Hierarchical Conditional Random Fields for Outlier Detection:
An Application to Detecting Epileptogenic Cortical Malformations

Bilal Ahmed[*]   BAHMED01@CS.TUFTS.EDU
Thomas Thesen[†]   THOMAS.THESEN@MED.NYU.EDU
Karen E. Blackmon[†]   KAREN.BLACKMON@NYUMC.EDU
Yijun Zhao[*]   YIJUN@CS.TUFTS.EDU
Orrin Devinsky[†]   OD4@MED.NYU.EDU
Ruben Kuzniecky[†]   RUBEN.KUZNIECKY@NYUMC.EDU
Carla E. Brodley[*]   BRODLEY@CS.TUFTS.EDU

[*]Department of Computer Science, Tufts University, Medford, Massachusetts, USA.
[†]Comprehensive Epilepsy Center, Department of Neurology, School of Medicine, New York University, New York, USA.

Abstract

We cast the problem of detecting and isolating regions of abnormal cortical tissue in the MRIs of epilepsy patients in an image segmentation framework. Employing a multiscale approach we divide the surface images into segments of different sizes and then classify each segment as being an outlier, by comparing it to the same region across controls. The final classification is obtained by fusing the outlier probabilities obtained at multiple scales using a tree-structured hierarchical conditional random field (HCRF). The proposed method correctly detects abnormal regions in 90% of patients whose abnormality was detected via routine visual inspection of their clinical MRI. More importantly, it detects abnormalities in 80% of patients whose abnormality escaped visual inspection by expert radiologists.

1. Introduction

The ultimate goal of our interdisciplinary research is to improve detection of structural brain lesions in people with medication-resistant epilepsy, so that the lesional tissue can be surgically removed. The first step is to locate the lesion generating the seizures using the patient's clinical MRI. If the lesion is detected via visual inspection (*MRI-positive*), then inter-cranial electrodes (IEEG) are implanted in the target region to monitor seizure activity to further localize the region responsible for causing the seizures. Yet, in

Focal Cortical Dysplasia (FCD), the most common form of medication-resistant epilepsy, 70-80% of histologically verified cases escape visual detection by expert radiologists (Thesen et al., 2011); in the absence of a visually detected lesion (*MRI-negative*), the success of surgical resection drops from 66% to 29% (Bell et al., 2009). In this case IEEG is used, but without a specific target it loses its efficacy in locating the lesion.

In this paper, we present an automated method for detecting FCD lesions using surface-based morphometry. The proposed method first applies a multiscale segmentation of the cortical surface and then combines the results via a hierarchical conditional random field (HCRF) (Reynolds & Murphy, 2007). Because we do not have accurately labeled training data, we cast the problem as an outlier detection task and thus extend the HCRF framework proposed in (Reynolds & Murphy, 2007) to perform outlier detection. The resulting outlier regions (lesions), sorted by their probability and surface area, are shown to a team of neuroradiologists and neurosurgeons who can combine the MRI lesion-detection results with other information such as the pattern of seizure onset and the patient's IEEG to determine the final candidate resection zone. In a dataset consisting of patients who have undergone resective surgery for whom the resected tissue was histopathologically verified to contain FCD lesions, the proposed method correctly identified lesions in 90% of the MRI-positive patients. But more importantly, for 80% of the MRI-negative patients who were seizure-free after resective surgery, it detected lesions within the resected region.

HCRFs have been effectively applied in computer vision for semantic image labeling (Plath et al., 2009), figure-ground segmentation (Reynolds & Murphy, 2007) and object detection (Awasthi et al., 2007). To train the HCRF, previ-

ous applications required either that each pixel in the image have a label or that a bounding box around the object(s) of interest is provided. Consequently, the accuracy of the labels directly impacts the final performance of the HCRF. In this paper, we propose an extension of the HCRF-based image segmentation and object detection framework for problems in which pixel-level labels are not available. For our task, we are given the MRIs of patients and of normal controls, but information about which pixels form a lesion in the patient MRIs is either missing or highly noisy. Thus, we have a global label for each image indicating either that the image contains no lesions (our healthy controls) or that it contains one or more lesions (our FCD patients). In Section 2 we explain in more detail why we cannot use the lesions/resection zones of previously treated patients as labels during training.

