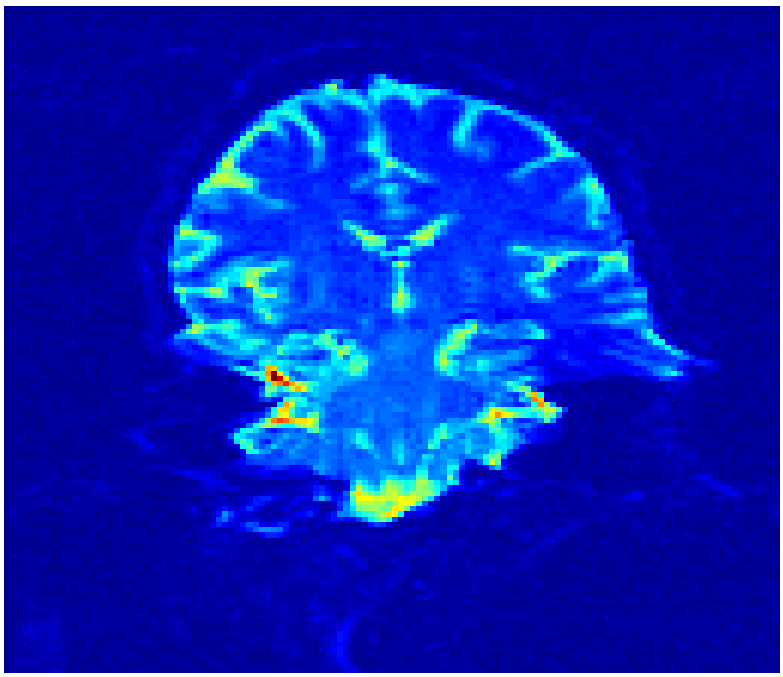

Figure 3: A slice acquired by Diffusion Weighting Imaging technique.

was developed. The program 3D Slicer[1], a free open source software for visualization and image computing, is employed for the graphic interface. The process was performed in several steps.

The source DWI study (*moving image*, see Figure 4) is aligned through a roto-translational matrix with the T1-w data (*fixed image*); the two studies are acquired in straight succession with the same MR unit without patient repositioning; the parameters related to algorithm implementation are automatically defined; then, by applying a multi-resolution pyramid we are able to reach a registration within eight iterations avoiding local minimal solution.

The results of the registration are visually inspected in a checkerboard, where each block alternately displayed data from each study, verifying alignment of anatomical landmarks (ventricles, etc.) for confirmation. In Figure 5, we can see registered DWI and sMRI images.

This procedure is needed because sMRI images have better resolution and the anatomy can better be seen for manual ROI segmentation. We use this procedure to extract ADC values for each of the ROIs instead of the whole image.

Once the co-registration is carried out a direct voxel-by-voxel comparison between the two data modalities become feasible and therefore any joint feature can be extracted.

[1]http://www.slicer.org/

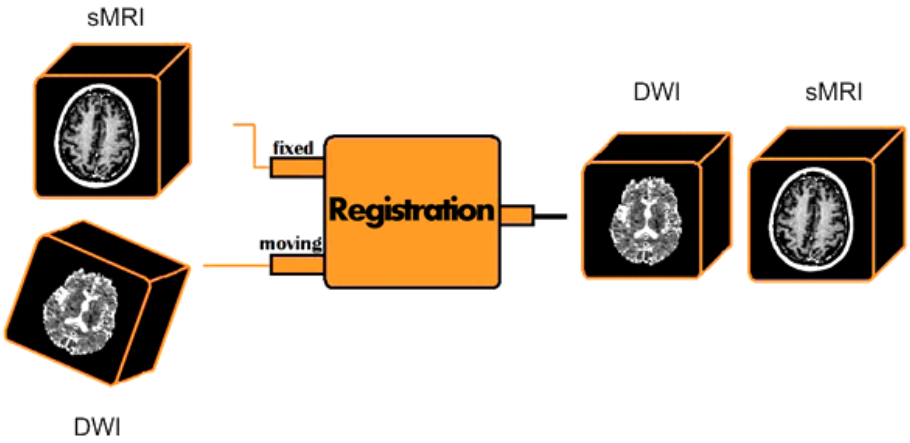

Figure 4: Registration of sMRI to DWI.

4 Region Selection

The brain is a complex organ composed of different kinds of tissues related to different physiological properties of brain matter. Moreover, the brain can be segmented into well defined anatomical structures which are associated to specific functions of the brain. In order to improve the search of brain abnormalities, it is important to take into account of such kind of brain subdivisions. Two main paradigms are in general defined: i) White matter (WM), Gray matter (GM), and Cerebro-Spinal Fluid (CSF) segmentation, and ii) brain parcellation.

4.1 WM-GM-CSF segmentation

WM-GM-CSF segmentation aims at decomposing the brain into its main kind of tissues. In particular, white matter encloses mainly the axons which connect different parts of the brain, while gray matter contains neural cell bodies. Cerebrospinal fluid is a clear, colorless bodily fluid, that occupies the ventricular system around and inside the brain, and sulci.

4.2 Brain Parcellation

The raw images are acquired using a 1.5 tesla MRI machine and 144 slices are acquired using 384×512 resolution. These images are then transferred to PC workstations in order to be processed for ROI *tracing*. Based on manual identification of landmarks, these slices are resampled and realigned by the medical personnel using the Brains2[2] software. The same software is used to manually trace the ROIs by manually drawing contours enclosing the intended region (See Appendix C for further details on ROI tracing). This was carried out

[2]http://www.psychiatry.uiowa.edu/mhcrc/IPLpages/BRAINS.htm

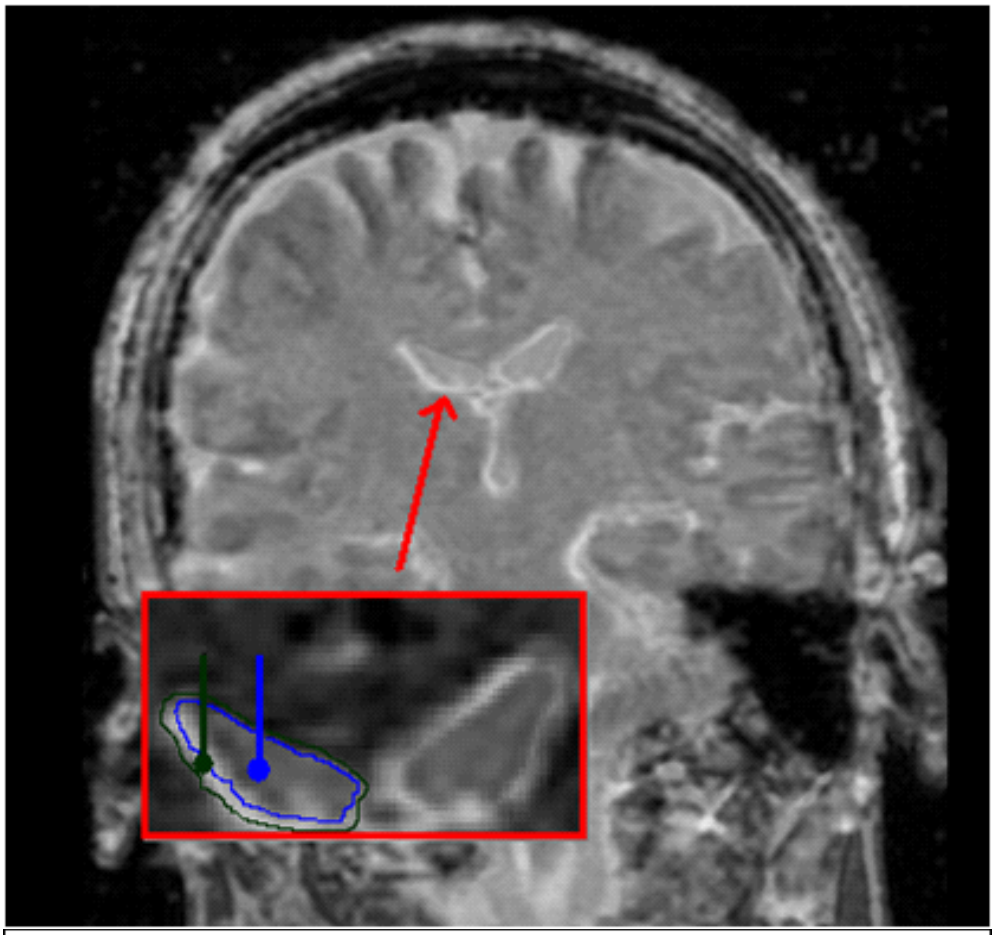

Figure 5: Example of registered sMRI and DWI images.

Figure 6: Example of brain segmentation among White Matter (WM), Gray Matter(GM), and Cerebro-Spinal Fluid (CSF).

by a trained expert following a specific protocol for each ROI (Baiano et al., 2008) without knowledge of the class labels. There are methods which automatically segment the ROIs, but their accuracy is lower than the manual methods so manual segmentation was preferred. The

ROIs traced are 7 pairs (for the left and the right hemisphere respectively) of disconnected image areas:

- Amygdala (*lamyg* and *ramyg* in short);
- Dorso-lateral PreFrontal Cortex (*ldlpfc* and *rdlpfc*);
- Entorhinal Cortex (*lec* and *rec*);
- Heschl's Gyrus (*lhg* and *rhg*);
- Hippocampus (*lhippo* and *rhippo*);
- Superior Temporal Gyrus (*lstg* and *rstg*);
- Thalamus (*lthal* and *rthal*).

We select these ROIs because they have consistently been found to be impaired in schizophrenia and in a recent work, some of them have been found to support a specific altered neural network (Corradi-DellAcqua et al., 2011). The Inter Rater Reliability (IRR) values for each brain hemisphere and ROI can be seen in Table 1 which shows us the reliability of the segmentation. Higher value means the segmentation is more reliable.

Table 1: IRR values for ROI segmentation.

ROI	*left*	*right*
amyg	0.91	0.98
dlpfc	0.93	0.98
ec	0.94	0.96
hg	0.96	0.98
hippo	0.96	0.96
stg	0.93	0.99
thal	0.95	0.96

In Fig. 7, we show a sample from the data set, specifically the ROI volume of the right superior temporal gyrus for subject 11. This volume is made up of 35 slices of size 41×40 and can be arranged as a montage of images (ordered from left to right, top to bottom). Within this bounding box, the ROI itself is irregularly shaped, as can be clearly seen from the corresponding binary masks on the right, artificially colored to highlight the ROI shape.

Additionally, another important ROI that is traced is the *intracranial volume* (ICV), that is the volume occupied by the brain in the cranial cavity leaving out the brainstem and the cerebellum. This information is extremely useful for normalizing volume values against differing overall brain sizes.

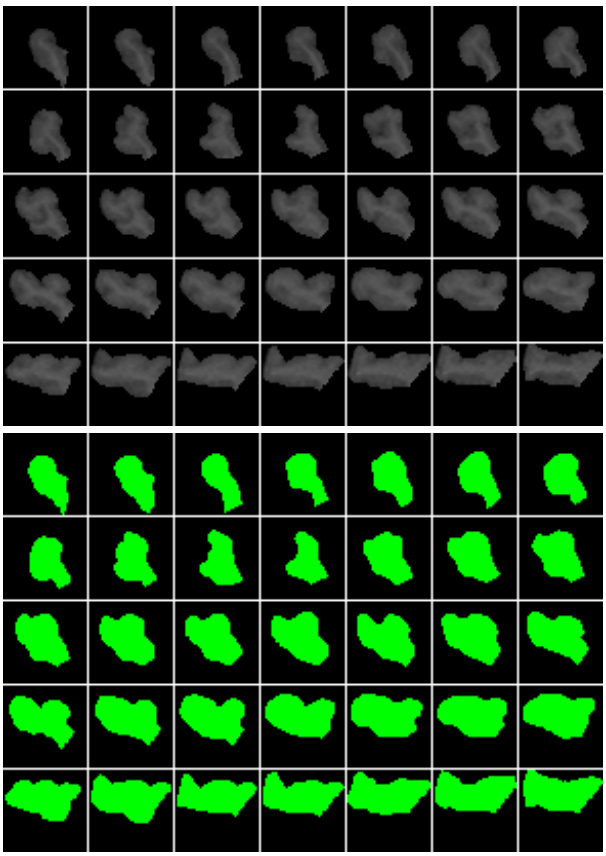

Figure 7: Montage of the slices in the ROI volume ($41 \times 40 \times 35$) of *rstg* for subject 11. At the top, the MRI values; at the bottom, the corresponding binary masks.

5 Data description

In order to encode useful information in a compact representation data descriptors are employed. The overall idea consists in representing the brain with a signature which summarizes brain characteristics, and using such signature for comparison purposes. Several kinds of brain characteristics have been exploited in this project each of them is focusing on a specific aspect of the brain. In particular, we have employed histogram of image *intensities* to encode tissue characteristics, and *geometric* features to concentrate the analysis on shape properties of brain structures. We highlight that, according to standard feature-based approach, such descriptors could be directly used for brain classification. In this project, since we aim at going beyond features, we have exploited new paradigm to deal with such brain characteristics by proposing new approaches for data description (as we will explain in Sections 6 and 7).

In the following we introduce i) Intensity Histograms of sMRI, ii) Histograms of Apparent Diffusion Coefficient values, iii) basic geometric shape descriptor, and iv) spectral shape descriptor.

5.1 Intensity Histograms of Structural MRI Images

From MRI data we compute histograms of image intensities. In particular, we compute a histogram for each ROI. A major disadvantage of MRI compared to other imaging techniques is the fact that its intensities are not standardized. Even MR images taken for the same patient on the same scanner with the same protocol at different times may differ in content due to a variety of machine-dependent reasons, therefore, image intensities do not have a fixed meaning (Nyúl et al., 2000). This implies a significant effect on the accuracy and precision of the following image processing, analysis, segmentation and registration methods relying on intensity similarity.

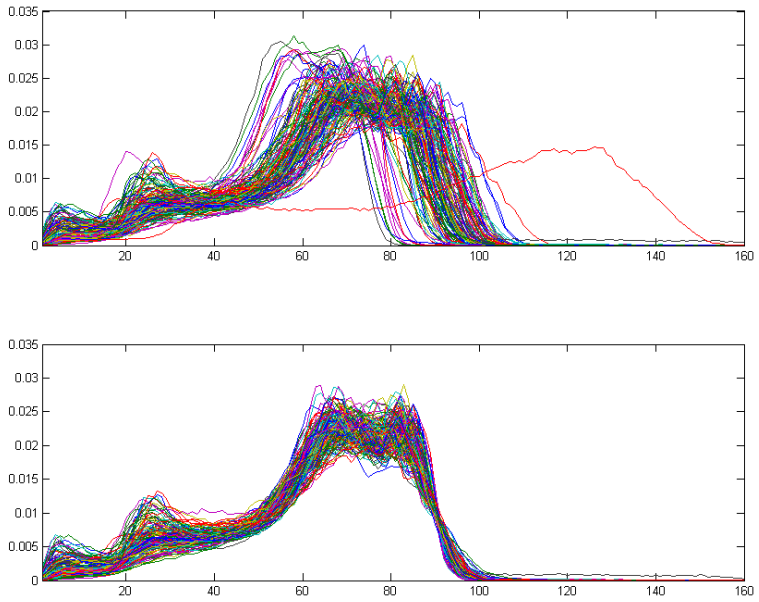

Figure 8: ICV intensity histograms (treated like probability density functions), before and after the normalization process.

A successful technique used to calibrate MR signal characteristics at the time of acquisition employs *phantoms* (Edelstein et al., 1984), by placing physical objects with known attributes within the scanning frame. Unfortunately, this technique is not always exploited, which is our present case. Alternatively, it is possible to obtain good results by retrieving deformation mappings for the image intensities, that is, by developing histogram mappings (Jager and Hornegger, 2009; Nyúl et al., 2000).

In this work, we retrieve the rescaling parameters to form intensity histograms from the ICV histograms (Figure 8). In this way, we focus on the interesting content of the images, which usually contain "noise" in the form of bone and muscle tissue surrounding the brain matter proper (Cheng et al., 2009a). It is also easier to identify landmarks on the histograms that match the canonical subdivision of intracranial tissue into white matter, gray matter and cerebrospinal fluid. We opt to select a simple rescaling mapping that conserves most of the signal in the gray matter - white matter area, corresponding to the two highest bumps in the range 60-90, since ROIs primarily contain those kinds of tissue. With this technique,

we form the histograms of intensity values of images using the whole ROI and use them as bases for our experiments.

5.2 Histograms of Apparent Diffusion Coefficient values

Although we don't have manually segmented ROIs for DWI images, we used deformable registration to segment DWI images into ROIs. For this purpose, every subject's DWI image was registered into the corresponding structural MRI image. Then Apparent Diffusion Coefficient (ADC) values are calculated using these images. We form the histograms of ADC values and use them in our experiments. Since the ADC values are already normalized, we don't need to do another step of normalization on ADC histograms. We can see in Figure 9 examples of ADC histograms of a patient and a healthy control.

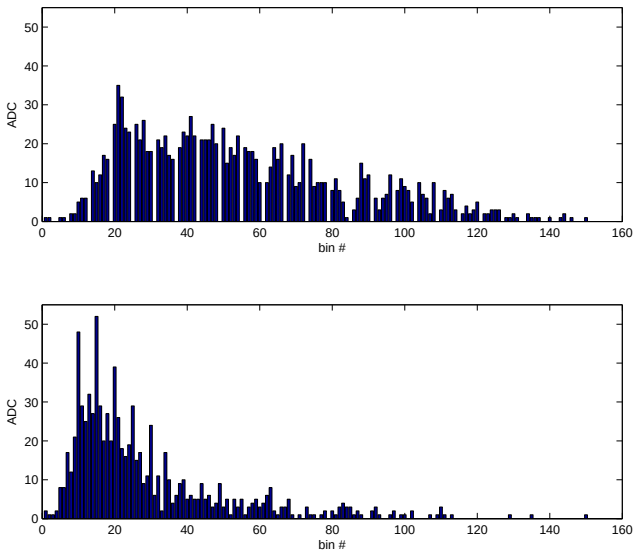

Figure 9: ADC histograms of a patient (top) and a healthy control (bottom).

5.3 Basic geometric shape descriptor

From the set of 2D ROIs of the shapes (slices) the 3D surface is computed as triangle mesh using marching cubes. A minimal smoothing operation is applied to remove noise and voxelization effect. We encode geometric properties of the surface using the *Shape Index* (Koenderink and van Doorn, 1992), which is defined as:

$$si = -\frac{2}{\pi} \arctan\left(\frac{k_1 + k_2}{k_1 - k_2}\right) \quad k_1 > k_2,$$

where k_1, k_2 are the principal curvatures of a generic surface point. The Shape Index varies in $[-1, 1]$ and provides a local categorization of the shape into primitive forms such as spherical cap and cup, rut, ridge, trough, or saddle (Koenderink and van Doorn, 1992).

Shape index is pose and scale invariant (Koenderink and van Doorn, 1992) and it has already been successfully employed in biomedical domain (Awate et al., 2009). The shape index is computed at each vertex of the extracted mesh. Then, all the values are quantized and a histogram of occurrences is computed. Such histograms represent the descriptor of a given subject and it basically encodes the brain local geometry of a subject, disregarding the spatial relationships.