In our approach, we use surface-based morphometry (SBM) (Dale et al., 1999) to extract the cortical surface from the T1 weighted structural MRI scans. The key advantage of using SBM is that the cortical surfaces of different individuals can be registered to an average surface allowing for point-wise comparisons between different individuals. Section 2 briefly introduces SBM, the different morphological features that can be extracted within the SBM framework, and describes the current state of the art in lesion detection using SBM and its shortcomings. In Section 3 we detail how we pose the lesion detection task within the object detection/segmentation framework, where the saliency of the target object is defined based on its "outlier-ness". We evaluate our results within an information-retrieval framework where the detected clusters are ranked based on their probability and surface area. Section 4, provides the detail of our ranking and evaluation methodology and it also provides an empirical comparison of our approach and a recent approach reported in the epilepsy literature. We also compare the performance of the proposed approach with a senior neuroradiologist at one of the world's leading tertiary epilepsy treatment centers. In this experiment the findings of the expert and the HCRF based approach were found to be highly correlated in the case of MRI-positive patients. For MRI-negative patients, the expert was unable to detect any lesion with or without the use of the HCRF approach.

Our work makes a significant contribution toward the detection of FCD lesions. Our empirical evaluation demonstrates that not only are we able to detect lesions in a higher number of MRI-negative patients as compared to the state of the art, but we are able to detect lesions that escape detection by one of the world's most expert neuroradiologists. Thus this work has the potential to increase the number of patients who are referred to resective surgery, which

remains underutilized[‡] (Tllez-Zenteno et al., 2005) despite a growing number of studies that demonstrate its efficacy.

Our contribution to machine learning in addition to making progress on a challenging and important application is a new method for using HCRFs for binary object detection/segmentation for which only image captions are available. A caveat to this contribution is that the images must be able to be accurately registered such that a one-to-one correspondence can be made between sub-regions.

2. SBM and Lesion Detection

Surface based morphometry (SBM) provides the means to characterize and analyze the human brain by explicitly modeling the cortex using a suitable geometric model (Dale et al., 1999). The cortical surface represents the outer layer of the brain modeled as a folded two-dimensional surface. It is extracted by delineating the boundary between the gray and white matter using T1-weighted MRI images (Dale et al., 1999). After extraction, different morphological transforms can be applied to register the cortical surface to a standard surface also known as a group-atlas. Registration is achieved by aligning specific sulcal and gyral features across the *reconstructed* cortical surfaces allowing for a more precise comparison of individual cortical structures across subjects (Fischl & Dale, 2000). SBM has been used successfully for analyzing and detecting neurological abnormalities in various neurological disorders such as schizophrenia (Rimol et al., 2012), autism (Nordahl et al., 2007), and epilepsy (Thesen et al., 2011).

Different morphological features such as cortical thickness, curvature, gray-white contrast (GWC) etc., can be extracted to characterize the cortex. Recently, SBM has been used in conjunction with machine learning and statistical techniques to identify lesions in FCD patients. Besson et al., (Besson et al., 2008) use texture, GWC and a number of morphological features including cortical thickness to represent each vertex on the reconstructed cortical surface. They then train a neural network to classify each vertex as being normal or lesional. Similarly, Thesen et al., (Thesen et al., 2011) use a uni-variate z-score based thresholding approach on registered SBM data to classify each vertex as being lesional or normal, using cortical thickness, GWC, curvature, sulcal depth and Jacobian-distortion, individually. They nominate cortical thickness along with GWC as being the most informative features for FCD lesion detection. In our empirical evaluation we use this latter z-score approach (Thesen et al., 2011) as the baseline from which to compare results from our HCRF method; therefore, we

---

[‡]Patients who lack an MRI-visible lesion are less likely to be referred to specialized epilepsy centers by neurologists (Benbadis et al., 2003) and a high number of epilepsy specialists are reluctant to operate without a well-defined lesion.