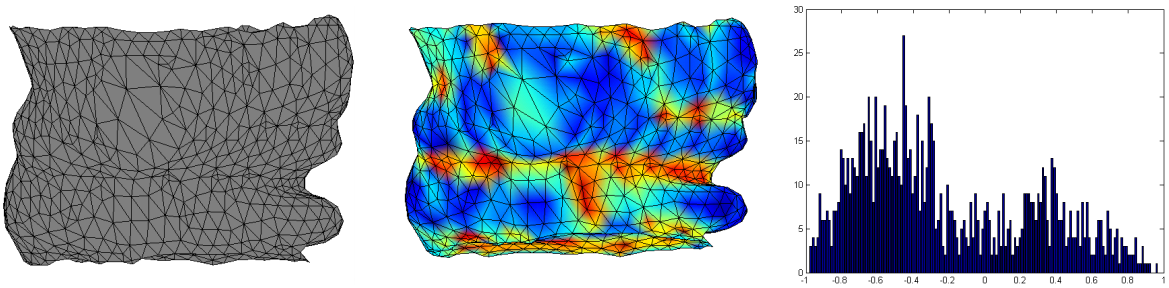

Figure 10: Geometric feature extraction: 3D surface of the left-Amygdala (left), the surface colored according with Shape Index values (center), and the histogram of Shape Index occurrences (right).

Figure 10 shows the 3D surface of the left-Amygdala (left), the surface colored according with Shape Index values (center), and the histogram of Shape Index occurrences (right). It is worth noting that convex regions (in blue) are clearly distinguished from concave regions (in red) by the Shape Index values. As a further step we also calculate the mean curvature using the same methodology):

$$m = \frac{k1 + k2}{2}$$

5.4 Spectral shape descriptor

In this section, we describe a new shape descriptor proposed within the project, which is based on advanced *diffusion* geometry techniques. Local geometric properties are encoded by the so-called *Heat Kernel* (Sun et al., 2009) which exploits heat diffusion characteristics at different scales. The general idea consists of capturing information about the neighborhood of a point on the shape by recording the dissipation of heat over time from that point onto the rest of the shape. In this way, *local* shape characteristics are highlighted through the behavior of heat diffusion over short time periods, and, conversely, *global* shape properties are observed while considering longer periods (Sun et al., 2009; Gebal et al., 2009). So doing, simply varying a single parameter (the time), it is possible to characterize the properties of a shape at different scales. Therefore, local heat kernel values observed at each point are accumulated into a histogram for a fixed number of scales leading to the so-called *Global Heat Kernel Signature* (GHKS). The method is inspired from (Sun et al., 2009) which proposed the *Heat Kernel signature* (HKS) for a single vertex of a mesh. We extend the HKS for the whole shape for both surface mesh (i.e., external surface) and volumetric representation. The proposed descriptor has several nice properties which are shared with very few other works. GHSK allows for shape comparisons using minimal shape preprocessing, in particular, no

registration, mapping, or remeshing is necessary. GHKS is robust to noise since it implicitly employs surface smoothing by neglecting higher frequencies of the shape. Finally, GHKS is able to encode isometric invariance properties of the shape (Sun et al., 2009) which are crucial to deal with shape deformations.

The heat diffusion process. Given a shape M as a compact Riemannian manifold, the heat diffusion on shape[3] is defined by the *heat* equation:

$$(\Delta_M + \frac{\partial}{\partial t})u(t,x) = 0; \tag{1}$$

where u is the distribution of heat on the surface, Δ_M is the *Laplace-Beltrami* operator which, for compact spaces, has discrete eigendecomposition of the form $\Delta_M = \lambda_i \phi_i$. In this fashion the *heat kernel* has the following eigendecomposition:

$$k_t(x,y) = \sum_{i=0}^{\infty} e^{-\lambda_i t}\phi_i(x)\phi_i(y), \tag{2}$$

where λ_i and ϕ_i are the i^{th} eigenvalue and the i^{th} eigenfunction of the Laplace-Beltrami operator, respectively. The heat kernel $k_t(x,y)$ is the solution of the heat equation with point heat source at x at time $t = 0$, i.e., the heat value at point y after time t. The heat kernel is *isometric invariant*, *informative*, *multi-scale*, and *stable* (Sun et al., 2009). In order to estimate the Laplace-Beltrami and the heat kernel in discrete domains several strategies can be employed (Raviv et al., 2010). In the following we describe the cases of surface meshes and volumetric representations.

Heat kernel on surface meshes. In the case of surface mesh only the boundary of the shape is considered. In order to work on a discrete space, we estimate the Laplace-Beltrami operator by employing linear Finite Elements Methods (FEM) (Reuter et al., 2009). More in detail, given a triangular mesh composed by $v_1, \cdots, v_m$ vertices, with *linear* finite elements the *generalized eigendecomposition* problem (Reuter et al., 2009) becomes:

$$A_{cot}\Phi = -\Lambda B \Phi, \tag{3}$$

where Λ is the diagonal matrix of the Laplace Beltrami eigenvalues λ_i, and Φ is the matrix of corresponding eigenfunctions ϕ_i. The matrices A_{cot} and B are defined as:

$$A_{cot}(i,j) = \begin{cases} \frac{cot\alpha_{i,j} + cot\beta_{i,j}}{2} & if \quad (i,j) \in E, \\ -\sum_{k \in N(i)} A_{cot}(i,k) & if \quad i = j, \\ 0 & otherwise. \end{cases} \tag{4}$$

$$B(i,j) = \begin{cases} \frac{|t_1| + |t_2|}{12} & if \quad (i,j) \in E, \\ \frac{-\sum_{k \in N(i)} |t_k|}{6} & if \quad i = j, \\ 0 & otherwise. \end{cases} \tag{5}$$

[3]In this section, we borrow the notation from (Sun et al., 2009; Raviv et al., 2010)

where E is the set of edges of the triangular mesh, $\alpha_{i,j}$ and $\beta_{i,j}$ are the two angles opposite to the edge between vertices v_i and v_j in the two triangles sharing the edge (i, j), $|t_i|$ is the area of the triangle t_i, and t_1, t_2 are the triangles that shares the edge (i, j). Indeed, the *heat kernel* can be approximated on a discrete mesh by computing Equation 2 and retaining the k smallest eigenvalues and the corresponding eigenfunctions.

Heat kernel on volumetric representations. In the case of volumetric representations, the interior part of the shape is also considered. The volume is sampled by a regular Cartesian grid composed of voxels, which allows the use of standard Laplacian in R^3 as the Laplace-Beltrami operator. We use finite differences to evaluate the second derivative in each direction of the volume. The heat kernel on volumes is invariant to volume isometries, in which shortest paths between points inside the shape do not change. Note that in real applications exact volume isometries are limited to the set of rigid transformations (Raviv et al., 2010). However, also non-rigid deformations can faithfully be modelled as approximated volume isometries in practice. Moreover, differently from spectral surface representation, volumetric approach is able to capture volume atrophy. It is worth noting that, as observed in (Sun et al., 2009; Raviv et al., 2010), for small t the heat kernel $k_t(x, x)$ of a point x with itself is directly related to the *scalar* curvature $s(x)$ (Raviv et al., 2010). More formally:

$$k_t(x, x) = (4\pi t)^{-3/2}(1 + \frac{1}{6}s(x)). \tag{6}$$

Note that in the case of surface meshes $s(x)$ can be interpreted as the Gaussian curvature (Sun et al., 2009). In practice, Equation 6 states that heat tends to diffuse slower at points with positive curvature, and vice versa. This gives an intuitive explanation about the geometric properties of $k_t(x, x)$ an leads the idea of using it to build a shape descriptor (Sun et al., 2009).

Global Heat Kernel Signature. Once data are collected, a strategy to encode the most informative properties of the shape M can be devised. To this end, a global shape descriptor is proposed, which is inspired by the so-called *Heat Kernel Signature*(HKS) defined as:

$$HKS(x) = [k_{t_0}(x, x), \cdots, k_{t_n}(x, x)]. \tag{7}$$

where x is a point of the shape (i.e., a vertex of a mesh or a voxel) and $(t_0, t_1, \cdots, t_n)$ are n time values. To extend this approach to the whole shape, we introduce the following global shape descriptor:

$$GHKS(M) = [hist(K_{t_0}(M)), \cdots, hist(K_{t_n}(M))], \tag{8}$$

where $K_{t_i}(M) = \{k_{t_i}(x, x), \forall x \in M\}$, and $hist(\cdot)$ is the histogram operator. Note that our approach combines the advantages of (Bronstein and Bronstein, 2011; Raviv et al., 2010) since it encodes the distribution of local heat kernel values and it works at multiscales. Figure 11 shows a schema of the proposed descriptor. Each point of the shape is colored according to $k_{t_i}(x, x)$. Such values are collected into a histogram for each scale t_i. Finally, histograms are concatenated leading to the global signature.

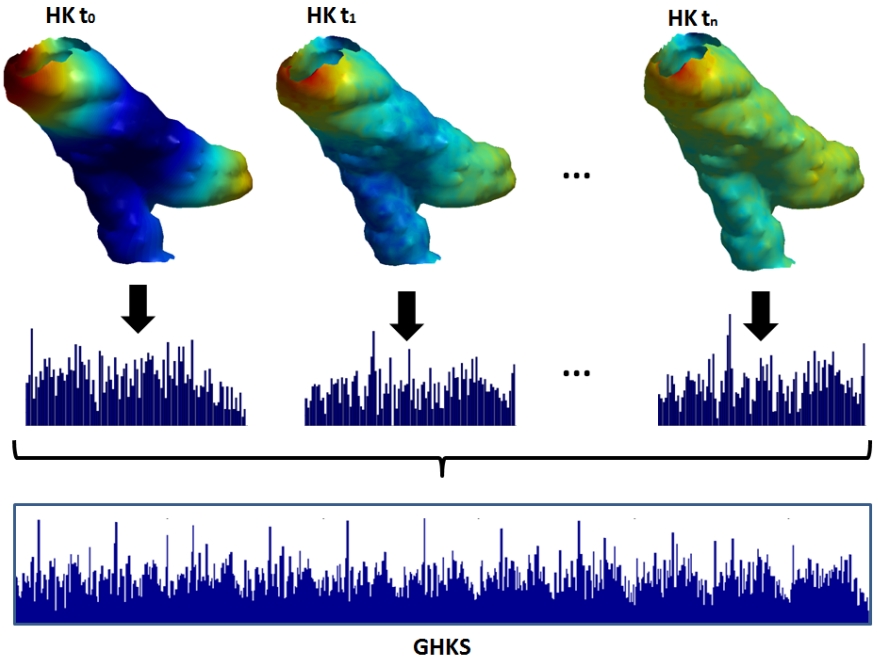

Figure 11: GHKS: Each point of the shape is colored according to $k_{t_i}(x, x)$. Such values are collected into a histogram for each scale t_i. Finally, histograms are concatenated leading to the global signature.

6 Descriptors on Dissimilarity Space

In this section we describe data descriptors generated by employing similarity-based approach. In general, similarity-based approach aims at exploiting the discriminative properties of similarity measures per se, as opposed to standard feature-based approach. In fact, the similarity-based paradigm differs from typical pattern recognition approaches where objects to be classified are represented by feature vectors. Devising pattern recognition techniques starting from similarity measures is a real challenge, and the main goal of this project. Among the different proposed techniques, in this project we investigated the use of the dissimilarity-based representation paradigm, introduced by Ela Pekalska and Bob Duin (Pekalska and Duin, 2005). Withih this approach, objects are described using pair wise (dis)similarities to a representation set of objects . This offers the analyst a different way to express application background knowledge as compared to features. In a second step the dissimilarity representation is transformed into a vector space in which traditional statistical classifiers can be employed. Unlike the related kernel approach, whose application is often restrained by technicalities like fulfilling Mercer's condition, basically any dissimilarity measure can be used.

Similarity-based approach is more versatile in dealing with different data representations (i.e., images, MRI volume, graphs, DNA strings, and so on) since for each kind of data the most suitable (dis)similarity measure can be chosen. In the following we introduce several dissimilarity measures and define the dissimilarity space.

6.1 Dissimilarity measures

Up to this level of the pipeline, data are characterized by hisograms. Therefore we can use histograms to devise similarity measures to be employed in the dissimilarity-based representation scheme. There are various dissimilarity measures that can be applied to measure the dissimilarities between histograms (Cha and Srihari, 2002; Serratosa and Sanfeliu, 2006). Moreover, histograms can be converted to pdfs and dissimilarity measures between two discrete distributions can be used as well. All in all, we decided to study measures below.

Given two histograms S and M with n bins, we define the number of elements in S and M as $|S|$ and $|M|$ respectively.

Histogram intersection It measures the number of intersecting values in each bin (Swain and Ballard, 1991):

$$\text{sim}(S, M) = \frac{\sum_{i=1}^{n} \min(S_i, M_i)}{\min(|S|, |M|)} .$$

Since this is a similarity measure, we convert it to a dissimilarity using $D = \min(|M|, |S|) \times (1 - \text{sim}(S, M))$.

Diffusion distance In diffusion distance (Ling and Okada, 2006)[4], the distance between two histograms is defined as a temperature field $T(x, t)$ with $T(x, 0) = S(x) - M(x)$. Using the heat diffusion equation

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}$$

which has a unique solution

$$T(x, t) = T(x, 0) \times \phi(x, t)$$

where

$$\phi(x, t) = \frac{1}{(2\phi)^{1/2} t} \exp -\frac{x^2}{2t^2},$$

we can compute D as:

$$D = \int_0^r k(|T(x, t)|)dt$$

χ^2 distance This metric is based on the χ^2 test for testing the similarity between histograms. It is defined as

$$D = \sum_{i=1}^{n} \frac{(S_i - M_i)^2}{S_i + M_i} .$$

It is a standard measure for histograms.

[4]The code has been taken from the author's home page: http://www.ist.temple.edu/~hbling/code_data.htm

Earth mover's distance This distance was originally proposed by Rubner et al. (2000). It's basically defined as the cost to transform one distribution into another. It is calculated using linear optimization by defining the problem as a transportation problem. For 1D histograms, it reduces to a simple calculation (Cha and Srihari, 2002) which was implemented in this study.

$$C_i = \left| \sum_{j=1}^{i} (S_j - M_j) \right| \, , \, D = \sum_{i=1}^{n} C_i \, .$$

Similarly, we have considered the following dissimilarities between pdfs:

Bhattacharyya It is used to measure the similarity of discrete probability distributions p and q. It is defined as:

$$D(p,q) = -\log BC(p,q) \, ,$$

where

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)} \, .$$

KullbackLeibler (KL) divergence KullbackLeibler divergence is defined as

$$D(p,q) = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i} \, .$$

This measure is not a distance metric but a relative entropy since $D(p,q) \neq D(q,p)$, i.e., the dissimilarity matrix is not symmetric. There are various ways to symmetrize this dissimilarity. We simply used $D = D(p,q) + D(q,p)$ and the so-called Jensen-Shannon divergence: $D = \frac{1}{2}D(p,r) + \frac{1}{2}D(q,r)$, where r is the average of p and q.

6.2 Dissimilarity space

Suppose that we have n objects and we have a dissimilarity matrix D of size $n \times n$. And suppose that the dissimilarity between two objects o and $\hat{o}$ are denoted by $D(o, \hat{o})$. There are several ways to transform an $n \times n$ dissimilarity matrix D with elements $D(o, \hat{o})$ into a vector space with objects represented by vectors $X = \{x'_1, \ldots, x'_o, \ldots, x'_{\hat{o}}, \ldots, x'_n\}$ (Pekalska and Duin, 2005). Classical scaling (for proper Euclidean dissimilarities) and pseudo-Euclidean embedding (for arbitrary symmetric dissimilarities) yield vector spaces in which vector dissimilarities can be defined that produce the given dissimilarities D. As almost all dissimilarity measures studied here are non-Euclidean, classification procedures for these pseudo-Euclidean spaces are ill-defined, as for instance the corresponding kernels are indefinite.

A more general solution is to work directly in the *dissimilarity space* (see Figure 12). It postulates an Euclidean vector space using the given dissimilarities to a representation set as features. As opposed to the previously mentioned techniques, it is not true anymore that dissimilarities in this space are identical to the given dissimilarities, but this is an advantage in case it is doubtful whether they really represent dissimilarities between the physical objects. As this holds in our case we constructed such a dissimilarity space using all available objects by taking X equal to D. In the dissimilarity space basically any traditional

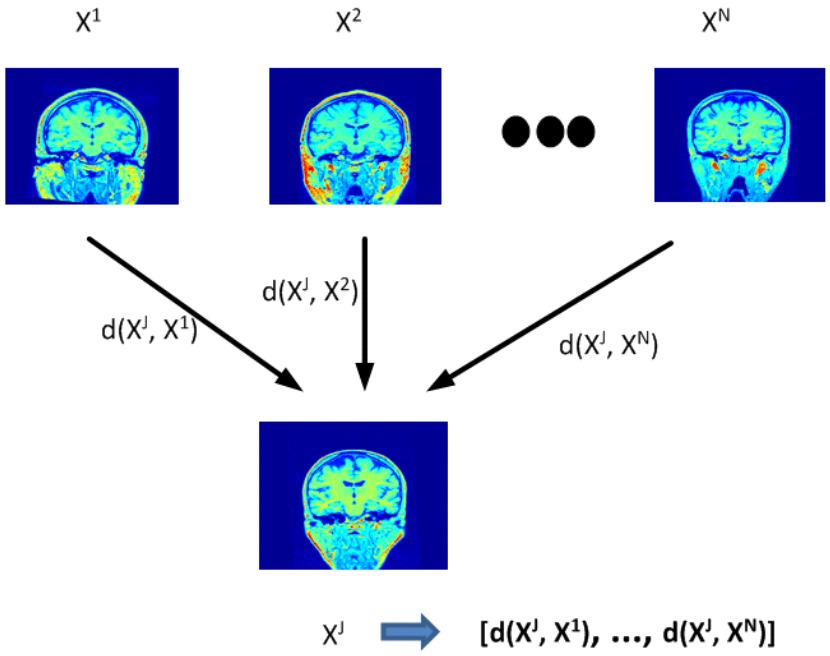

Figure 12: Computation of the dissimilarity space.

classifier can be used. The number of dimensions, however, equals the number of objects. Many classifiers will need dimension reduction techniques or regularization to work properly in this space.

A further refining of the scheme can be obtained by considering at the same time differente dissimilarities (we have many, linked to different modalities, different zones of the brain or different methods to compute them), trying to combine them in a single dissimilarity space. Combined dissimilarity spaces can be constructed by combining dissimilarity representation. As in normal classifier combination (Kittler et al., 1998), a simple and often effective way is using an (weighted) average of the given dissimilarity measures:

$$D_{combined} = \frac{\sum \alpha_i D_i}{\sum \alpha_i} . \tag{9}$$

It is related to the sum-rule in the area of combining classifiers. The weights can be optimized for some overall performance criterion, or determined from the properties of the dissimilarity matrix D_i itself, e.g. its maximum or average dissimilarity. In this work, we used equal weights while combining multiple dissimilarity matrices and all the dissimilarity matrices are scaled such that the average dissimilarity is one, i.e.:

$$\frac{D(o,\hat{o})}{\frac{1}{n(n-1)}\sum_{o,\hat{o}} D(o,\hat{o})} = 1 \tag{10}$$

This is done to assure that the results are comparable over the dissimilarities as we deal with dissimilarity data in various ranges and scales. Such scaled dissimilarities are denoted as $\tilde{D}$. In addition, we assume here that the dissimilarities are symmetric. So, every dissimilarity $\tilde{D}(i,j)$ has been transformed by

$$\tilde{D}(i,j) := \frac{\tilde{D}(i,j) + \tilde{D}(j,i)}{2} \tag{11}$$

7 Descriptors by Generative Embedding

In this section we define a new class of data descriptors basing on generative embedding procedure. The overall idea consists of fitting a generative model on training data and using the generative process to define new data-dependent representations. Then, such representation can be plugged into a standard discriminative classifier for classification purposes. This approach is pursued by hybrid architectures of discriminative and generative classifiers which is currently one of the most interesting, useful, and difficult challenges for Machine Learning. The underlying motivation is the proved complementariness of discriminative and generative estimations: asymptotically (in the number of labeled training examples), classification error of discriminative methods is lower than for generative ones (Ng and Jordan, 2002). On the other side, generative counterparts are effective with less, possibly unlabeled, data; further, they provide intuitive mappings among structure of the model and data features. Among these hybrid generative-discriminative methods, "generative embeddings" (also called generative score space) grow in the recent years their importance in the literature (Jaakkola and Haussler, 1998; Tsuda et al., 2002; Smith and Gales, 2002a; Perina et al., 2009a; Bosch et al., 2006; Li et al., 2011; Smith and Gales, 2002b; Bicego et al., 2009).

Generative score space framework consists of two steps: first, one or a set of generative models are learned from the data; then a score (namely a vector of features) is extracted from it, to be used in a discriminative scenario. The idea is to extract fixed dimensions feature vectors from observations by subsuming the process of data generation, projecting them in highly informative spaces called score spaces. In this way, standard discriminative classifiers such as support vector machines, or logistic regressors are proved to achieve higher performances than a solely generative or discriminative approach.

Using the notation of (Smith and Gales, 2002a; Perina et al., 2009a), such spaces can be built from data by mapping each observation x to the fixed-length score vector $\varphi^f_{\hat{F}}(x)$,

$$\varphi^f_{\hat{F}}(x) = \varphi_{\hat{F}} f(\{P_i\left(x|\theta_i\right)\}), \tag{12}$$

where $P_i\left(x|\theta_i\right)$ represents the family of generative models learnt from the data, f is the function of the set of probability densities under the different models, and $\hat{F}$ is some operator applied to it. In general, the generative score-space approaches help to distill the relationship between a model parameters θ and the particular data sample.

Generative score-space approaches are strictly linked to generative kernels family, namely kernels which compute similarity between points through a generative model – the most famous example being the Fisher Kernel (Jaakkola and Haussler, 1998): Typically, a generative kernel is obtained by defining a similarity measure in the score space, e.g. the inner product.

Score spaces are also called model dependent feature extractors, since they extract features from a generative model.

In order to apply the generative embedding scheme to the MRI data we should define a generative model able to explain and model what we have. In this project we adopted as generative model probabilistic the Latent Semantic Analysis (pLSA – citeHofmann01), a tool widely applied in the linguistic and in the computer vision community.

In the following we first describe the basics of the pLSA, then explaining how this model can be applied to our problem, finally describing the kind of generative embeddings we exploited in this project.

7.1 Probabilistic Latent Semantic Analysis

In the Probabilistic Latent Semantic Analysis (PLSA – (Hofmann, 2001)) the input is a set of D documents, each one containing a set of words taken from a vocabulary of cardinality W. The documents are summarized by an occurrence matrix of size $W \times D$, where $n(w_j, d_i)$ indicates the number of occurrences of the word w_j in the document d_i. In PLSA, the presence of a word w_j in the document d_i is mediated by a latent *topic* variable, $z \in Z = \{z_1,..., z_Z\}$, also called *aspect* class, *i.e.*,

$$P(w_j, d_i) = \sum_{k=1}^{Z} P(z_k) \cdot P(w_j|z_k) \cdot P(d_i|z_k) \tag{13}$$

In practice, the topic z_k is a probabilistic co-occurrence of words encoded by the distribution $\beta(w) = p(w|z_k)$, $w = \{w_1,..., w_N\}$, and each document d_i is compactly (usually, $Z < W$) modeled as a probability distribution over the topics, *i.e.*, $p(z|d_i)$, $z = \{z_1,..., z_Z\}$ (note that this formulation, derived from $p(d_i|z)$, provides an immediate interpretation).

The hidden distributions of the model, $p(w|z)$, $p(d|z)$ and $p(z)$, are learnt using Expectation-Maximization (EM) (Dempster et al., 1977), maximizing the model data-log likelihood $\mathcal{L}$:

$$\mathcal{L} = \prod_{j=1}^{W} \prod_{i=1}^{D} n(w_j, d_i) \cdot \log\left(p(w_j, d_i)\right) \tag{14}$$

The E-step computes the posterior over the topics, $p(z|w, d)$, and the M-step updates the hidden distributions. Once the model has been learnt one can estimate the topic proportion of an unseen document. Here, the learning algorithm is applied by fixing the previously learnt parameters $p(w|z)$ and estimating $p(d|z)$ for the document in hand. For a deeper review of PLSA, see Hofmann (2001). It is important to note that d is a dummy index into the list of documents in the training set. Thus, d is a multinomial random variable with as many possible values as there are training documents and the model learns the topic mixtures $p(d|z)$ only for those documents on which it is trained. For this reason, PLSA is not a well-defined generative model of documents; there is no natural way to assign probability to a previously unseen document.

Even if pLSA is a model for documents, it has been largely applied in other contexts, especially in computer vision (Cristani et al., 2008; Bosch et al., 2006) but also in the medical informatics domain (Bicego et al., 2010a; Castellani et al., 2010; Bicego et al., 2010b). The idea under its application in the MRI domanin is straightforward. In particular, we can assume that a given brain (or the particular ROI) represents the documents d, whereas the words w_j are the local features previous described. With such a point of view, the extracted histograms represent the counting vectors, able to describe how much a visual feature (namely a word) is present in a given image (namely a document). Given the histograms, pLSA is trained following the procedure described in Section 2.

7.2 PLSA-based generative embeddings

Once trained a generative model, different spaces can be obtained. Generally speaking, we can divide them into two families: parameters-based and hidden variable-based. The former

class derives the features on the basis of differential operations linked to the parameters of the probabilistic model, while the latter seeks to derive feature maps on the basis of the log likelihood function of a model, focusing on the random variables rather than on the parameters.

7.2.1 Parameters based score space

These methods derive the features on the basis of differential operations linked to the parameters of the probabilistic model.

The Fisher score

Fisher kernel (Jaakkola and Haussler, 1998) was the first example of generative score space. At first, a parameter estimate $\hat{\theta}$ is obtained from training examples. Then, the tangent vector of the data log likelihood $\log p(x|\theta)$ is used as a feature vector. Referring to the notation of (Smith and Gales, 2002a; Perina et al., 2009a), the score function is the data log likelihood, while the score argument is the gradient.

The fisher score for the PLSA model has been introduced in (Hofmann, 2000), starting from the asymmetric formulation of PLSA. In this case, the log-probability of a document d_i is defined by

$$l(d_i) = \frac{\log P(d_i, w)}{\sum_m n(d_i, w_m)} = \sum_{j=1}^{W} \hat{P}(w_j|d_i) \log \sum_{k=1}^{Z} P(w_j|z_k)P(d_i|z_k)P(z_k), \tag{15}$$

where $\hat{P}(w_j|d_i) \equiv n(d_i, w_j)/\sum_m n(d_i, w_m)$ and where $l(d_i)$ represents the probability of all the word occurrences in d_i normalized by document length.

Differentiating Eq. 15 with respect to $P(z)$ and $P(w|z)$, the pLSA model parameters, we can compute the score. In formulae:

$$\frac{\partial l(d_i)}{\partial P(w = r|z = t)} = n(d_i, w = t) \cdot \frac{P(d_i|z = t)P(z = t)}{\sum_k P(w = r|z = k)P(d_i|z = k)P(z = k)} \tag{16}$$

$$\frac{\partial l(d_i)}{\partial P(z = t)} = \sum_{r=1}^{M} n(d_i, w = j) \cdot \frac{P(d_i|z = t)P(w = j|z = t)}{\sum_k P(z = k)P(w = j|z = k)P(d_i|z = k)} \tag{17}$$

As visible from Eq. 16-17, the samples are mapped in a space of dimension $W \times Z + Z$. The fisher kernel is defined as the inner product in this space. We will refer to it as FSH.

TOP Kernel scores

Top Kernel and the tangent vector of posterior log odds score space were introduced in (Tsuda et al., 2002). One of the aim of the paper was to introduce a performance measure for score spaces. They considered the estimation error of the posterior probability by a logistic regressor and they derived the TOP kernel in order to maximize the performance.

Whereas the Fisher score is calculated from the marginal log-likelihood, TOP kernel is derived from Tangent vectors Of Posterior log-odds. Therefore the two score spaces have the same score function (i.e., the gradient) but different score argument, which, for TOP kernel

$f(p(x|\theta)) = \log P(c = +1|x, \theta) - logP(c = -1|x, \theta)$ where, c is the class label. We will refer to it as TOP.

Log likelihood ratio score space

The log likelihood ratio score space is introduced in (Smith and Gales, 2002b). Its dimensions are similar to the Fisher score, except that the procedure is repeated for each class: a model per class is learnt θ_c and the gradient is applied to each $\log p(x|\theta_c)$. The dimensionality of the resulting space is $C\times$ the dimensionality of the original Fisher score. We will refer to it as LLR.

7.2.2 Random variable based methods

These methods, starting from considerations in (Perina et al., 2009a), seek to derive feature maps on the basis of the log likelihood function of a model, focusing on the random variables rather than on the parameters in their derivation (as done in the parameter-based score spaces).

Free Energy Score Space (FESS)

In the Free Energy Score Space (Perina et al., 2009a), the score function is the free energy while the score argument is its unique decomposition in addends that composes it[5]. Free energy is a popular score function representing a lower bound of the negative log-likelihood of the visible variables used in the variational learning. For pLSA it is defined by the following equation:

$$\begin{aligned} \mathcal{F}(d_i) &= \sum_w n(d_i, w) \cdot \sum_z P(z|d, w) \cdot \log P(z|d, w) \\ &- \sum_w n(d_i, w) \cdot \sum_z P(z|d, w) \cdot \log P(d, w|z) \cdot P(z) \end{aligned} \tag{18}$$

where the first term represents the entropy of the posterior distribution and the second term is the cross-entropy. For further details on the free energy and on variational learning see (Frey and Jojic, 2005), on the pLSA's free energy see (Hofmann, 2001).

As visible in Eq. 18 both terms are composed of $Z \times W$ addends $\{f_j\}_{j=1}^{Z\times W}$, and their sum is equal to the free energy. In generative classification, a test data is assigned to the class which gives the lower free energy (i.e., higher log likelihood). The idea of FESS is to decompose the free energy of each class in its addends, i.e., $\mathcal{F}(d_i)^c = \sum_j \{f_{j,c}\}$ and to add a discriminative layer by estimating a set of weights $\{w_{j,c}\}$ through a discriminative method.

For pLSA this results in a space of dimension equal to $C \times 2 \times Z \times W$; in (Perina et al., 2009a) the authors point out that, if the dimensionality is too high, some of the sums can be carried out to reduce the dimensionality of the score vector before learning the weights. The choice of the addend to optimize is intuitive but guided by the particular application. In our case, as previously done in (Li et al., 2011; Perina et al., 2009b), we perform the sums over the word indices, optimizing the topics contribute. The resulting score space has dimension equal to $C \times 2 \times Z$; we will refer to this score space FESS.

[5]This is true once a family for the posterior distribution is given. See the original paper for details.

Posterior Divergence

Posterior Divergence score space is described in (Li et al., 2011). Like FESS it takes into account how well a sample fits the model (cross entropy terms in FESS) and how uncertain the fitting is (entropy terms in FESS, Eq. 18) but it also assesses the change in model parameters brought on by the input sample, i.e. how much a sample affects the model. These three measures are not simply stacked together, but they are derived from the incremental EM algorithm which, in the E-step only looks at one or few selected samples to update the model at each iteration. Details on posterior divergence score vector for pLSA and on its relationships with FESS case can be found in (Li et al., 2011). We will refer to this score space as PD.

Classifying with the mixture of topics of a document

Very recently, pLSA has been used as a dimensionality reduction method in several fields, like computer vision, bioinformatics and medicine (Bosch et al., 2006; Bicego et al., 2010b; Castellani et al., 2010). The idea is to learn a pLSA model to capture the co-occurrence between visual words (Bosch et al., 2006; Castellani et al., 2010), or gene expressions (Bicego et al., 2010b), which represent the (usually) high-dimensional data description; co-occurrences are captured by the topics. Subsequently, the classification is performed using the topic distribution that defines a document as sample descriptor.

Since we are extracting features from a generative model, we are defining a score space which is the Z-dimensional simplex. In this case, the score argument f, a function of the generative model, is the topic distribution $P(z|d)$ (using Bayes' formula, one can easily derive $P(z|d)$ starting from $P(d|z)$), while the score function is the identity. We will refer to this score space as TPM.

In our experiments, for the two score spaces FESS and TPM, we include two versions. The first version is where we train one pLSA per class and concatenate the resulting feature vectors (we will refer these as FESS-1 and TPM-1), the second one is where we train a pLSA for the whole data without looking at the class label (we will refer these as FESS-2 and TPM-2). All in all, we have eight different score spaces: TPM-1, TPM-2, FESS-1, FESS-2, LLR, FSH, TOP, PD.

8 Classification

After data description step a learning-by-example procedure is employed for brain classification in order to discriminate between healthy subjects and patients affected by schizophrenia. As basic approach, when a single source of information is considered, a standard single classifier can be employed. From the medical point of view this means that the relevance of a particular source of information is considered to characterize the brain abnormality. On the other hand, when several factors can be the possible cause of the disease, a multi-source classification strategy may be employed. For this project we have exploited two paradigms: i) multi-classification, and ii) Multiple Kernel Learning (MKL).

8.1 Multi-classifier

It is a well-known fact that there is no single most accurate classification algorithm so methods have been proposed to combine classifiers based on different assumptions (Kuncheva,

2004; Alpaydın, 2004). Classifier combination (also called ensemble construction) can be done at different levels and in different ways: (i) sensor fusion, (ii) representation fusion, (iii) algorithm fusion, (iv) decision fusion and others. Each classifier (<algorithm/parameter set/data representation> triplet) makes a different assumption about the data and makes errors on different instances and by combining multiple classifiers; the overall error can be decreased. Classifiers making different errors on different parts of the space is called "diversity" (in a broad definition) and to achieve diversity, it has been proposed (Kuncheva, 2004) to use different (i) learning algorithms, (ii) hyperparameters, (iii) input features, and (iv) training sets.

There are various methods to combine classifiers; the simple method is to use voting (Kittler et al., 1998) (or taking an average over the outputs) which corresponds to fixed rules which we usually applied during the course of the project when the classifiers created posterior probability outputs, i.e. $P(C_k|\boldsymbol{x}, E) = \sum_{i=1}^{L} P(C_k|\boldsymbol{x}, M_i)$, where E denotes the ensemble, $P(C_k|\boldsymbol{x}, E)$ is the posterior of the ensemble for class C_k, L is the number of classifiers to combine, $M_i, i = 1 \dots L$ are the individual classifiers to combine, and $P(C_k|\boldsymbol{x}, M_i)$ is the posterior of classifier M_i. Voting does not require any parameter to be optimized and is simple. Other methods such as weighted averaging or more advanced methods require the estimation of other parameters. Bagging (Breiman, 1996) uses bootstrapping to generate different training sets and takes an average, the random subspace method (Ho, 1998) trains different classifiers with different subsets of a given feature set. The simple idea of AdaBoost is to find instances that are incorrectly classified in one iteration and give them a higher probability to be selected in the next iteration. In a mixture of experts architecture, models are local and a separate gating network selects one of the local experts based on the input (Jacobs et al., 1991). In our preliminary works (Cheng et al., 2009b; Ulaş et al., 2011a), we used single classifier and multi-classifier approaches to schizophrenia detection with correlation analysis which were reported in the first deliverable WP7.1 (Ulaş et al., 2010a) to serve as a baseline for our dissimilarity based analysis.

8.2 Multiple Kernel Learning (MKL)

The main idea behind SVMs (Vapnik, 1998) is to transform the input feature space to another space (possibly with a greater dimension) where the classes are linearly separable. After training, the discriminant function of SVM becomes $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle + b$, where $\boldsymbol{w}$ is the vector of weights, b is the threshold, and $\Phi(\cdot)$ is the mapping function. Using the dual formulation and the kernel trick, one does not have to define this mapping function explicitly and the discriminant function can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

where $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$ is the kernel function that calculates a similarity metric between data instances. Selecting the kernel function is the most important issue in the training phase; it is generally handled by choosing the best-performing kernel function among a set of kernel functions on a separate validation set.

In recent years, MKL methods have been proposed (Bach et al., 2004; Lanckriet et al.,

2004), for learning a combination k_η of multiple kernels instead of selecting only one:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = f_\eta(\{k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)_{m=1}^P\}; \boldsymbol{\eta}) \tag{19}$$

where the combination function f_η forms a single kernel from P base kernels using the parameters $\boldsymbol{\eta}$. Different kernels correspond to different notions of similarity and instead of searching which works best, the MKL method does the picking for us, or may use a combination of kernels. MKL also allows us to combine different representations possibly coming from different sources or modalities.

Linear Multiple Kernel Learning

There is significant work on the theory and application of MKL and most of the proposed algorithms use a linear combination function such as convex sum or conic sum. Fixed rules use the combination function in (19) as a fixed function of the kernels, without any training. Once we calculate the combined kernel, we train a single kernel machine using this kernel. For example, we can obtain a valid kernel by taking the mean of the combined kernels.

Instead of using a fixed combination function, we can also have a function parameterized by a set of parameters and then we have a learning procedure to optimize these parameters as well. The simplest case is to parameterize the sum rule as a weighted sum:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)$$

with $\eta_m \in \mathbb{R}$. Different versions of this approach differ in the way they put restrictions on the kernel weights: (Bach et al., 2004; Lanckriet et al., 2004; Rakotomamonjy et al., 2008). For example, we can use arbitrary weights (i.e., linear combination), nonnegative kernel weights (i.e., conic combination), or weights on a simplex (i.e., convex combination).