similarly employ cortical thickness to facilitate comparison. A comprehensive review of automated FCD lesion detection can be found in Bernasconi et al.,(Bernasconi et al., 2011). These studies classify individual vertices of the cortical surface as lesional or normal, using labeled training data from MRI-positive patients and controls. There are four crucial issues that these studies fail to address:

(1) The goal of resective surgery is to remove the entire lesion. If any part of the lesion is left behind, the outcome will not be successful. This introduces label noise, because the expert-marked lesion can contain normal vertices; the margin around the lesion is marked in a "generous" manner so as to increase the chances of capturing the entire lesion.

(2) These studies assume that individual vertices are i.i.d., completely ignoring the spatial correlation that exists between neighboring vertices. It has been shown in other domains such as object detection and segmentation in natural images, that modeling spatial correlations leads to superior performance (Reynolds & Murphy, 2007; Plath et al., 2009).

(3) Usually a post-processing method is employed to reduce the false positive rate. In this strategy a portion of the vertices labeled lesional by the classifier are relabeled as normal. This can be done by training a second-level classifier to classify the detected clusters as lesional or nonlesional (Besson et al., 2008). Also different heuristics can also be used such as the surface area of the detected clusters (Thesen et al., 2011). Discarding any detected region based on its size or surface area can result in discarding the actual lesion or part of the lesion, because FCD can be located in any part of the cortex, is highly variable in size, and may occur in multiple lobes (Blumcke et al., 2011).

(4) Results are evaluated on MRI-positive patients, but the real challenge is to find lesions in MRI-negative patients.

Our proposed method is designed to explicitly address these issues. First, we model lesion detection as an outlier detection problem. The assumption is that a lesional region is an outlier in a suitable feature space when compared to the same region across a control population. This view eliminates the use of noisy class labels. Second, instead of classifying individual vertices we classify segmented patches of the cortex. The patches are obtained using unsupervised segmentation of the flattened cortex that isolates regions of homogeneous feature values. As the size of the FCD lesions has a wide range, using a single scale to isolate the lesion may not be effective. To minimize the chances of missing the lesion, we employ a multiscale strategy where the segmentation is carried out at different scales of varying granularity. The interplay between the patches obtained in this scale hierarchy is modeled as a tree structured HCRF, rooted at the most crude scale and having leaves at the finest scale. Third, we rank the detections based on a criterion

that takes into account both the probability that a detected cluster is abnormal along with its surface area. Therefore, larger clusters with a higher probability of being abnormal are ranked most highly. This ranking approach eliminates the need to post process the results, and provides a natural way of presenting the results to a radiologist to function as a focus of attention mechanism. Finally, we evaluate our approach on both MRI-positive and MRI-negative patients. The next section describes the details of the HCRF construction and inference.

3. Outlier Detection using HCRFs

In the original HCRF framework proposed for figureground segmentation (Reynolds & Murphy, 2007), an image is first segmented into a number of patches at different scales. Each patch is then classified as being part of the background or foreground, using a suitable binary classifier based on image features such as texture, SIFT, etc. Exploiting the fact that the labels assigned to overlapping patches between different scales should agree, an HCRF (a tree-structured conditional random field) is constructed to model these inter-scale interactions. The image is thus modeled as a forest, where the root node for each tree corresponds to a patch obtained at the coarsest scale, while the leaves reside at the finest scale. The joint probability of all the patch labels is estimated by running inference on the HCRFs. *The image is segmented by thresholding the final probabilities at the leaves.* Plath et al.,(Plath et al., 2009) extend this framework to work with more than two classes. Mutliclass image labeling using HCRFs is also done in (Awasthi et al., 2007), where instead of obtaining image patches using segmentation, the authors impose a grid structure on the image at different scales and model the HCRF as a quad-tree structure. These multiscale methods are highly sensitive to the accuracy of pixel-level labels. For example in (Reynolds & Murphy, 2007) the bounding boxes around the ROI in training images were manually refined to eliminate extraneous pixels and this resulted in a significant increase in accuracy.