Nonlinear Multiple Kernel Learning

A linear combination may be restrictive and nonlinear combinations are also possible (Cortes et al., 2010; Gönen and Alpaydın, 2008; Lewis et al., 2006). Cortes et al. (2010) developed a nonlinear kernel combination method based on kernel ridge regression (KRR) and polynomial combination of kernels. The nonlinear combination can be formulated as

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{\boldsymbol{q} \in \mathcal{Q}} \eta_{q_1 q_2 \ldots q_P} k_1(\boldsymbol{x}_i^1, \boldsymbol{x}_j^1)^{q_1} k_2(\boldsymbol{x}_i^2, \boldsymbol{x}_j^2)^{q_2} \ldots k_P(\boldsymbol{x}_i^P, \boldsymbol{x}_j^P)^{q_P}$$

where $\mathcal{Q} = \{\boldsymbol{q} \colon \boldsymbol{q} \in \mathbb{Z}_+^P, \ \sum_{m=1}^P q_m \leq d\}$ and $\eta_{q_1 q_2 \ldots q_P} \geq 0$. The number of parameters to be learned is too large and the combined kernel is simplified in order to reduce the learning complexity:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{\boldsymbol{q} \in \mathcal{R}} \eta_1^{q_1} \eta_2^{q_2} \ldots \eta_P^{q_P} k_1(\boldsymbol{x}_i^1, \boldsymbol{x}_j^1)^{q_1} k_2(\boldsymbol{x}_i^2, \boldsymbol{x}_j^2)^{q_2} \ldots k_P(\boldsymbol{x}_i^P, \boldsymbol{x}_j^P)^{q_P}$$

where $\mathcal{R} = \{\boldsymbol{q} \colon \boldsymbol{q} \in \mathbb{Z}_+^P, \ \sum_{m=1}^P q_m = d\}$ and $\boldsymbol{\eta} \in \mathbb{R}^P$. For example, when $d = 2$, the combined kernel function becomes

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{P} \sum_{h=1}^{P} \eta_m \eta_h k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m) k_h(\boldsymbol{x}_i^h, \boldsymbol{x}_j^h). \tag{20}$$

The combination weights are optimized by solving the following min-max optimization problem:

$$\underset{\boldsymbol{\eta}\in\mathcal{M}}{\text{mininimize}} \;\; \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\text{maximize}} \;\; \boldsymbol{y}^\top\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top(\mathbf{K}_\eta + \lambda\mathbf{I})\boldsymbol{\alpha}$$

where $\mathcal{M}$ is a positive, bounded, and convex set. Two possible choices for the set $\mathcal{M}$ are the l_1-norm and l_2-norm bounded sets defined as

$$\begin{aligned}\mathcal{M}_1 &= \{\boldsymbol{\eta}\colon \boldsymbol{\eta}\in\mathbb{R}_+^P,\;\; \|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|_1 \le \Lambda\} \\ \mathcal{M}_2 &= \{\boldsymbol{\eta}\colon \boldsymbol{\eta}\in\mathbb{R}_+^P,\;\; \|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|_2 \le \Lambda\}\end{aligned} \tag{21}$$

where $\boldsymbol{\eta}_0$ and Λ are two model parameters. A projection-based gradient-descent algorithm can be utilized to solve this min-max optimization problem. At each iteration, $\boldsymbol{\alpha}$ is obtained by solving a KRR problem with the current kernel matrix and $\boldsymbol{\eta}$ is updated with the gradients calculated using $\boldsymbol{\alpha}$ while considering the bound constraints on $\boldsymbol{\eta}$ due to $\mathcal{M}_1$ or $\mathcal{M}_2$.

We formulate a variant of this method by replacing KRR with SVM as the base learner. In that case, the optimization problem becomes

$$\underset{\boldsymbol{\eta}\in\mathcal{M}}{\text{mininimize}} \;\; \underset{\boldsymbol{\alpha}\in\mathcal{A}}{\text{maximize}} \;\; J_\eta = \mathbf{1}^\top\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top((\boldsymbol{y}\boldsymbol{y}^\top)\odot\mathbf{K}_\eta)\boldsymbol{\alpha}$$

where $\mathcal{A}$ is defined as

$$\mathcal{A} = \{\boldsymbol{\alpha}\colon \boldsymbol{\alpha}\in\mathbb{R}_+^P,\;\; \boldsymbol{y}^\top\boldsymbol{\alpha} = 0,\;\; \boldsymbol{\alpha}\le C\}.$$

We solve this optimization problem again using a projection-based gradient-descent algorithm. When updating the kernel parameters at each iteration, the gradients of J_η with respect to $\boldsymbol{\eta}$ are used. These gradients can be written as

$$\frac{\partial J_\eta}{\partial \eta_m} = -\frac{1}{2}\sum_{h=1}^{P}\eta_h\boldsymbol{\alpha}^\top((\boldsymbol{y}\boldsymbol{y}^\top)\odot\mathbf{K}_h\odot\mathbf{K}_m)\boldsymbol{\alpha}.$$

9 Case study 1: Brain classification on dissimilarity space

After presenting all the possible choices we made in the different parts of the pipeline, let us present some concrete systems. In particular here we describe a method based on the dissimilarity-representation paradigm, whereas in Section 10 a method based on the generative embeddings is presented.

Concerning the taxonomies presented in Figure 1, here we are using both sMRI and DWI approach (namely a multimodal scheme), starting from the brain parcellation, employing the dissimilarity-based description and dissimilarity combination by classifying with a single classifier.

In particular these experiments (Ulaş et al., 2010b, 2011c) which is a joint work of Verona and Delft groups, we use a 114 subject subset of the original data set (59 patients, 55 healthy controls). We used the intensity histograms from sMRI images (SMRI), ADC histograms

from DWI images (ADC) and two geometric shape descriptors: shape index (SH) and mean curvature (MCUR). We used all the ROIs and used the dissimilarity space by computing the distances between the histograms and their corresponding pdfs. In summary, for each ROI and representation we use the following 13 measures:

- *hist-euclid*: Euclidean distance between histograms.
- *hist-l1*: L1 distance between histograms.
- *hist-intersect*: Intersection between histograms.
- *hist-diffusion*: Diffusion distance between histograms.
- *hist-chi*: χ^2 distance between histograms.
- *hist-emd*: Earth mover's distance between histograms.
- *pdf-euclid*: Euclidean distance between pdfs.
- *pdf-l1*: L1 distance between pdfs.
- *pdf-emd*: Earth mover's distance between pdfs.
- *pdf-bs*: Bhattacharyya distance between pdfs.
- *pdf-kl*: Symmetrized KL divergence between pdfs.
- *pdf-kl-orig*: Original, asymmetric KL divergence.
- *pdf-js*: Jensen-Shannon divergence between pdfs.

All in all there are 14 ROIs and 13 different dissimilarity measures per modality, which yields a total of 182×4 dissimilarity matrices. In addition to these, we propose to merge the different dissimilarity matrices into one overall dissimilarity matrix per modality potentially exploiting complementary information useful to improve the classification accuracy. We also test the accuracy of these combinations against combining classifiers in the original feature space (histograms and pdfs for each of the four modalities). For each test we evaluated the leave-one-out error. All differences in accuracy reported in this case study are significant at $p = 0.05$ using the paired t-test. The dissimilarity spaces have been built in a transductive way by using all available subjects for dissimilarity (of course labels are ignored in this phase). Three classifiers are considered to compare the performances:

- Linear SVM classifier on the original feature space (called *svm*)
- the 1-Nearest Neighbour (NN) rule on the dissimilarity matrices (called *1nn*)
- Linear SVM classifier on the dissimilarity space (called *sv0*)

The linear SVM in dissimilarity space avoids complications that could arise from the dissimilarity measures being non-Euclidean because we treat the dissimilarities as features in this new space. While combining dissimilarities, we use for α_i in (Eq. 9) the reciprocal of the number of dissimilarity matrices to be combined (Lee et al., 2010). On the original feature space, the SVM classifiers produce posterior probability outputs, and these outputs are combined using the SUM rule (Kittler et al., 1998). So, on the original feature space, we combine after training the classifiers, whereas on the dissimilarity space, we combine before we do classification. The experiments are carried out using the Matlab package PRTools (Duin, 2005). We designed three experiments to show the improvements of dissimilarity-based pattern recognition techniques and combination of dissimilarities using multiple ROIs and modalities:

1. ROI-based classification: for each modality, we report the highest accuracy that a classifier reaches without combination (on the original feature space and on the dissimilarity space). We use these results as baseline for comparison.

2. Multi-ROI classification: in these set of experiments, for each modality, we fix the dissimilarity measure and combine all ROIs using this dissimilarity measure.

3. Multimodal classification: in this experiment, we go one further step to combine information coming from different sources by combining different dissimilarity matrices from different modalities.

We note that, through this section, we will use the following notation: every dissimilarity representation will be referred to as MODALITY-*roi-dissimilarity-measure* (i.e. SMRI-*ldlpfc-pdf-js* shows the dissimilarity matrix for the structural MRI of ROI *ldlpfc* using the dissimilarity measure of Jensen-Shannon divergence). The modality, ROI, or the dissimilarity measure will be omitted when its clear from the context.

9.1 ROI-based classification

We evaluate the classification accuracies for each of the original dissimilarity matrices. Table 2 summarizes the results for structural MRI. For each ROI the best performance is reported with respect to various dissimilarity measures. First column reports the accuracy estimates for *svm* using the original feature space (histograms and pdfs). Second column reports the maximum accuracy of *1nn* on different dissimilarity measures. Third column reports the leave-one-out accuracy estimates of the linear SVM in dissimilarity space. It shows clearly the improvements of our dissimilarity-based approach. Except for two ROIs (*rhg* and *rthal*) SVM classifier in the dissimilarity space is always better than classifiers in the standard space. While the best accuracy of standard approaches is 68.42 per cent, we can reach 78.07 % accuracy on dissimilarity space and dissimilarity space accuracies are more stable.

Table 3 shows the same results for the ADC values extracted from DWI images. We can again see that when we switch to dissimilarity based classification, we get better accuracies (either *1nn* or *sv0*) except for two ROIs (*lamyg* and *rec*). We can again see that with a single ROI and dissimilarity measure, we can reach 70.18 % whereas the highest accuracy we can obtain in the original space is 64.04 %. The same pattern can be observed when we investigate Tables 4 and 5. Also in these modalities, the best accuracy can be achieved using

Table 2: Best single ROI accuracies for each ROI on histograms of intensities.

ROI	*svm*	*1nn*	*sv0*
lamyg	68.42 (pdf)	64.04 (hist-l1)	**78.07 (pdf-bc)**
ramyg	54.39 (hist)	65.79 (pdf-l1)	66.67 (hist-chi)
ldlpfc	60.53 (hist)	62.28 (pdf-kl-orig)	76.32 (pdf-js)
rdlpfc	64.04 (hist)	57.89 (hist-intersect)	68.42 (pdf-kl-orig)
lec	64.04 (pdf)	56.14 (pdf-emd)	64.91 (hist-l1)
rec	64.91 (pdf)	65.79 (pdf-l1)	71.05 (hist-intersect)
lhg	51.75 (pdf)	60.53 (pdf-l1)	63.16 (pdf-bc)
rhg	50.00 (hist)	*63.16 (hist-l1)*	59.65 (hist-emd)
lhippo	63.16 (pdf)	60.53 (hist-intersect)	72.81 (pdf-kl-orig)
rhippo	60.53 (pdf)	64.04 (pdf-emd)	66.67 (pdf-bc)
lstg	55.26 (pdf)	59.65 (hist-intersect)	69.30 (hist-chi)
rstg	63.16 (hist)	57.02 (pdf-emd)	64.91 (pdf-kl)
lthal	58.77 (pdf)	64.91 (pdf-l1)	67.54 (hist-l1)
rthal	*64.91 (pdf)*	59.65 (pdf-l2)	64.04 (pdf-emd)

dissimilarities. We can see that on SH, we reach 68.42 % using *1nn* and 65.79 % using *sv0*. The best accuracy using the features on the original space is 55.26 %. Also on MCUR, best accuracy is reached using *sv0*.

Table 3: Best single ROI accuracies for each ROI on histograms of ADC values.

ROI	*svm*	*1nn*	*sv0*
lamyg	*64.04 (hist)*	57.89 (hist-emd)	62.28 (hist-emd)
ramyg	54.39 (pdf)	56.14 (pdf-l2)	59.65 (pdf-bc)
ldlpfc	56.14 (pdf)	51.75 (pdf-l2)	61.40 (pdf-kl)
rdlpfc	54.39 (hist)	56.14 (hist-emd)	65.79 (pdf-emd)
lec	53.51 (pdf)	62.28 (pdf-emd)	61.40 (hist-emd)
rec	*64.04 (pdf)*	58.77 (pdf-l2)	60.53 (pdf-kl-orig)
lhg	55.26 (hist)	61.40 (pdf-l1)	54.39 (pdf-l2)
rhg	50.88 (hist)	58.77 (hist-emd)	58.77 (hist-emd)
lhippo	52.63 (hist)	57.02 (hist-emd)	59.65 (pdf-l1)
rhippo	48.25 (hist)	55.26 (hist-emd)	52.63 (pdf-kl)
lstg	54.39 (pdf)	56.14 (hist-l2)	60.53 (hist-intersect)
rstg	64.04 (hist)	60.53 (hist-emd)	**70.18 (pdf-bc)**
lthal	57.89 (hist)	57.89 (pdf-l1)	58.77 (pdf-kl-orig)
rthal	53.51 (pdf)	60.53 (pdf-bc)	59.65 (pdf-js)

Table 4: Best single ROI accuracies for each ROI on shape index histograms.

ROI	*svm*	*1nn*	*sv0*
lamyg	45.61 (hist)	**68.42 (hist-emd)**	64.91 (hist-emd)
ramyg	49.12 (hist)	53.51 (hist-l2)	57.89 (pdf-kl-orig)
ldlpfc	52.63 (hist)	62.28 (hist-l2)	57.89 (hist-chi)
rdlpfc	54.39 (hist)	59.65 (hist-intersect)	60.53 (hist-chi)
lec	46.49 (hist)	51.75 (pdf-emd)	54.39 (pdf-emd)
rec	52.63 (hist)	60.53 (hist-emd)	57.02 (hist-chi)
lhg	47.37 (hist)	54.39 (pdf-js)	65.79 (hist-intersect)
rhg	43.86 (hist)	55.26 (pdf-emd)	55.26 (hist-intersect)
lhippo	55.26 (hist)	50.00 (hist-diffusion)	57.02 (hist-chi)
rhippo	47.37 (hist)	59.65 (pdf-emd)	57.02 (pdf-kl-orig)
lstg	40.35 (hist)	58.77 (pdf-js)	52.63 (pdf-emd)
rstg	53.51 (hist)	55.26 (pdf-bc)	57.89 (pdf-emd)
lthal	46.49 (hist)	53.51 (pdf-l2)	57.89 (hist-l2)
rthal	54.39 (hist)	57.89 (hist-emd)	59.65 (hist-l1)

Table 5: Best single ROI accuracies for each ROI on mean curvature histograms

ROI	*svm*	*1nn*	*sv0*
lamyg	43.86 (pdf)	61.40 (hist-intersect)	57.02 (hist-l1)
ramyg	46.49 (hist)	58.77 (hist-chi)	57.89 (pdf-kl-orig)
ldlpfc	49.12 (pdf)	53.51 (hist-chi)	53.51 (hist-emd)
rdlpfc	56.14 (pdf)	57.02 (hist-l1)	60.53 (hist-chi)
lec	47.37 (pdf)	53.51 (pdf-kl)	53.51 (pdf-emd)
rec	52.63 (hist)	65.79 (hist-emd)	63.16 ([df-l2)
lhg	61.40 (hist)	56.14 (pdf-bc)	62.28 (pdf-kl-orig)
rhg	57.89 (pdf)	64.04 (pdf-emd)	61.40 (pdf-js)
lhippo	53.51 (hist)	58.77 (hist-intersect)	62.28 (pdf-emd)
rhippo	54.39 (pdf)	55.26 (pdf-emd)	58.77 (pdf-js)
lstg	50.00 (hist)	50.00 (pdf-bc)	51.75 (pdf-l1)
rstg	41.23 (pdf)	63.16 (hist-emd)	57.89 (hist-intersect)
lthal	53.51 (pdf)	56.14 (pdf-emd)	58.77 (pdf-kl-orig)
rthal	48.25 (pdf)	54.39 (hist-chi)	**67.54 (pdf-bc)**

9.2 Multi-ROI classification

In this subsection, we will show our experiments where we combine multiple ROIs, fixing the modality and distance measure. We also conducted experiments by fixing the ROIs and combining multiple dissimilarity matrices using the same ROI. We see that the accuracy does not increase as compared to combining ROIs with fixed dissimilarity measure. This conforms to our previous studies, therefore in this work, we do not report combination of distance measures with fixed ROI.

In this experiment, a multi-ROI approach is adopted to use all ROIs at the same time. All the dissimilarity matrices for each ROI are combined by averaging the normalized dissimilarity matrices. Second and third columns of Table 6 reports the results on intensity histograms, using 1-NN rule on the dissimilarity matrices and the support vector classifiers in the dissimilarity spaces. Also in this case, the classification on the dissimilarity space clearly outperforms the standard approach. Moreover, the multi-ROI approach brings an improvement by confirming the complementary information enclosed onto the different brain subparts when we use *sv0* on the dissimilarity space. In most of the cases, the results from the averaged similarity matrices are better than the respective best single ROI results. The row average in Table 6 reports the error estimates computed on the overall dissimilarity matrix (combining all the measures and ROIs), which has the highest accuracy 76.32% (same as combining all ROIs for *pdf-l1*) for both standard approach and dissimilarity-based approach, respectively. The last row reports the accuracy of combining all SVM classifiers in the original feature space. When we combine all the SVM classifiers on the original space, we get 71.93% accuracy. This shows us that, the dissimilarity space produces better results also when we consider classifier combination. We repeated the same experiments also with the other modalities. In Tables 7, 8, 9, we see the results using the other modalities. We observe that again we get the most accurate results when we combine ROIs in the dissimilarity space using *sv0* except mean curvature histograms where the best results are obtained using *1nn* (using dissimilarities again).

9.3 Combining Different Modalities

As a further step to understand how information from multiple sources can be combined to reach better classification accuracy, we develop another experiment where we combine information from multiple modalities. We have 182 dissimilarity matrices from each of the four modalities. It is not possible to exhaustively search the whole solution space to find the best solution (optimum subset for combination), so instead, we choose the most accurate four ROI-dissimilarity pairs from each modality and do an exhaustive search on the combination of these matrices to get the best result. We can see the selected dissimilarity matrices and their base accuracies in Table 10. With a total of 16 dissimilarity matrices (modality-ROI-dissimilarity triples), we can get the best accuracy 86.84% (last row in Table 10), which contains two dissimilarity matrices from intensities (*ldlpfc-pdf-kl-orig* and *ldlpfc-pdf-bc* both having 75.44% accuracy) and one dissimilarity matrix from shape index (*rdlpfc-hist-chi* with 60.53% accuracy). This accuracy is the best accuracy, which has been reached using dissimilarity combination and cannot be reached using only one modality. Applying the same methodology, we can reach only 76.32% accuracy with *1nn* and 83.33% accuracy with svm on the original feature space. This also shows us why it is important to combine

Table 6: Best accuracies for each dissimilarity measure combining all ROIs on histograms of intensities

Measure	*1nn*	*sv0*
hist-l2	62.28	71.05
hist-l1	62.28	74.56
hist-intersect	66.67	68.42
hist-diffusion	62.28	74.56
hist-chi	57.02	71.05
hist-emd	52.63	58.77
pdf-l2	57.02	74.56
pdf-l1	60.53	**76.32**
pdf-emd	59.65	75.44
pdf-bc	65.79	69.30
pdf-kl	66.67	70.18
pdf-kl-orig	64.04	64.91
pdf-js	65.79	71.93
average	60.53	**76.32**
svm	71.93	

Table 7: Best accuracies for each dissimilarity measure combining all ROIs on histograms of ADC values

Measure	*1nn*	*sv0*
hist-l2	50.00	60.53
hist-l1	46.49	**64.91**
hist-intersect	43.86	61.40
hist-diffusion	46.49	**64.91**
hist-chi	50.88	55.26
hist-emd	58.77	51.75
pdf-l2	57.02	60.53
pdf-l1	54.39	61.40
pdf-emd	57.89	53.51
pdf-bc	53.51	53.51
pdf-kl	55.26	48.25
pdf-kl-orig	49.12	51.75
pdf-js	52.63	54.39
average	51.75	60.53
svm	63.16	

Table 8: Best accuracies for each dissimilarity measure combining all ROIs on shape index histograms

Measure	*1nn*	*sv0*
hist-l2	57.89	57.89
hist-l1	58.77	**60.53**
hist-intersect	40.35	53.51
hist-diffusion	58.77	**60.53**
hist-chi	59.65	57.02
hist-emd	55.26	56.14
pdf-l2	50.88	55.26
pdf-l1	50.88	58.77
pdf-emd	**60.53**	**60.53**
pdf-bc	48.25	57.89
pdf-kl	52.63	59.65
pdf-kl-orig	57.02	59.65
pdf-js	48.25	**60.53**
average	54.39	**60.53**
svm	51.75	

Table 9: Best accuracies for each dissimilarity measure combining all ROIs on mean curvature histograms

Measure	*1nn*	*sv0*
hist-l2	49.12	50.88
hist-l1	50.00	51.75
hist-intersect	53.51	50.88
hist-diffusion	50.00	51.75
hist-chi	**55.26**	48.25
hist-emd	43.86	53.51
pdf-l2	57.02	51.75
pdf-l1	54.39	46.49
pdf-emd	50.00	52.63
pdf-bc	44.74	52.63
pdf-kl	48.25	49.12
pdf-kl-orig	**55.26**	46.49
pdf-js	53.51	48.25
average	54.39	49.12
svm	47.37	

useful information from different sources to come up with better accuracy. We see that the accuracy can be increased when complementary information using different modalities are combined.

In a medical application, besides increasing accuracy, the interpretability of the results is also important. We use this experiment to deduce information on the use of ROIs, their complementary information, and how each modality relates to the detection of schizophrenia. For this purpose, we select all the combinations of distance matrices with accuracies above 82% (we have 69 different combinations) and count the occurrences of dissimilarity matrices for every combination. From Table 10, we can see that most of the combinations include *ldlpfc* of SMRI and the shape index of *rthal*. This shows us that these two modality-ROI pairs contribute and complement other dissimilarity matrices and by using these two in combination, we increase accuracy. After these two dissimilarity matrices, we see that mean curvature of *rthal* and shape index of *rdlpfc* are used in combination the most. These are followed by *ldlpfc* of histogram intensities and the mean curvature of *rec*. With ADC, we see that most used ROI is *rstg*, which has been selected 38 times. This also shows us that the DWI information is the least complementary modality in this scenario and one can design experiments without this modality, focusing on the other modalities. We can use this information to decrease the costs of the operation, that is, not performing DWI analysis. Also we see that the most accurate dissimilarity matrix (SMRI-*lamyg-pdf-bc*) is the eighth most used dissimilarity when we consider combination. This interesting fact shows us that when doing combination, the complementary information is more important than individual accuracies.

Another interesting fact is that some ROIs are more discriminative when the structural information is considered, and some are more discriminative when we consider DWI. The ROIs selected from the structural analysis in this experiment are those, considered crucial for the impaired neural network in schizophrenia and comply with current studies in the literature (Corradi-DellAcqua et al., 2011), in contrast DWI is particularly keen in exploring the microstructural organization of white matter therefore providing intriguing information on brain connectivity (Brambilla and Tansella, 2007) but does not have complementary contribution in this context.

With this analysis, we can open a new perspective of how to use each of these modalities to get better accuracies. One can use this information to setup new experiments considering the contributions of these ROIs on these modalities.

9.4 Discussion

In this case study a novel approach based on dissimilarity-based pattern recognition is proposed for the detection of schizophrenic brains. Several dissimilarity measures are proposed to deal with histograms of different types for different ROIs. ROI-based classification onto the dissimilarity space shows improvements of the standard NN rule and the support vector classifier on the original space. Moreover, a Multi-ROI classification strategy is obtained by simply averaging the similarity matrices observed in each ROI. Such approach improves the single-ROI one, by highlighting the complementary information enclosed in the several ROIs. This confirms the benefit of combining dissimilarity information and fusing information from various regions in the brain.

Table 10: Most accurate four dissimilarity matrices from each modality, their single performances, and number of occurrences in the combination of most accurate results

Selected dissimilarity	Accuracy	Occurrences
SMRI-ldlpfc-pdf-js	76.32	60
SH-rthal-hist-l1	59.65	57
MCUR-rthal-pdf-bc	67.54	52
SH-rdlpfc-hist-chi*	60.53	50
SMRI-ldlpfc-pdf-bc*	75.44	48
SMRI-ldlpfc-pdf-kl-orig*	75.44	48
MCUR-rec-pdf-l1	63.16	47
SMRI-lamyg-pdf-bc	78.07	42
ADC-rstg-hist-l2	65.79	38
SH-lamyg-hist-emd	64.91	38
ADC-rstg-pdf-bc	70.18	20
MCUR-rec-pdf-l2	63.16	17
ADC-rdlpfc-pdf-emd	65.79	14
ADC-rstg-pdf-js	65.79	9
SH-lhg-hist-intersect	65.79	8
MCUR-lhippo-pdf-emd	62.28	1
Dissimilarities with * are in the optimum combination	86.84	

We investigate further to combine information from multiple modalities such as intensities, ADC values and geometric information. We can see that, some ROIs are discriminative when we use intensities; some are useful when DWI data is considered. Geometric properties of some ROIs play a part in schizophrenia detection. We show that we get the best accuracy when we combine multiple modalities.

We can interpret the results of combining multiple modalities to set up further experiments in this context. Our results show that the least contributing modality is the DWI. With this information, one can skip using this modality and focus more on histograms of intensities and geometric information. Also, one can use this result to reduce the costs of this operation, by not performing DWI measurements and without the patient to undergoing further medical operations.

We would like to emphasize that in building the (combined) dissimilarities no parameters are optimized w.r.t. performance. The proposed approach of combining dissimilarities on the dissimilarity space opens new perspectives in neuroanatomy classification by allowing the possibility to exploit dissimilarity measures where one does not have to deal with technical difficulties such as the metric requirements of distance based classification and kernel restrictions of support vector machines.

9.5 Work in progress

As part of finding different ways to combine dissimilarities and adapting different normalization schemes to obtain better classification accuracies, we adapted Ho's random subspace method (Ho, 1998) to dissimilarity construction (Ulaş et al., 2011b). In this case study, we use *Heat Kernel Signatures* to extract histogram based features (see also (Castellani et al., 2011)) from sMRI scans of 30 schizophrenic patients and 30 healthy controls to detect schizophrenia. We first create several dissimilarity matrices using histogram based dissimilarity measures (using L1 norm ((`L1`), Earth mover's distance (`EMD`), KullbackLeibler (KL) divergence (`KL`), and Jensen-Shannon divergence (`JS`) and compare our results with the baseline original space accuracies using the support vector machines with the linear kernel. We focus on only one ROI, namely left thalamus.

9.5.1 Random subspace method and adaptation to dissimilarity computation

In classical pattern recognition, for the combination of classifiers to be effective, one has to create diverse classifiers and combine them in a proper way. Most combination methods aim at generating uncorrelated classifiers, and it has been proposed (Kuncheva, 2004) to use different (i) learning algorithms, (ii) hyperparameters, (iii) input features, and (iv) training sets. For example Bagging (Breiman, 1996) uses bootstrapping to generate different training sets and takes an average, the random subspace method (Ho, 1998) trains different classifiers with different subsets of a given feature set. In dissimilarity based classification, since most of the time the k-nearest classifier is used, the research has focused on prototype selection (Pekalska et al., 2006). In prototype selection, the aim is to select a subset of the data which explains the whole set of data in a proper way. In this work, we aim for the transpose of this problem, instead of selecting a set of instances, we aim to select a set of features and combine them as in (Ho, 1998). This can be done after we have the dissimilarities which was also targeted in (Pekalska et al., 2006, 2000) or before the dissimilarities are actually computed.

In this work, we propose the latter approach and adopt Ho's random subspace method to dissimilarity computation. What we do is we select a random subset of the original set of bins and compute the dissimilarities according to these subsets. Then we combine these subsets to get the final classification. In this way we enrich the dissimilarity space by using more information. This is possible to achieve because we calculate the dissimilarities using histogram bins and is logical because some bins may be more descriptive of the data and computing the dissimilarities using only these bins may increase the classification accuracy. Dividing the feature sets into two and combining Euclidean distances has been investigated recently in (Lee et al., 2010). Once these dissimilarities are computed, they are combined using averaging and concatenation. We show the results using both techniques.

9.5.2 Results

Table 11 reports the accuracies of the SVM with the linear kernel on the original space (`LIN`), linear SVM on the dissimilarity spaces (`DIS`, where `-AVG` shows combining using averaging and `-CONC` shows combination using concatenation) and random subspace accuracies (`RAND-`N, where N is the selected number of bins. For this example, we used 20 random subspaces; the results for other numbers are similar. See also Figures 13 and 14). We can see that the dissimilarity space accuracies are always better than the original space accuracies. This also corresponds with our previous work (Ulaş et al., 2011c). We can also see that with a suitable number of random bins, one can always achieve a better accuracy than the full dissimilarity space accuracies. We can see that the average operation usually creates more accurate results (though the best result is using concatenation on `KL`). We believe that this is because of the curse of dimensionality. When one uses the concatenation operator, the number of features increase and the classification accuracy may decrease accordingly.

Table 11: Accuracies using the dissimilarity space and the random subspace. 20 random subspaces are created.

`LIN`				66.67				
`DIS`	76.67		61.67		75.00		68.33	
	`L1-AVG`	`L1-CONC`	`EMD-AVG`	`EMD-CONC`	`KL-AVG`	`KL-CONC`	`JS-AVG`	`JS-CONC`
`RAND-50`	76.67	75.00	70.00	58.33	73.33	68.33	70.00	75.00
`RAND-25`	78.33	70.00	65.00	58.33	71.67	58.33	71.67	60.00
`RAND-20`	**86.67**	83.33	65.00	55.00	71.67	70.00	68.33	71.67
`RAND-15`	80.00	73.33	73.33	68.33	75.00	**88.33**	73.33	63.33
`RAND-10`	71.67	56.67	65.00	60.00	71.67	66.67	70.00	53.33

We can see that the random subspace method applied to the creation of the dissimilarity space achieves better accuracies than the full dissimilarity space and the original space. One disadvantage of this method is to choose the proper number of bins and proper number of random subspaces. One can use cross validation to decide these values but we leave the exploration of this topic as a future work. Nevertheless, we want to present the performance of these methods when we apply different number of bins and different number of random subspaces. In Figure 13, we can see the change in accuracy versus the number of subspaces created. The numbers on the plots show the number of bins chosen and the peak accuracy using that number of bins. Although there is no clear correlation between accuracy and the

number of subspaces, we can still observe that with a fairly small number of subspaces, one can achieve relatively high accuracies. We can also observe that (as expected), with the increasing number of subspaces, the diversity decreases and the accuracy converges. We can observe that the dissimilarity space is always superior to the original space and (except for `KL` where the choice of the number of subspaces is critical), the random subspace accuracy is higher than full dissimilarity space accuracy.

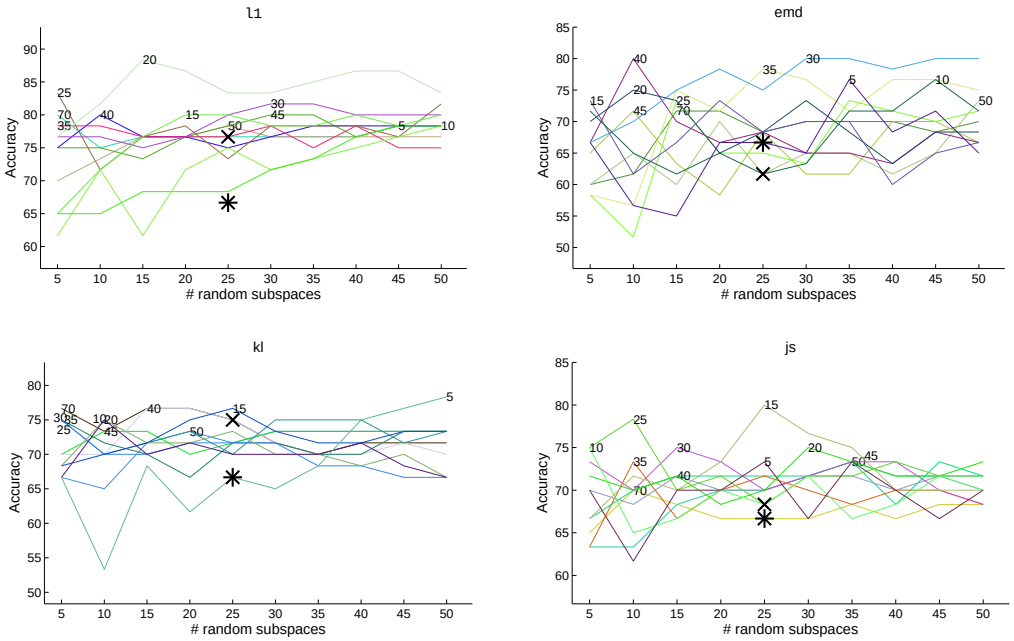

Figure 13: # of subspaces vs accuracy. The numbers in the plot show the number of histogram bins used. X is the accuracy on the full dissimilarity matrix of the corresponding dissimilarity measure and * is the base-line original space accuracy.

In the second set of experiments, we fix the number of random subspaces and see the change in accuracy when we change the number of selected bins. These results are presented in Figure 14. Again, the numbers on the figure show the number of random subspaces and peak accuracy obtained by this selection. We can see a clearer picture in this setup because when we increase the number of bins, we get to a peak point and then the accuracy decreases and in the end levels off as expected. The peak point is usually between 20 and 30 which we suggest to use for a 100 bin histogram representation. This shows us that the selection of number of bins is the critical point in this methodological setup and the proper selection of this parameter yields the best accuracy.

9.5.3 Discussions

In this work as the novel part of this study, we propose to adapt the random subspace method (Ho, 1998) to the creation of dissimilarity matrices and show that with the combination of these matrices, we achieve higher detection accuracies. Random subspaces have been used

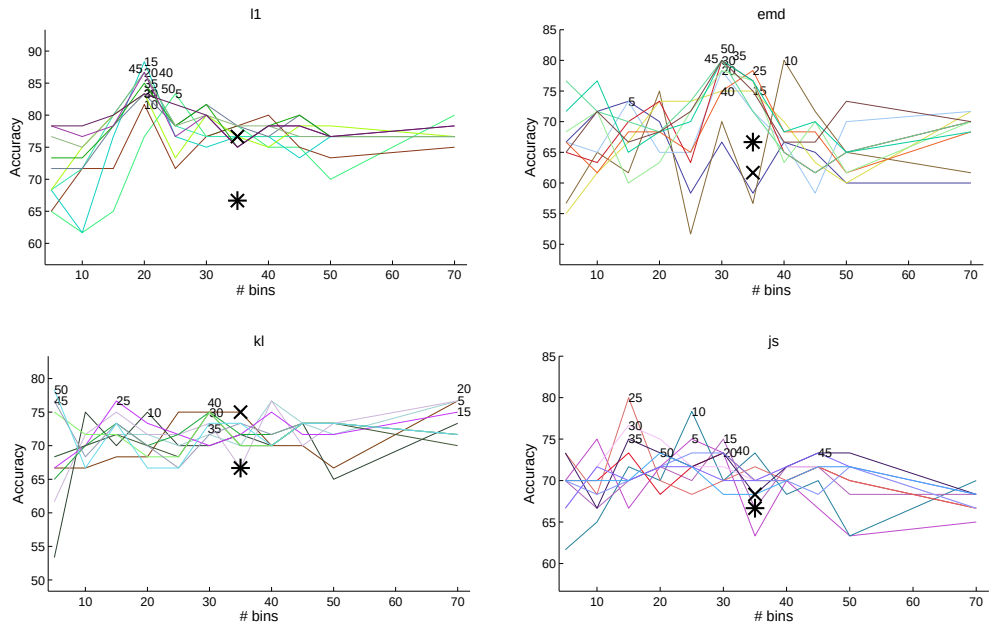

Figure 14: # of bins vs accuracy. The numbers in the plot show the number of random subspaces used. X is the accuracy on the full dissimilarity matrix of the corresponding dissimilarity measure and * is the base-line original space accuracy.

also in the dissimilarity based pattern recognition paradigm (Pekalska et al., 2000, 2006), but the method is applied after the dissimilarity matrices are constructed. The novelty of our method is that we use the random subspaces technique before creating the dissimilarity matrices to combine the useful information of the data.

In our previous work, we used the dissimilarity space and dissimilarity combination (Ulaş et al., 2011c) using intensity and apparent diffusion coefficient (ADC) based histogram representations. The main aim of that study was to show the effect of dissimilarity space using dissimilarity combination and multiple modalities. In this work, we use only one modality and one ROI; our aim is not to combine data from different sources but to get the best from the data as hand as was also targeted in (Pekalska et al., 2006). For this purpose, we use four different dissimilarity measures and analyze the effect of using the dissimilarity space and the dissimilarity space constructed using random subspaces.

We have seen that the dissimilarity space constructed using the dissimilarity measures mentioned in this work always outperform the base-line original space accuracies. Our novel contribution is the adaptation of Ho's random subspace methodology for selecting a subset of bins used in the computation of the dissimilarity matrices. With this method, we see that we can achieve significantly higher results when proper number of bins and random subspaces are selected. For the combination purpose, we compare combining using averaging and concatenation and we see that averaging usually outperforms concatenation (though the best result is achieved using concatenation) and we believe that this is due to the curse of dimensionality.

Our analysis of number of bins and number of random subspaces show that the important parameter is the number of bins. With relatively small number of random subspaces, one can achieve good results. As expected, when the number of subspaces increases, the individual dissimilarity matrices become less diverse and the combination accuracy converges. Same occurs with the number of bins, when we increase the number of bins, the diversity decreases and the accuracy converges. In this setup, the number of bins becomes more important because of the peak points. In this study we used a source histogram of 100 bins and we have observed that we get the peak accuracies in the 20-30 bin range. As a future study, we would like to apply this methodology to other data sets and artificial data sets with different number of source bins and draw a theoretical/experimental formula for the selection of the number of bins.

9.6 Classification using multiple instance learning

Multiple Instance Learning (MIL) (Dietterich et al., 1997) is a promising paradigm which is becoming more popular in the recent years. The idea is to represent objects by a set of descriptors instead of a single descriptor which is called a *bag*. This especially useful in cases where an image can be defined using several regions or a web page is represented using the web pages linking to it. The usual approach is to choose a representative of the bag by using *concepts* and apply classical pattern recognition algorithms. Recently, the Delft group developed a MIL algorithm which works using dissimilarities and on the dissimilarity space (Tax et al., 2011). In a previous work (Castellani et al., 2009), we used sparse point-based local features for brain classification and had promising results. Few and significant landmarks are detected and characterized by local region descriptors, estimated by the Scale Invariant Feature Transform (SIFT) operator (Lowe, 2004). After landmarks extraction, feature points are properly clusterized in order to obtain the *visual words*. As a future work, we would like to apply the multiple instance learning paradigm to these set of features and instead of clusterizing the feature points to obtain the *visual words*, we would like to define the feature points as multiple instances and apply the MIL paradigm and dissimilarity based MIL methodology proposed by Tax et al. (2011).

10 Case study 2: Brain classification by generative embeddings

In this section we present a method based on the generative embeddings. We use only the sMRI data and use spectral shape descriptors as data representation. We classify the score spaces created by the generative embeddings by using the IT kernels and single support vector machines.