In our problem domain we have training data from two distinct types of epilepsy patients: MRI-positive and MRI-negative. For the MRI-positive patients we could use the radiologist-traced lesion as local vertex-level labels, and for MRI-negative patients the resected region could serve the same purpose. However, as explained previously these labels tend to be highly noisy and using them to train the classifiers will result in noisy predictions. To ameliorate this problem we extend the HCRF framework proposed in (Reynolds & Murphy, 2007) to perform semi-supervised outlier detection on registered image data. Thus, we define an FCD lesion as a region of the brain which will be considered an outlier when compared to the same region across

a population of normal controls. Before describing the extension we first describe our approach to segmentation.

3.1. Segmentation

Recall that in SBM the cortex is modeled as a two-dimensional surface, which on average contains approximately 0.15 million vertices. Even though it is possible to flatten the entire cortex, segmentation and the resulting HCRFs for the resulting image would require significant computational resources. Thus, to reduce the processing overload we have chosen to subdivide the detection task into smaller regions of the cortical surface defined by a standard atlas, which outlines cortical regions based on their morpho-functional properties (Fischl et al., 2002). These regions are also known as *parcellations*. Instead of segmenting the entire cortical surface at once, we isolate these parcellations one at a time and flatten them individually to obtain a standard two dimensional image, which we then segment at multiple scales. Any morphological feature (e.g., cortical thickness), can be used to represent the intensity values of the resulting image. Figure 1 illustrates the overall HCRF construction process for a given parcellation image. We have chosen to model the cortex using cortical thickness (Fischl & Dale, 2000), which has been established as one of the most informative features for the detection of FCD lesions (Thesen et al., 2011).

We use quick shift (Vedaldi & Soatto, 2008b) for unsupervised segmentation. One of the main advantages for using quick shift is that the number and size of segments need not be specified. Additionally, quick shift does not penalize for boundary regions, and produces a diverse set of segments having different shapes and sizes. It should be noted that any segmentation method can be used, as long as it has the ability to segment the image at different scales.

The standard quick shift algorithm is a fast mode seeking algorithm similar to mean shift (Comaniciu & Meer, 2002). It performs a hierarchical segmentation of the image, where the sub-trees represent image segments. It has two parameters namely the size of the Gaussian kernel (σ) used by a Parzen window density estimator, and the maximum distance (Δ) between two pixels permitted while remaining part of the same segment. We vary the scale parameter σ to change the average size of segments, and set Δ to be a multiple of σ (Vedaldi & Fulkerson, 2008a). Thus, higher values of σ produce larger segments.

3.2. Multiscale HCRFs

Once the multiscale segmentation is complete for a particular subject, we obtain a set of patches at different scales. Let I_p^k be the pth patch obtained at the kth scale. We can collect the corresponding patches from all controls and then estimate a label $y \in \{0, 1\}$ for I_p^k, where $y = 1$ indicates
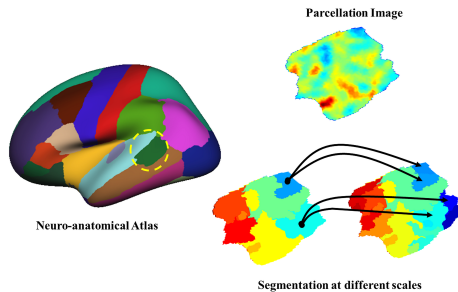

Figure 1. Construction of the HCRF. A standard neuroanatomical atlas (*left*), and the isolated parcellation image (*top-right*). Cortical thickness values are used to represent the image. On the bottom we have the super-pixels obtained at two different scales using Quickshift. Each super-pixel on the coarser scale (*bottom-left*) becomes a root having children at the finer scale (*bottom-right*).

that I_p^k is an outlier. This label cannot be considered independent from the labels of other patches that overlap with I_p^k at other scales.