In this case study, we use *Heat Kernel Signatures* to extract histogram based features from SMRI and use the generative embedding score spaces mentioned in Section 7 and apply IT kernels which were developed as part of WP2. The work will be submitted to the Pattern Recognition Journal and is a joint work of the Verona and Lisbon groups. In all the problems, we used average hold out methodology with 30 repetitions using stratification. For estimating the C value of the SVM and q value for the IT kernels, we used 5-fold cross validation on the training set. To estimate the number of topics, we used the Bayesian

Information Criterion (BIC) (Schwarz, 1979), which penalizes the likelihood with a penalty term on the number of free parameters in a way that larger models which do not increase the likelihood significantly are discouraged. In the pLSA model, the number of free parameters is calculated as $(D-1).Z + (W-1).Z + (Z-1)$. Then the BIC becomes:

$$BIC = \frac{1}{2}.((D-1).Z + (W-1).Z + (Z-1)).\log \sum_{j=1}^{W}\sum_{i=1}^{D} n(d_i, w_j)$$

10.1 Information Theoretic Kernels

Kernels on probability measures have been shown very effective in classification problems involving text, images, and other types of data (Cuturi et al., 2005; Jebara et al., 2004). Given two probability measures p_1 and p_2, representing two objects, several information theoretic kernels (ITKs) can be defined (Martins et al., 2009). The Jensen-Shannon kernel is defined as

$$k^{\,\mathrm{JS}}(p_1, p_2) = \ln(2) - JS(p_1, p_2), \tag{22}$$

with $JS(p_1, p_2)$ being the Jensen-Shannon divergence

$$JS(p_1, p_2) = H\left(\frac{p_1 + p_2}{2}\right) - \frac{H(p_1) + H(p_2)}{2}, \tag{23}$$

where $H(p)$ is the usual Shannon entropy.

The Jensen-Tsallis (JT) kernel is given by

$$k_q^{\,\mathrm{JT}}(p_1, p_2) = \ln_q(2) - T_q(p_1, p_2), \tag{24}$$

where $\ln_q(x) = (x^{1-q} - 1)/(1-q)$ is the q-logarithm,

$$T_q(p_1, p_2) = S_q\left(\frac{p_1 + p_2}{2}\right) - \frac{S_q(p_1) + S_q(p_2)}{2^q} \tag{25}$$

is the Jensen-Tsallis q-difference, and $S_q(r)$ is the Jensen-Tsallis entropy, defined, for a multinomial $r = (r_1, ..., r_L)$, with $r_i \geq 0$ and $\sum_i r_i = 1$, as

$$S_q(r_1, ..., r_L) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{L} r_i^q\right).$$

In (Martins et al., 2009), versions of these kernels applicable to unnormalized measures were also defined. Let $\mu_1 = \omega_1 p_1$ and $\mu_2 = \omega_2 p_2$ be two unnormalized measures, where p_1 and p_2 are the normalized counterparts (probability measures), and ω_1 and ω_2 arbitrary positive real numbers (weights). The weighted versions of the JT kernels are defined as follows:

- The weighted JT kernel (version A) is given by

$$k_q^A(\mu_1, \mu_2) = S_q(\pi) - T_q^\pi(p_1, p_2), \tag{26}$$

  where $\pi = (\pi_1, \pi_2) = \left(\frac{\omega_1}{\omega_1+\omega_2}, \frac{\omega_2}{\omega_1+\omega_2}\right)$ and

$$T_q^\pi(p_1, p_2) = S_q\left(\pi_1 p_1 + \pi_2 p_2\right) - \left(\pi_1^q S_q(p_1) + \pi_2^q S_q(p_2)\right).$$

- The weighted JT kernel (version B) is defined as

$$k_q^B(\mu_1, \mu_2) = \left(S_q(\pi) - T_q^\pi(p_1, p_2)\right)(\omega_1 + \omega_2)^q. \tag{27}$$

10.2 Proposed Approach

Once the generative model is estimated, the generative score spaces are calculated. Then the kernels are computed. Since results were similar, we omit the weighted JT kernel (version B) – we will refer to weighted JT kernel (version A) as JT-W.

The approach herein proposed consists in defining a kernel between two observed objects x and x' as the composition of the score function with one of the JT kernels presented above. Formally,

$$k(x, x') = k_q^i \left(\phi_\Theta(x), \phi_\Theta(x') \right), \tag{28}$$

where $i \in \{\text{JT, A, B}\}$ indexes one of the Jensen-Tsallis kernels (24), (26), or (27), and ϕ_Θ is one of the generative embeddings defined in Section 7. Notice that this kernel is well defined because all the components of ϕ_Θ^{FE} are non-negative.

We consider two types of kernel-based classifiers: K-NN and SVM. Recall that positive definiteness is a key condition for the applicability of a kernel in SVM learning. It was shown in (Martins et al., 2009) that k_q^A is a positive definite kernel for $q \in [0, 1]$, while k_q^B is a positive definite kernel for $q \in [0, 2]$. Standard results from kernel theory (Shawe-Taylor and Cristianini, 2004, Proposition 3.22) guarantee that the kernel k defined in (28) inherits the positive definiteness of k_q^i, thus can be safely used in SVM learning algorithms.

10.3 Results

We compare the results of our proposed approach with the linear kernel as a reference (which is the most used solution in the hybrid generative discriminative approach case, e.g. the Fisher Kernel). As classifiers we used Support Vector Machines and K-Nearest Neighbor (with K set to 1, i.e. the nearest neighbor rule). When possible, the classifiers have been applied also in the original domain (namely without the application of the generative embedding step).

Results are displayed in Table 12. In the table, "NN" stands for nearest neighbor results, while "SVM" refers to SVM results. "Linear" is the linear kernel, whereas "JS", "JT" and "JT-W" stands for Jensen-Shannon, Jensen-Tsallis and Weighted Jensen Tsallis kernels, respectively, as described in section 10.1. The acronyms of the generative embeddings follow the notation described in Section 7: "TPM-1" is the posterior topic mixture for a single pLSA, "TPM-2" is the posterior topic mixture starting from one pLSA per class,"FESS-1" is the Free Energy Score Space for a single pLSA, "FESS-2" is the Free Energy Score Space obtained starting from one pLSA per class, "LLR" is the Log Likelihood Ratio score space, "FSH" is the Fisher Score space, "TOP" is the TOP kernel score space and "PD" is the Posterior Divergence Score space. The standard errors of means, in all runs, were all less than 0.0252.

From the table different observations may be done:

- in almost all cases, the use of IT kernels over generative embeddings outperforms the linear kernel over the same embeddings, this being really evident in some cases
- at the same time the intermediate use of a generative embedding is almost always beneficial with respect to use the linear and the IT kernels on the original space

Embedding	Linear		JS		JT		JT-W	
	NN	SVM	NN	SVM	NN	SVM	NN	SVM
TPM-1	0.516	0.500	0.542	0.596	0.503	0.627	0.584	0.643
TPM-2	0.610	0.686	0.589	0.689	0.543	0.658	0.631	0.702
FESS-1	0.561	0.500	0.569	0.500	0.500	0.369	0.584	0.500
FESS-2	0.629	0.500	0.627	0.600	0.500	0.674	0.601	0.720
LLR	0.573	0.500	0.588	0.616	0.500	0.638	0.614	0.636
FSH	0.618	0.500	0.584	0.702	0.553	0.673	0.619	0.699
TOP	0.519	0.500	0.519	0.500	0.500	0.500	0.500	0.500
PD	0.752	0.500	0.748	0.500	0.627	0.806	0.726	0.808
ORIG	0.610	0.770	0.602	0.743	0.503	0.706	0.500	0.738

Table 12: Results on the Brain classification task. See the text for details.

- it is evident from the table that the best generative embedding is the very recently proposed Posterior Divergence Score Space. It seems this generative embedding has a slight preference to be used with the IT kernels
- there is not a significant difference among the various IT kernels, even if it may be argued that the Weighted Jensen Tsallis one is the most positive
- comparing the classifiers, there is not a so huge difference between the SVM and the Nearest Neighbor performances, thus confirming the goodness of the devised similarity measure

10.4 Work in progress

In this case study, we use *Heat Kernel Signatures* to extract histogram based features (see also (Castellani et al., 2011)) using different scales and using these as different sources for Multiple Kernel Learning paradigm. The data is extracted from sMRI scans of the left thalamus of 30 schizophrenic patients and 30 healthy controls. Several kernels are computed (i.e., one kernel per scale) and a set of weights are estimated for the kernel combination. In this fashion, we can choose the most discriminative scales by selecting those associated to the highest weights, and vice versa. Moreover, kernel combination leads to a new similarity measure which increases the classification accuracy. It is important to note that in our appraoch we aim at selecting the best shape characteristics for classification purposes, hence, our selection is driven by the performance of a Support Vector Machine (SVM) classifier.

10.4.1 Methodology

The contribution of geometric features extracted at each scale are combined by employing the MKL strategy as described in Section 8.2. Each shape representation r_i is associated to a kernel k_m by leading to $n = P$ kernels. Indeed, both the weights $(\eta_1 \cdots \eta_P)$ and the SVM parameters are estimated. In order to obtain the best classification accuracy according to the *max-margin* paradigm an *alternating* approach is used between the optimization of kernel weights and the optimization of the SVM classifier. In each step, given the current

solution of kernel weights, MKL solves a standard SVM optimization problem with the combined kernel. Then, a specific procedure is applied to update the kernel weights. Once the MKL procedure is completed, we obtain a two-fold advantage: i) we can select the best scale contributions by keeping only the scales associated to the highest weighs, and ii) we can compose a new kernel from the weighted contributions of the best scales, which can be evaluated for classification purposes.

10.4.2 Experimental protocol

In our experiments, we apply leave-one-out (LOO) cross-validation to assess the performance of the technique. Since LOO is used as the cross validation technique, we do not report standard deviations or variances. We compare our results using k-fold paired t-test at $p = 0.05$. We collect geometric features at 11 scales generating different shape representations $r_0 1, \cdots, r_1 1$. In practice, each representation r_i is a feature vector x_i which is plugged in the MKL framework. We employ the dot product as basic kernel function (i.e., linear kernel) since it avoids the estimation of free kernel parameters. Different strategies to combine the different shape representations have also been evaluated:

- Single Best Kernel (`Single-best`): an SVM is trained separately per each representation. Therefore, the performances of the classification are evaluated separately at each scale. So doing, we can evaluate the independent contributions coming from the different sources of information and select the best one.

- Feature concatenation (`SVM-con`): the contributions coming from the different sources are concatenated into a single feature vector. Then, a single SVM is employed for classification [6].

- Rule-based MKL (`RBMKL`): as baseline MKL approach, the so called rule-based method is evaluated: the kernels computed at each scale are combined by simply taking their average (i.e., $\forall m, \boldsymbol{\eta_m} = 1/P$).

- Simple MKL (`SMKL`): a simple but effective MKL algorithm is employed (Rakotomamonjy et al., 2008) by addressing the MKL problem through a weighted 2-norm regularization formulation with additional constraint on the weights that encourages sparse kernel combination. It is a popular approach and its code is publicly available[7].

- Group Lasso MKL (`GLMKL`): it denotes the group Lasso-based MKL algorithms proposed by (Kloft et al., 2011; Xu et al., 2010). A closed form solution for optimizing the kernel weights based on the equivalence between group-lasso and MKL is proposed. In our implementation, we used l_1-norm on the kernel weights and learned a convex combination of the kernels.

10.4.3 Results

The first evaluation scores are shown in Table 13, which reports the single-best kernel accuracies for all feature representations. We can observe that the best performance is obtained

[6] We use LIBSVM software (Chang and Lin, 2001) to train the SVM.
[7] http://asi.insa-rouen.fr

at 78.33 % using r02 which is shown as bold face in the table. The entries marked with "*" show the accuracies which are statistically significantly less accurate than the best algorithm using k-fold paired t-test at $p = 0.05$.

Table 13: Single-kernel SVM accuracies.

r01	r02	r03	r04	r05	r06	r07	r08	r09	r10	r11
75.00	**78.33**	76.67	76.67	73.33	*66.67	68.33	70.00	76.67	71.67	70.00

Second, concatenating the features in a single vector leads to 83.33 % accuracy. Third, using the proposed three different MKL algorithms, we combined the eleven kernels by introducing the weights $\boldsymbol{\eta_m}$. Table 14 reports the results of the best single-kernel SVM, the accuracy of the concatenated feature set, and the three MKL-based algorithms trained. The values in parentheses show the percentage of controls classified as schizophrenia and the percentage of patients classified as healthy respectively. We achieve an accuracy of 86.67%, reached by combining eleven kernels with the SMKL approach. This result is better than all other MKL settings and single-kernel SVMs. Further, GLMKL achieves 85% accuracy which is still higher than that reached by the feature concatenation method. We can also note that we cannot overcome SVM-con when we use RBMKL, as the latter gives equal weight to each kernel. In fact, if there are inaccurate representations in the given set, the overall mean combination accuracy may be less of that reached using the single best. Conversely, when the weights are automatically estimated, such as in SMKL and GLMKL the selection of the most reliable information is carried out by the MKL procedure and the overall performance improves.

Table 14: MKL accuracies.

SVM	SVM-con	RBMKL	SMKL	GLMKL
*78.3 (10, 11.7)	83.3 (8.3, 8.3)	*81.7 (10, 8.3)	**86.7 (6.7, 6.7)**	85.0 (8.3, 6.7)

In Figure 15, we plotted the weights of MKL for both SMKL and GLMKL algorithms. Note that the estimated weights are coherent in the two algorithms. As expected, the best representation is r02, which has the highest weights. Although the other representations with high weights (r08, r11 and r05) do not provide much accurate single-kernel SVMs results, their contributions to the overall accuracy in the combination is higher than those given by the other kernels. This demonstrates that when considering combinations, even a representation which does not lead to very precise results may contribute to raise the overall combination accuracy. Moreover, we can also deduce that these four representations are the most useful in discriminating between healthy and schizophrenic subjects, and we may focus the attention on these properties only.

Using this information, we also performed the above pipeline using only these four representations, and we can observe the results in Table 15. Using this subset, we get the highest accuracy with SMKL[8], reaching 88.33% of accuracy. We can also observe an increase in RBMKL.

[8]Note that in principle the same result should have been obtained automatically from MKL algorithms on all representations. In practice, this is not the case in our experiment due to the fact that the estimated solution is trapped into a local minimum.

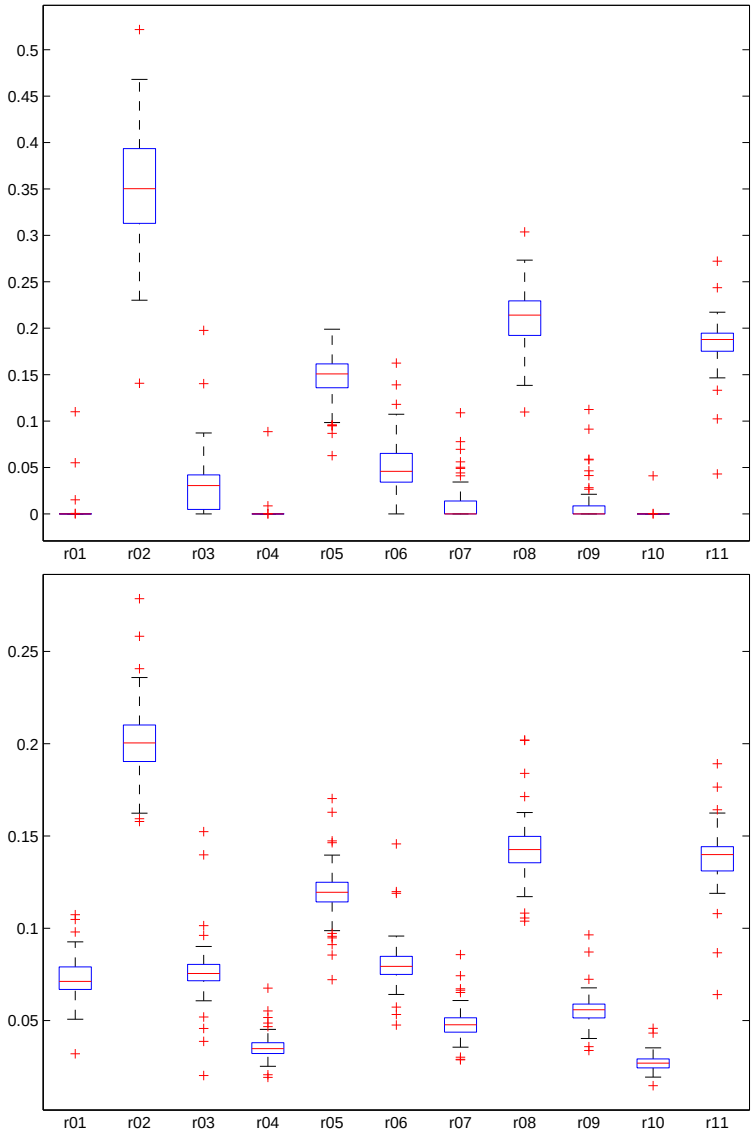

Figure 15: Combination weights in MKL using the linear kernel. Top: using `SMKL`, Bottom: using `GLMKL`.

Table 15: MKL accuracies on the selected subset of representations.

SVM	SVM-con	RBMKL	SMKL	GLMKL
*78.3 (10, 11.7)	*83.3 (6.7, 10)	*83.3 (6.7, 10)	**88.3 (6.7, 5)**	85.0 (6.7, 8.3)

10.4.4 Discussion

In this work, we have shown in general that MKL algorithms perform better than both single-best kernel SVMs and feature concatenation strategies. We have also observed that `RBMKL` (which does not compute weights while combining kernels) does not outperform the feature concatenatenation approach. Conversely, when the kernel combination is carried out by estimating proper weights, a drastic improvement is instead obtained. The kernel weights also allow us to extract useful information: it is interesting to observe that, for both MKL algorithms with the highest accuracy, four representations have the maximum effect (i.e., the highest weights), i.e., `r02`, `r08`, `r11`, and `r05`, with `r02` being the best single-kernel. We use this information to select a smaller number of representations to reduce the costs of the feature extraction phase. Finally, we can also observe that by using such subset we can reach the best accuracy overall.

11 Conclusions and future work

In this project, we have defined a set of new approaches to deal with schizophrenia detection from MRI images. We have proposed a working pipeline which takes into account different aspects of the disease. We have successfully exploited dissimilarity-based techniques developed in other WPs to our medical application. In particular, we have shown that brain classification on dissimilarity space reaches a drastic improvement over standard feature-based approaches. Moreover, we have shown that combining dissimilarities represents a natural and effective approach to merge different sources of information. In this fashion we were able to exploit complementary information about different parts of the brain, different acquisition modalities, and different brain properties. Moreover, we have shown that our new paradigm to define data descriptors by generative embedding is very effective and works well on our medical application. This research has opened new perspectives in the medical application of this WP which have been envisaged by our work in progress. In particular, we have shown that a further improvement can be obtained by adapting random subspace method to create the dissimilarity space.

Another promising direction is to use Multiple Instance Learning (MIL). In our previous studies, we reached promising results using SIFT features by clustering the visual words. As a future work, we would like to work on MIL paradigm following the work of Delft group on dissimilarity based MIL.

Furthermore, we are working on employing advanced dissimilarity-based techniques to encode shape properties. Our preliminary results have shown a drastic improvement by using Multiple Kernel Learning to improve the diffusion based shape description.

Nevertheless, some issues have remained open. Direct use of dissimilarity measures was not successful (See Appendix D). We applied the Iterative Closest Point (ICP) distances

(Zhang, 1994) and used non-rigid registration and extracted distances based on the deformation field. Although preliminary experiments did not have promising results, different registration procedures can be exploited to improve this approach. Moreover, also the use of Multiple Kernel Learning to combine dissimilarities was not a success. A further investigation in this direction will be exploited.

Finally, we have shown in our experiments that DWI data was important to improve the classification accuracy when multimodal approach was employed. This encourages us to exploit more advanced imaging techniques such as Diffusion Tensor MRI or Functional MRI to further improve schizophrenia detection.

During this project we have strongly collaborated with other partners. We have used the dissimilarity combination paradigm from the Delft group and published two papers: (Ulaş et al., 2010b, 2011c). We have collaborated with the Lisbon group using IT kernels and generative embeddings and we are about to submit a paper to the Pattern Recognition journal. Although our collaborations with the York group and Venice group (especially on graph distances, skeletons, and brain registrations) did not bear fruit, the exchange of information shall guide us in our future studies. We have collaborated with the Zurich group using WP6 data and published four papers ((Bicego et al., 2011; Ulaş et al., 2011d; Gönen et al., 2011; Schüffler et al., 2011)). One of these papers ((Bicego et al., 2011)) was also in collaboration with the Lisbon group.

12 Appendices

A Publications

Below we list the publications of our group during the course of the project on WP7 and WP6.

WP7

- Aydın Ulaş, Robert P. W. Duin, Umberto Castellani, Marco Loog, Manuele Bicego, Vittorio Murino, Marcella Bellani, Stefania Cerruti, Michele Tansella, Paolo Brambilla
  Dissimilarity-based Detection of Schizophrenia,
  International Journal of Imaging Systems and Technology, Vol. 21, No. 2, pp:179-192, 2011.

- Aydın Ulaş, Umberto Castellani, Vittorio Murino, Marcella Bellani, Michele Tansella, Paolo Brambilla,
  Heat Diffusion Based Dissimilarity Analysis for Schizophrenia Classification,
  PRIB 2011, LNBI Vol: 7036, pp: 306-317, November 2-4, Delft, The Netherlands.

- Umberto Castellani, Pasquale Mirtuono, Vittorio Murino, Marcella Bellani, Michele Tansella, Paolo Brambilla,
  A new shape diffusion descriptor for brain classification,
  MICCAI 2011, accepted.

- Umberto Castellani, Aydın Ulaş, Vittorio Murino, Marcella Bellani, Michele Tansella, Paolo Brambilla,
  Selecting scales by Multiple Kernel Learning for shape diffusion analysis,
  MICCAI 2011, Workshop on "Mathematical Foundations of Computational Anatomy", pp:148-158, September 18-22, Toronto, Canada.

- Aydın Ulaş, Umberto Castellani, Pasquale Mirtuono, Manuele Bicego, Vittorio Murino, Stefania Cerruti, Marcella Bellani, Manfredo Atzori, Gianluca Rambaldelli, Michele Tansella, Paolo Brambilla,
  Multimodal Schizophrenia Detection by Multiclassification Analysis,
  CIARP 2011, LNCS Vol: 7042, pp: 491-498, November 15-18, Pucon, Chile.

- Umberto Castellani, Alessandro Perina, Vittorio Murino, Marcella Bellani, Gianluca Rambaldelli, Michele Tansella, Paolo Brambilla
  Brain Morphometry by Probabilistic Latent Semantic Analysis
  MICCAI 2010, LNCS Vol: 6362, pp: 177-184, 2010.

- Aydın Ulaş, Robert P. W. Duin, Umberto Castellani, Marco Loog, Manuele Bicego, Vittorio Murino, Marcella Bellani, Stefania Cerruti, Michele Tansella, Paolo Brambilla
  Dissimilarity-based Detection of Schizophrenia,
  ICPR 2010, Workshop on "Brain Decoding: Pattern Recognition Challenges in FMRI Neuroimaging", pp: 32-35, August 22, 2010, Istanbul, Turkey.

- Umberto Castellani, Elisa Rossato, Vittorio Murino, Marcella Bellani, Gianluca Rambaldelli, Michele Tansella, Paolo Brambilla,
  Local Kernels for Brains Classification in Schizophrenia,
  AI*IA 2009, LNAI Vol: 5883, pp: 112-121, 2009.

- Dong Seon Cheng, Manuele Bicego, Umberto Castellani, MArco Cristani, Stefania Cerruti, Marcella Bellani, Gianluca Rambaldelli, Manfredo Atzori, Paolo Brambilla, Vittorio Murino,
  A hybrid generative/discriminative method for classification of regions of interest in schizophrenia brain MRI,
  MICCAI 2009, Workshop on "Probabilistic Models for Medical Image Analysis", pp: 174-184, 2009.

- Dong Seon Cheng, Manuele Bicego, Umberto Castellani, Stefania Cerruti, Marcella Bellani, Gianluca Rambaldelli, Manfredo Atzori, Paolo Brambilla, Vittorio Murino,
  Schizophrenia classification using regions of interest in brain MRI,
  IDAMAP 2009, pp: 47-52.

WP6

- Manuele Bicego, Aydın Ulaş, Peter J. Schüffler, Umberto Castellani, Pasquale Mirtuono, Vittorio Murino, André Martins, Pedro M. Q. Aguiar, Mário A. T. Figueiredo, Renal Cancer Cell Classification Using Generative Embeddings and Information Theoretic Kernels,
  PRIB 2011, November 2-4, Delft, The Netherlands.

- Mehmet Gönen, Aydın Ulaş, Peter J. Schüffler, Umberto Castellani, Vittorio Murino, Combining Data Sources Nonlinearly for Cell Nucleus Classification of Renal Cell Carcinoma,
  1st International Workshop on Similarity-Based Pattern Analysis, 2011, LNCS Vol: 7005, pp: 250-260, September 28-30, Venice, Italy.

- Aydın Ulaş, Peter J. Schüffler, Manuele Bicego, Umberto Castellani, Vittorio Murino, Hybrid Generative-Discriminative Nucleus Classification of Renal Cell Carcinoma,
  1st International Workshop on Similarity-Based Pattern Analysis, 2011, LNCS Vol: 7005, pp:77-88, September 28-30, Venice, Italy.

- Peter J. Schüffler, Aydın Ulaş, Umberto Castellani, Vittorio Murino,
  A Multiple Kernel Learning Algorithm for Cell Nucleus Classification of Renal Cell Carcinoma,
  ICIAP 2011, LNCS Vol: 6978, pp:413-422, September 14-16, Ravenna, Italy.

B Data Set

The data set involves a 64 patient database cared by the Research Unit on Brain Imaging and Neuropsychology (RUBIN) at the Department of Medicine and Public Health-Section of Psychiatry and Clinical Psychology of the University of Verona. The data set is composed of MRI brain scans of 64 patients recruited from the area of South Verona (i.e., 100,000 inhabitants) through the South Verona Psychiatric Case Register (Tansella and Burti, 2003).

All had received a diagnosis of schizophrenia according to the criteria of the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (American Psychiatric Association, 1994) and were being treated by the South Verona Community-based Mental Health Service (for a detailed description please refer to Andreone et al. (2007)) and by other clinics reporting to the South Verona Psychiatric Care Register (Amaddeo and Tansella, 2009). Diagnoses for schizophrenia were obtained using the Item Group Checklist of the Schedule for Clinical Assessment in Neuropsychiatry (World Health Organization, 1992), administered by research clinical psychologists who had extensive experience with it. They were required to show inter-rater reliability both blindly and independently with those of a senior investigator also trained in the procedure by achieving similar diagnosis for at least 8 of 10 assessments.

Moreover, the psycho-pathological item groups completed by the two raters were compared, in order to discuss any major symptom discrepancies. The reliability of the IGC-SCAN diagnoses was also ensured by holding regular consensus meetings with the psychiatrists treating the patients and a senior investigator. The Italian version of the SCAN was edited by the RUBIN group (World Health Organization, 1996), and our investigators attended specific training courses in order to learn how to administer the IGC-SCAN.

Subsequently, diagnoses for schizophrenia according to the DSM-IV criteria were corroborated by the clinical consensus of two staff psychiatrists. Patients with comorbid psychiatric disorders, alcohol or substance abuse within the 6 months preceding the study, history of traumatic head injury with loss of consciousness, epilepsy or other neurological diseases were excluded.

All but two patients were receiving antipsychotic medication at the time of imaging. More specifically, 25 patients were on typical antipsychotic drugs (16 on haloperidol, three

on chlorpromazine, two on fluphenazine, two on clotiapine, one on thioridazine, one on zuclopenthixol) and 45 on atypical antipsychotic medication (25 on olanzapine, nine on clozapine, nine on risperidone, two on quetiapine). Patients' clinical information was retrieved from psychiatric interviews, the attending psychiatrist and medical charts. The Brief Psychiatric Rating Scale (24-item version) (Ventura et al., 2000) was used to characterize clinical symptoms. Again, it was administered by trained research clinical psychologists following the same reliability procedure as outlined above for the IGC-SCAN.

Additionally, 60 individuals without schizophrenia (control subjects) were also recruited. They had no DSM-IV axis I disorders, as determined by a modified version of the Structured Clinical Interview for DSM-IV nonpatient version. As well, they had no history of psychiatric disorders among first-degree relatives, no history of alcohol or substance abuse and no current major medical illness. Typical control subjects were hospital / university staff volunteers or individuals undergoing imaging for dizziness whose MRI showed no evidence of central nervous system abnormalities when reviewed by the neuroradiologist. Any dizziness was due to benign paroxysmal positional vertigo or to nontoxic labyrinthitis. Participants in the control group were scanned only after a full medical history was taken and general neurological, otoscopic, and physical examinations were carried out; they had completely recovered from the dizziness. Also, none of these participants was taking medication, including drugs for nausea or vertigo.

This research study was approved by the Biomedical Ethics Committee of the Azienda Ospedaliera of Verona. All participants provided signed informed consent after they understood all aspects of study participation.

Table 16 shows relevant demographic and clinical characteristics of both groups.

Table 16: Some demographic and clinical characteristics of the study groups. The Student's t-test of the age means rejects (at a two-tailed significance level of $p < 0.05$) the hypothesis that the study groups are significantly different in age, and Pearson χ^2 confirms the same for the gender differences.

Characteristic	Group mean (and SD)*		Statistics		
	Control $n = 60$	Schizophrenia $n = 64$	Test	df	p
Age, yr	39.95 (11.25) [23-60]	38.84 (11.96) [18-62]	$t = 0.53$	122	0.60
Male/female	32/28	43/21	$\chi^2 = 2.49$	1	0.11
Age at onset, yr		26.28 (9.17)			
Duration of illness, yr		13.37 (10.30)			

SD = standard deviation; df = degrees of freedom; p = value of significance
* Unless otherwise indicated.

C Guidelines for ROI Tracing

C.1 Hippocampus

The first slice to be traced was the one where the superior colliculus completely connected with the thalamus bilaterally. Moving anteriorly through the brain, we traced around the hippocampi using first the corona radiata, and then the ambient cistern as the superior border. The white matter acted as the inferior border, and the inferior horn of the lateral ventricle as the lateral one. The anterior limit was one slice posterior to the slice where the mamillary body became visible. On average, 16 slices were traced.

C.2 Amygdala

The first slice to be traced was the one where the mamillary body becomes visible. The superior and lateral borders were defined by the temporal lobe white matter, and the inferior one by the parahippocampal gyrus white matter. Moving forward, the anterior limit, either right or left, was defined by the point when the amygdale became too diffuse to be resolved from the temporal lobe gray matter. In average, seven slices were traced.

C.3 Entorhinal Cortex

The entorhinal cortex was traced on MRI coronal slices. The frontotemporal stem delimited the region of interest anteriorly. The intersection of the line along the grey-white junction with the medial bank of the collateral sulcus defined the inferolateral border. The superomedial border was defined rostrally by the sulcus semiannularis and caudally by the uncal cleft. The intersection of the line along the grey-white junction with the cortical surface was used to improve the definition of these structures. The most anterior slice in which the body of the hippocampus first became clearly visible was chosen as the posterior limit. It should be noted that the prior methods were slightly modified since the lateral geniculate body was poorly detectable in most of our scan and therefore it was not to used as a posterior limit. Also, the perirhinal cortex was included in our tracing.

C.4 Dorsolateral Prefrontal Cortex

The DLPFC was defined as slices anterior to the posterior border of the genu till the anterior border of the the Sylvian horizontal ramus; the superior border was the superior frontal sulcus, the inferior border was the upper border of the Sylvian fissure posteriorly and the horizontal ramus of the Sylvian fissure anteriorly, the lateral boundary was the edge of the brain, and the medial boundary was the line connecting the most medial point of the superior frontal sulcus and the Sylvian fissure/horizontal ramus .

C.5 Thalamus

The tracing of the thalamus was performed on the T1-weighted MP-RAGE sequence, beginning at the coronal slice where the anterior pillars of the fornix merge into the mammillary

bodies and continuing to the slice in which it was no longer possible to distinguish the thalamus from the surrounding brain matter. The lateral ventricles at the superior border, the red nucleus and the substantia nigra at the inferior border, the posterior limb of the internal capsule at the lateral border separating the thalamus from the adjacent lentiform nucleus and the third ventricle at the medial border demarcated the limits of the thalamus. The presence of the adhesio interthalamica was also detected.

C.6 Superior Temporal Gyrus

Superior temporal gyrus (STG) was traced bilaterally in the coronal plane. The anterior border was defined by the first slice where the temporal stem appeared. Posteriorly, it was traced to the end of the Sylvian fissure. The superior border was the Sylvian fissure and the inferior one was the superior temporal sulcus.

C.7 Heschls Gyrus

The HG was anatomically identified as an omega or heart-shaped protrusion in the supratemporal plane. It is defined medially by the first transverse sulcus of temporal lobe and laterally by Heschl's sulcus. If there are two complete Heschl's sulci defining two gyri, then the anterior gyrus was used.

D Discontinued and Inconclusive Works

Since computing histograms of intensities and calculating dissimilarities from these features involves preprocessing of data and feature intermediation, we also prepared some dissimilarity matrices based on direct computation of distances between brains.

D.1 Dissimilarities based on Iterative Closest Point distance

To extract dissimilarities directly without pre-processing or without resorting to intermediate feature representations, we first tried to use the Iterative Closest Point (ICP) algorithm (Zhang, 1994). We computed the dissimilarities between pairwise brains using the ICP algorithm. Preliminary results can be seen in Table 17. Because of the discouraging results, we left this line of work.

D.2 Dissimilarities based on registration of MRIs

Second, we applied dissimilarity computation based on deformable registration of brain MRIs for each subject pair. For every pair of subjects we used the CAMP[9] software to register the subjects and calculated dissimilarities based on the deformation field using the ideas presented in (Klein et al., 2010) by computing the standard deviation of the log Jacobian of the deformation field. We used two measures for calculating the dissimilarity between two subjects based on the deformation field:

[9]The software is available from http://www.mrf−registration.net

Table 17: 1nn results of dissimilarities based on ICP distances (%).

	Accuracy
lamyg	53.23
ramyg	48.39
ldlpfc	N/A
rdlpfc	N/A
lec	44.35
rec	49.19
lhg	55.65
rhg	48.39
lhippo	53.23
rhippo	62.10
lstg	46.77
rstg	52.42
lthal	50.81
rthal	46.77

- The sum of the deformations on every point (i, j, k) summed over all points.
- The length of the deformations on every point (i, j, k) summed over all points.

The registration was carried out on the whole MRIs and the portion that concerns every ROI has been extracted to calculate the dissimilarity between the ROIs of each pair of subjects. In Table 18 we can see the $1nn$ results of the dissimilarity matrices using registration based dissimilarities using LOO methodology. Because of the extensive time and disk space needed to create the registrations, tune the parameters and save the pair wise registration results; and the unpromising preliminary results on the deformation fields, we discontinued this line of work.

D.3 Dissimilarity combination using MKL

In another line of study, instead of normalizing and taking the average of dissimilarity matrices on the dissimilarity space, we used the multiple kernel learning methodology to combine different dissimilarity matrices. The first basic advantage of this approach is that one does not need to normalize the dissimilarity matrices because the inherent optimization procedure included in MKL is going to assign the weights accordingly. Second, now that we have weights, the importance of each matrix can be seen to extract useful information.

For the preliminary experiments, we used the same data as presented in Section 9 ((Ulaş et al., 2011c)). We combined ROIs using the same representation and distance measure and using the linear kernel as the base kernel. We can see the results in Table 19. Except *hist-chi* on SMRI, the results are not promising but require further investigation.

Table 18: 1nn results of dissimilarities based on deformable registration (%).

	sum of deformations	sum of length of deformations
lamyg	49.19	50.00
ramyg	49.19	47.58
ldlpfc	56.45	57.26
rdlpfc	50.81	49.19
lec	48.39	52.42
rec	50.00	49.19
lhg	52.42	47.58
rhg	57.26	55.65
lhippo	53.23	53.23
rhippo	50.00	50.00
lstg	53.23	50.81
rstg	56.45	52.42
lthal	44.35	43.55
rthal	46.77	51.61
icv	55.65	53.23
whole-1	53.23	56.45
whole-2	58.87	58.06

Table 19: Combination of dissimilarity matrices using MKL.

	SMRI	ADC	MCUR	SH
hist-euclid	64.04	58.77	50.88	50.00
hist-l1	64.04	60.53	50.00	54.39
hist-intersect	63.16	57.89	46.49	46.49
hist-diffusion	64.04	60.53	50.00	54.39
hist-chi	**77.19**	56.14	42.98	63.16
hist-emd	48.25	35.96	50.00	50.88
pdf-euclid	50.00	50.00	57.89	50.88
pdf-l1	62.28	61.40	50.00	54.39
pdf-emd	66.67	54.39	45.61	64.91
pdf-bs	66.67	56.14	49.12	46.49
pdf-kl	63.16	58.77	47.37	44.74
pdf-kl-orig	68.42	47.37	50.88	55.26
pdf-js	65.79	55.26	47.37	55.26

References

Alpaydın, E., 2004. Introduction to machine learning. The MIT Press.

Amaddeo, F., Tansella, M., 2009. Information systems for mental health. Epidemiologia e Psichiatria Sociale 18 (1), 1–4.

American Psychiatric Association, 1994. Diagnostic and statistical manual of mental disorders, DSM-IV. Washington DC, 4th Edition.

Andreone, N., Tansella, M., Cerini, R., Versace, A., Rambaldelli, G., Perlini, C., Dusi, N., Pelizza, L., Balestrieri, M., Barbui, C., Nose, M., Gasparini, A., Brambilla, P., 2007. Cortical white-matter microstructure in schizophrenia. diffusion imaging study. British Journal of Psychiatry 191, 113–119.

Ashburner, J., Friston, K. J., 2000. Voxel-based morphometrythe methods. NeuroImage 11 (6), 805–821.

Awate, S. P., Yushkevich, P., Song, Z., Licht, D., Gee, J. C., 2009. Multivariate high-dimensional cortical folding analysis, combining complexity and shape, in neonates with congenital heart disease. In: Proceedings of the 21st International Conference on Information Processing in Medical Imaging, IPMI '09. pp. 552–563.

Bach, F. R., Lanckriet, G. R. G., Jordan, M. I., 2004. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning. pp. 41–48.

Baiano, M., Perlini, C., Rambaldelli, G., Cerini, R., Dusi, N., Bellani, M., Spezzapria, G., Versace, A., Balestrieri, M., Mucelli, R. P., Tansella, M., Brambilla, P., 2008. Decreased entorhinal cortex volumes in schizophrenia. Schizophrenia Research 102 (1–3), 171–180.

Bellani, M., Brambilla, P., 2008. The use and meaning of the continuous performance test in schizophrenia. Epidemiologia e Psichiatria Sociale 17 (3), 188–191.