We model the joint prediction of these mutually dependent labels of all the patches using a tree structured HCRF. Let I_p^{k+1} be an image patch at level $k + 1$, it has a parent I_q^k at the immediately coarser level k, such that I_q^k has maximal overlap with I_p^{k+1} (Reynolds & Murphy, 2007). We find the index q as follows:

$$q := \arg\max_q \frac{|I_p^{k+1} \cap I_q^k|}{|I_q^k|} \qquad (1)$$

Each patch at the coarsest scale is the root of a tree having leaves at the finest scale. Therefore, the image is represented by a forest, where each tree is modeled as an HCRF, as shown in Figure 1.

CRFs model the joint conditional probability distribution of all the patch labels $y = (y_1, \ldots, y_n)$ in the tree based on the input cortical thickness values (x). Generally, this can be written as:

$$p(y|x,\theta) = \frac{1}{Z(x,\theta)} \prod_i \phi(y_i|x,\theta) \prod_{i,\pi(i)} \psi(y_i, y_{\pi(i)}) \quad (2)$$

where, $\pi(.)$ represents the parent patch, and $Z(x, \theta)$ is the normalization constant also called the partition function. $\phi(.)$ is called the node potential and represents the local evidence for the label y_i based on the observed data x. The edge potentials that model the coupling between adjacent labels are represented by $\psi(.)$. As the graph is a tree we can efficiently calculate $Z(x, \theta)$ and the posterior probabilities of the patch labels at all scales using standard belief propagation (Pearl, 1988).

When labeled training data is available the node and edge potentials are parameterized, and the parameters are

learned jointly (see (Sutton & McCallum, 2010) for details). For our application, because the labels are noisy and we have chosen to work in an unsupervised manner, we set the node and edge potentials as described next.

3.2.1. NODE POTENTIALS

The node potential is modeled to reflect our belief about the abnormality of an individual image patch. We have chosen to work with local outlier probabilities (LoOP) (Kriegel et al., 2009), a standardized version of local outlier factor (LOF) (Breunig et al., 2000). LoOP produces standardized scores within the range $[0, 1]$ which can be treated as the probability that a data point is an outlier. This is an important design choice because running inference on non-standardized scores, which may not be comparable between different scales, can produce meaningless results.

3.2.2. EDGE POTENTIALS

Each edge in the HCRF represents the dependency between the "parent" image patch at scale t and the "child" patch at scale $t + 1$. We set the edge potential to reflect the visual similarity between the two patches, using the chi-squared distance between the histograms of scale invariant feature transform (SIFT) features (Lowe, 1999) of the parent and child patches. Thus the labels of image patches that bear close visual similarity to each other in the scale hierarchy are more strongly coupled than those with lower similarity. This heuristic is similar to one chosen by Reynolds and Murphy (Reynolds & Murphy, 2007).

To estimate the histograms of the SIFT features for each image, we initially learn a codebook of m codewords using the control data. For each control image in the subset we flatten and isolate the parcellation, and then calculate a SIFT feature vector at each pixel. These vectors are then clustered into m clusters using *k-means* clustering. A separate codebook is learned for each parcellation. The edge potential between two adjacent nodes in the tree is then calculated as (Reynolds & Murphy, 2007; Plath et al., 2009):

$$\psi(y_i, y_j) = \begin{bmatrix} e^{\gamma \cdot \eta_{ij}} & e^{-\gamma \cdot \eta_{ij}} \\ e^{-\gamma \cdot \eta_{ij}} & e^{\gamma \cdot \eta_{ij}} \end{bmatrix} \quad (3)$$

where, γ is a free parameter that represents the strength of coupling between adjacent levels in the CRF and $\eta_{ij} = e^{-\chi^2(x_i, x_j)}$. x_l represents the normalized histogram of SIFT features for the lth patch in the HCRF, and $\chi^2(., .)$ is the chi-squared distance between two normalized histograms.