Bicego, M., Lovato, P., Ferrarini, A., Delledonne, M., 2010a. Biclustering of expression microarray data with topic models. In: Proceedings of the International Conference on Pattern Recognition. pp. 2728–2731.

Bicego, M., Lovato, P., Oliboni, B., Perina, A., 2010b. Expression microarray classification using topic models. In: Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10. New York, NY, USA, pp. 1516–1520.

Bicego, M., Pekalska, E., Tax, D. M. J., Duin, R. P. W., November 2009. Component-based discriminative classification for hidden markov models. Pattern Recognition 42, 2637–2648.

Bicego, M., Ulaş, A., Schüffler, P. J., Castellani, U., Mirtuono, P., Murino, V., André Martins, P. M. Q. A., Figueiredo, M. A. T., November 2011. Renal cancer cell classification using generative embeddings and information theoretic kernels. In: et al., M. L. (Ed.), IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB '11. Vol. 7036 of Lecture Notes in Bioinformatics. Springer Berlin / Heidelberg, p. accepted.

Bosch, A., Zisserman, A., Munoz, X., 2006. Scene classification via pLSA. In: Proceedings of the European Conference on Computer Vision, ECCV '06. pp. 517–530.

Brambilla, P., Tansella, M., 2007. Can neuroimaging studies help us in understanding the biological causes of schizophrenia? International Review of Psychiatry 19 (4), 313–314.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

Bronstein, A. M., Bronstein, M. M., 2011. Shape recognition with spectral distances. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (5), 1065–1071.

Browne, A., Jakary, A., Vinogradov, S., Fu, Y., Deicken, R., 2008. Automatic relevance determination for identifying thalamic regions implicated in schizophrenia. IEEE Transactions on Neural Networks 19 (6), 1101–1107.

Castellani, U., Mirtuono, P., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P., 2011. A new shape diffusion descriptor for brain classification. In: Proceedings of the International Conference on Medical Image Computing, MICCAI '11. p. accepted.

Castellani, U., Perina, A., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P., 2010. Brain morphometry by probabilistic latent semantic analysis. In: Proceedings of the international conference on Medical image computing and computer-assisted intervention, MICCAI '10. MICCAI. pp. 177–184.

Castellani, U., Rossato, E., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P., 2009. Local kernel for brains classification in schizophrenia. In: AI*IA '09:: Proceedings of the XIth International Conference of the Italian Association for Artificial Intelligence Reggio Emilia on Emergent Perspectives in Artificial Intelligence. Springer-Verlag, Berlin, Heidelberg, pp. 112–121.

Cha, S.-H., Srihari, S. N., 2002. On measuring the distance between histograms. Pattern Recognition 35 (6), 1355–1370.

Chang, C. C., Lin, C. J., 2001. LIBSVM: a library for support vector machines.
URL `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Cheng, D. S., Bicego, M., Castellani, U., Cerruti, S., Bellani, M., Rambaldelli, G., Atzori, M., Brambilla, P., Murino, V., 2009a. Schizophrenia classification using regions of interest in brain MRI. Tech. rep., Dipartimento di Informatica, University of Verona, Italy.

Cheng, D. S., Bicego, M., Castellani, U., Cerruti, S., Bellani, M., Rambaldelli, G., Atzori, M., Brambilla, P., Murino, V., 2009b. Schizophrenia classification using regions of interest in brain mri. In: Proceedings of Intelligent Data Analysis in Biomedicine and Pharmacology, IDAMAP '09. pp. 47–52.

Corradi-DellAcqua, C., Tomelleri, L., Bellani, M., Rambaldelli, G., Cerini, R., Pozzi-Mucelli, R., Balestrieri, M., Tansella, M., Brambilla, P., 2011. Thalamic-insular dysconnectivity in schizophrenia: Evidence from structural equation modeling. Human Brain Mapping, in press.

Cortes, C., Mohri, M., Rostamizadeh, A., 2010. Learning non-linear combinations of kernels. In: Advances in Neural Information Processing Systems 22. pp. 396–404.

Cristani, M., Perina, A., Castellani, U., Murino, V., 2008. Geo-located image analysis using latent representations. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8.

Cuturi, M., Fukumizu, K., Vert, J.-P., December 2005. Semigroup kernels on measures. Journal of Machine Learning Research 6, 1169–1198.

Davatzikos, C., 2004. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. NeuroImage 23 (1), 17–20.

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 (Methodological) 39 (1), 1–38.

Dietterich, T. G., Lathrop, R. H., Lozano-Pérez, T., January 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89, 31–71.

Duda, R. O., Hart, P. E., Stork, D. G., 2000. Pattern classification, 2nd Edition. Wiley-Intersciences.

Duin, R. P. W., 2005. Prtools, a matlab toolbox for pattern recognition version 4.0.14. available at http://www.prtools.org/.
URL `http://www.prtools.org/`

Edelstein, W. A., Bottomley, P. A., Pfeifer, L. M., 1984. A signal-to-noise calibration procedure for NMR imaging systems. Medical Physics 11, 180–185.

Fan, Y., Shen, D., Gur, R. C., Gur, R. E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. on Medical Imaging 26 (1), 93–105.

Frey, B. J., Jojic, N., 2005. A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (9), 1392–1416.

Gebal, K., Baerentzen, J. A., Aanaes, H., Larsen, R., 2009. Shape analysis using the auto diffusion function. In: Proceedings of the Symposium on Geometry Processing. Eurographics Association, Berlin, Germany, pp. 1405–1413.

Gerig, G., Styner, M., Shenton, M. E., Lieberman, J. A., 2001. Shape versus size: Improved understanding of the morphology of brain structures. In: Proceedings of the International Conference on Medical Image Computing, MICCAI '01. pp. 24–32.

Giuliani, N. R., Calhouna, V. D., Pearlson, G. D., Francis, A., Buchanan, R. W., 2005. Voxel-based morphometry versus region of interest: a comparison of two methods for analyzing gray matter differences in schizophrenia. Schizophrenia Research 74 (2–3), 135–147.

Gönen, M., Alpaydın, E., 2008. Localized multiple kernel learning. In: Proceedings of the 25th International Conference on Machine Learning. pp. 352–359.

Gönen, M., Ulaş, A., Schüffler, P. J., Castellani, U., Murino, V., September 2011. Combining data sources nonlinearly for cell nucleus classification of renal cell carcinoma. In: Pelillo, M., Hancock, E. R. (Eds.), Proceedings of the International Workshop on Similarity-Based Pattern Analysis, SIMBAD '11. Vol. 7005 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 250–260.

Ho, T. K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8), 832–844.

Hofmann, T., 2000. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In: Proceedings of the conference on advances in neural information processing systems, NIPS '02. pp. 914–920.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42 (1–2), 177–196.

Jaakkola, T. S., Haussler, D., 1998. Exploiting generative models in discriminative classifiers. In: Proceedings of the conference on advances in neural information processing systems, NIPS '98. Vol. 11. Cambridge, MA, USA, pp. 487–493.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., Hinton, G. E., 1991. Adaptive mixtures of local experts. Neural Computation 3, 79–87.

Jager, F., Hornegger, J., January 2009. Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging. IEEE Transactions on Medical Imaging 28 (1), 137–150.

Jebara, T., Kondor, R., Howard, A., December 2004. Probability product kernels. Journal of Machine Learning Research 5, 819–844.

Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S.-Y., Nakamura, K., Matsui, M., Sumiyoshi, T., Seto, H., Kurachi, M., 2007. Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. NeuroImage 34 (1), 235–242.

Kittler, J., Hatef, M., Duin, R. P. W., Matas, J., 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3), 226–239.

Klein, S., Loog, M., van der Lijn, F., den Heijer, T., Hammers, A., de Bruijne, M., van der Lugt, A., Duin, R. P., Breteler, M. M. B., Niessen, W. J., 2010. Early diagnosis of dementia based on intersubject whole-brain dissimilarities. In: Proceedings of the 2010 IEEE international conference on Biomedical imaging: from nano to Macro, ISBI'10. pp. 249–252.

Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A., 2011. l_p-norm multiple kernel learning. Journal of Machine Learning Research 12, 953–997.

Koenderink, J. J., van Doorn, A. J., October 1992. Surface shape and curvature scales. Image and Vision Computing 10, 557–565.

Kuncheva, L. I., 2004. Combining pattern classifiers: methods and algorithms. Wiley-Interscience.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. E., Jordan, M. I., 2004. Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research 5, 27–72.

Lee, W.-J., Duin, R. P. W., Loog, M., Ibba, A., june 2010. An experimental study on combining euclidean distances. In: Cognitive Information Processing (CIP), 2010 2nd International Workshop on. pp. 304–309.

Lewis, D. P., Jebara, T., Noble, W. S., 2006. Nonstationary kernel combination. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 553–560.

Li, X., Lee, T. S., Liu, Y., 2011. Hybrid generative-discriminative classification using posterior divergence. In: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR '11. pp. 2713–2720.

Ling, H., Okada, K., 2006. Diffusion distance for histogram comparison. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '06. Vol. 1. pp. 246–253.

Liu, Y., Teverovskiy, L., Carmichael, O., Kikinis, R., Shenton, M., Carter, C. S., Stenger, V. A., Davis, S., Aizenstein, H., Becker, J. T., Lopez, O. L., Meltzer, C. C., 2004. Discriminative mr image feature analysis for automatic schizophrenia and alzheimer's disease classification. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention, MICCAI '04. pp. 393–401.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2), 91–110.

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., Figueiredo, M. A. T., 2009. Nonextensive information theoretic kernels on measures. Journal of Machine Learning Research 10, 935–975.

Ng, A. Y., Jordan, M. I., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: Proceedings of the conference on advances in neural information processing systems, NIPS '02. Vol. 14. pp. 841–848.

Nyúl, L. G., Udupa, J. K., Zhang, X., 2000. New variants of a method of mri scale standardization. IEEE Transactions on Medical Imaging 19 (2), 143–150.

Pekalska, E., Duin, R. P. W., 2005. The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific, Singapore.

Pekalska, E., Duin, R. P. W., Paclík, P., 2006. Prototype selection for dissimilarity-based classifiers. Pattern Recognition 39 (2), 189–208.

Pekalska, E., Skurichina, M., Duin, R. P. W., 2000. Combining fisher linear discriminants for dissimilarity representations. In: Proceedings of the First International Workshop on Multiple Classifier Systems. MCS '00. Springer-Verlag, London, UK, pp. 117–126.

Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N., 2009a. Free energy score space. In: Proceedings of the conference on advances in neural information processing systems, NIPS '09. Vol. 22. pp. 1428–1436.

Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N., 29 2009-oct. 2 2009b. A hybrid generative/discriminative classification framework based on free-energy terms. In: Proceedings of the IEEE International Conference on Computer Vision, ICCV '09. pp. 2058–2065.

Pohl, K. M., Sabuncu, M. R., 2009. A unified framework for mr based disease classification. In: IPMI '09: Proceedings of the 21st International Conference on Information Processing in Medical Imaging. pp. 300–313.

Pruessner, J., Li, L., Serles, W., Pruessner, M., Collins, D., Kabani, N., Lupien, S., Evans, A., 2000. Volumetry of hippocampus and amygdala with high-resolution mri and three-dimensional analysis software: Minimizing the discrepancies between laboratories. Cerebral Cortex 10 (4), 433–442.

Rakotomamonjy, A., Bach, F. R., Canu, S., Grandvalet, Y., 2008. SimpleMKL. Journal of Machine Learning Research 9, 2491–2521.

Raviv, D., Bronstein, A. M., Bronstein, M. M., Kimmel, R., 2010. Volumetric heat kernel signatures. In: Workshop on 3D Object Retrieval. pp. 39–44.

Ray, K. M., Wang, H., Chu, Y., Chen, Y. F., Bert, A., Hasso, A. N., Su, M. Y., 2006. Mild cognitive impairment: apparent diffusion coefficient in regional gray matter and white matter structures. Radiology 24, 197–205.

Reuter, M., Wolter, F.-E., Shenton, M., Niethammer, M., 2009. Laplace-Beltrami eigenvalues and topological features on eigenfuntions for statistical shape analysis. Computed-Aided Design 41 (10), 739–755.

Rovaris, M., Bozzali, M., Iannucci, G., Ghezzi, A., Caputo, D., Montanari, E., Bertolotto, A., Bergamaschi, R., Capra, R., Mancardi, G. L., Martinelli, V., Comi, G., Filippi, M., 2002. Assessment of normal-appearing white and gray matter in patients with primary progressive multiple sclerosis a diffusion-tensor magnetic resonance imaging study. Archives of Neurology 59, 1406–1412.

Rubner, Y., Tomasi, C., Guibas, L. J., 2000. The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40 (2), 99–121.

Rujescu, D., Collier, D. A., 2009. Dissecting the many genetic faces of schizophrenia. Epidemiologia e Psichiatria Sociale 18 (2), 91–95.

Schüffler, P. J., Ulaş, A., Castellani, U., Murino, V., September 2011. A multiple kernel learning algorithm for cell nucleus classification of renal cell carcinoma. In: Proceedings of the International Conference on Image Analysis and Processing, ICIAP '11. p. accepted.

Schwarz, G., 1979. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Serratosa, F., Sanfeliu, A., 2006. Signatures versus histograms: Definitions, distances and algorithms. Pattern Recognition 39 (5), 921–934.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press.

Shenton, M. E., Dickey, C. C., Frumin, M., McCarley, R. W., Apr 2001. A review of mri findings in schizophrenia. Schizophrenia Research 49 (1–2), 1–52.

Smith, N., Gales, M., 2002a. Speech recognition using SVMs. In: Proceedings of the conference on advances in neural information processing systems, NIPS '02. Vol. 14. pp. 1197–1204.

Smith, N. D., Gales, M. J. F., 2002b. Using SVMs to classify variable length speech patterns. Tech. Rep. CUED/F-INFENG/TR–412, Cambridge University Engineering Department.

Sun, J., Ovsjanikov, M., Guibas, L., 2009. A concise and provably informative multi-scale signature based on heat diffusion. In: Proceedings of the Symposium on Geometry Processing, SGP '09. pp. 1383–1392.

Swain, M. J., Ballard, D. H., 1991. Color indexing. Int. Journal of Computer Vision 7 (1), 11–32.

Tansella, M., Burti, L., 2003. Integrating evaluative research and community based mental health care in verona-italy. British J. of Psychiatry 183, 167–169.

Tax, D. M. J., Loog, M., Duin, R. P. W., Cheplygina, V., Lee, W.-J., September 2011. Bag dissimilarities for multiple instance learning. In: Pelillo, M., Hancock, E. R. (Eds.), Proceedings of the International Workshop on Similarity-Based Pattern Analysis, SIMBAD '11. Vol. 7005 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg.

Taylor, W. D., Hsu, E., Krishnan, K. R. R., MacFall, J. R., 2004. Diffusion tensor imaging: background, potential, and utility in psychiatric research. Biological Psychiatry 55 (3), 201–207.

Timoner, S. J., Golland, P., Kikinis, R., Shenton, M. E., Grimson, W. E. L., Wells III, W. M., 2002. Performance issues in shape classification. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI '02. pp. 355–362.

Toews, M., III, W. W., Collins, D., Arbel, T., 2009. Feature-based morphometry. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI '09. pp. 109–116.

Tomasino, B., Bellani, M., Perlini, C., Rambaldelli, G., Cerini, R., Isola, M., Balestrieri, M., Caligrave, S., Versace, A., Mucelli, R. P., Gasparini, A., Tansella, M., Brambilla, P., 2010. Altered microstructure integrity of the amygdala in schizophrenia: a bimodal MRI and DWI study. Psychological medicine 41 (2), 301–311.

Tsuda, K., Kawanabe, M., Rätsch, G., Sonnenburg, S., Müller, K.-R., October 2002. A new discriminative kernel from probabilistic models. Neural Computation 14, 2397–2414.

Ulaş, A., Bicego, M., Castellani, U., Cristani, M., Murino, V., Perina, A., 2010a. WP7 mid-term report. Tech. Rep. D7.1, SIMBAD.
URL `http://simbad-fp7.eu/images/D7.1_Mid Term Report.pdf`

Ulaş, A., Castellani, U., Mirtuono, P., Bicego, M., Murino, V., Cerruti, S., Bellani, M., Atzori, M., Rambaldell, G., Tansella, M., Brambilla, P., November 2011a. Multimodal schizophrenia detection by multiclassification analysis. In: Martín, C. S., Kim, S.-W. (Eds.), Proceedings of the Iberoamerican Congress on Pattern Recognition, CIARP '11. Vol. 7042 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 491–498.

Ulaş, A., Castellani, U., Murino, V., Bellani, M., Tansella, M., Brambilla, P., November 2011b. Heat diffusion based dissimilarity analysis for schizophrenia classification. In: et al., M. L. (Ed.), IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB '11. Vol. 7036 of Lecture Notes in Bioinformatics. Springer Berlin / Heidelberg, pp. 306–317.

Ulaş, A., Duin, R. P., Castellani, U., Loog, M., Bicego, M., Murino, V., Bellani, M., Cerruti, S., Tansella, M., Brambilla, P., August 2010b. Dissimilarity-based detection of schizophrenia. In: ICPR workshop on "Brain Decoding: Pattern Recognition Challenges in FMRI Neuroimaging", WBD'10. pp. 32–35.

Ulaş, A., Duin, R. P. W., Castellani, U., Loog, M., Mirtuono, P., Bicego, M., Murino, V., Bellani, M., Cerruti, S., Tansella, M., Brambilla, P., 2011c. Dissimilarity-based detection of schizophrenia. International Journal of Imaging Systems and Technology 21 (2), 179–192.

Ulaş, A., Schüffler, P. J., Bicego, M., Castellani, U., Murino, V., September 2011d. Hybrid generative-discriminative nucleus classification of renal cell carcinoma. In: Pelillo, M., Hancock, E. R. (Eds.), Proceedings of the International Workshop on Similarity-Based Pattern Analysis, SIMBAD '11. Vol. 7005 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 77–88.

Vapnik, V. N., 1998. Statistical Learning Theory. John Wiley and Sons.

Ventura, J., Nuechterlein, K. H., Subotnik, K. L., Gutkind, D., Gilbert, E. A., 2000. Symptom dimensions in recent-onset schizophrenia and mania: a principal component analysis of the 24-item brief psychiatric rating scale. Schizophrenia Research 97, 129–135.

Voets, N. L., Hough, M. G., Douaud, G., Matthews, P. M., James, A., Winmill, L., Webster, P., Smith, S., 2008. Evidence for abnormalities of cortical development in adolescent-onset schizophrenia. NeuroImage 43 (4), 665–675.

World Health Organization, 1992. Schedules for Clinical Assessment in Neuropsychiatry. WHO, Geneva.

World Health Organization, 1996. SCAN 2.1.: Schede di valutazione clinica in neuropsichiatria. Il Pensiero Scientifico Editore, Roma.

Xu, Z., Jin, R., Yang, H., King, I., Lyu, M. R., 2010. Simple and efficient multiple kernel learning by group Lasso. In: Proceedings of the 27th International Conference on Machine Learning, ICML '10. pp. 1175–1182.

Yoon, U., Lee, J.-M., Im, K., Shin, Y.-W., Cho, B. H., Kim, I. Y., Kwon, J. S., Kim, S. I., 2007. Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. NeuroImage 34 (4), 1405–1415.

Zhang, Z., October 1994. Iterative point matching for registration of free-form curves and surfaces. International Journal of Computer Vision 13, 119–152.