3.2.3. LESION DETECTION

For each subject, we calculate the posterior probabilities at each node of the HCRF for every parcellation by running belief propagation (Pearl, 1988). The final detection is
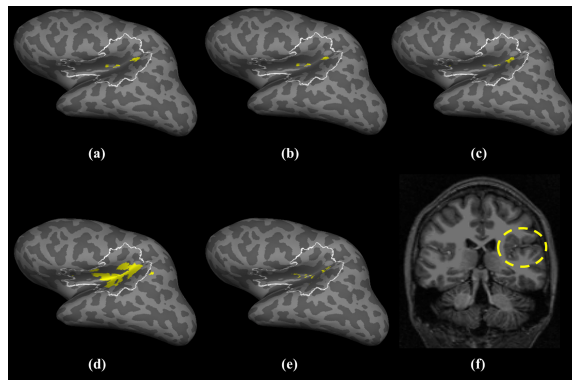

Figure 2. Detection results for an MRI positive subject shown on an inflated model of the lateral cortical surface. The actual lesion is delineated as the white circled region and the detection results are shown as filled yellow regions. Detected clusters after thresholding outlier probabilities at each individual scale (a)-(c), after running belief propagation (d), and using the z-score based approach (e). The results are shown for the highest ranking threshold without any post-processing. (f) shows the lesion highlighted on a T1 MRI slice.

obtained by thresholding the posterior beliefs at the leaves of each HCRF (Reynolds & Murphy, 2007; Plath et al., 2009). Different strategies for thresholding can be used, such as defining a single threshold across all subjects, or calculating a threshold for each subject individually. In this work we calculate an adaptive threshold for each patient separately. This decision is based on the observations that 1) FCD lesions can be manifested differently for different individuals, and 2) cortical thickness varies with different demographic factors such as gender and age (Salat et al., 2004). To this end, we sort the posterior probabilities and define the threshold as the lowest probability among the top K probability estimates. In practice the value of K can be left as a free parameter that the user can vary to see the different regions which are deemed lesional with varying levels of confidence. Thus, the radiologist has a knob to turn which shows more/fewer possible lesions.

4. Results

Our data consists of MRI-positive and MRI-negative patients who have undergone resective surgery and for whom their resected tissue was histopathologically verified to contain abnormal tissue. Each patient who undergoes surgery is assigned an "Engel" class. An Engel class of 1 represents complete seizure freedom while an Engel class of 4 represents no improvement. We selected only patients with an Engel class outcome of 1 for our experiments in order to verify that the region resected was indeed the primary lesion and that no additional epileptogenic lesions were present in other parts of the brain. This resulted in a dataset with fifteen MRI-negative patients. This may appear to be a

small dataset, but few patients proceed to surgery when no visible lesion is found on their MRI, and of those that do, less than a third experience complete seizure freedom (Bell et al., 2009). We also include experimental results on a small set of 10 MRI-positive patients, but want to reemphasize that our method is not needed for MRI-positive patients as their lesion is visible.

Imaging for both the subjects and the controls was performed on a Siemens Allegra 3T scanner. Image acquisitions included a conventional 3-plane localizer and a T1-weighted volume pulse sequence (TE=3.25 ms, TR =2530 ms, TI =1100 ms, flip angle =7 deg field of view (FOV) = 256 mm, matrix = 2566256, vertex size =$1 \times 1 \times 1.3$ mm, scan time: 8:07 min). Acquisition parameters were optimized for increased gray/white matter image contrast. The T1-weighted image was reoriented into a common space, roughly similar to alignment based on the AC-PC line. Images were corrected for nonlinear warping caused by no-uniform fields created by the gradient coils. In this study we have a total of 115 controls, 55 males (33.7 ± 12.5 years) and 60 females (32.0 ± 11.5 years). It should be noted that all the subjects were scanned on the same scanner, and the data used here is based on these research scans, which is different from their original clinical scans.

4.1. Data Pre-processing and Parameter Selection

After the surface has been reconstructed using the freesurfer software (http://surfer.nmr.mgh.harvard.edu/) we used the Desikan-Killiany atlas (Desikan et al., 2006) to isolate the different parcellations. It should be noted that any suitable neuroanatomical atlas can be used to subdivide the cortical surface. Each parcellation is flattened to obtain a standard 2-d image, where the intensity of each pixel is represented by cortical thickness.

To select the parameters for the various aspects of our method, we used a validation set consisting of two MRI-positive and two MRI-negative patients, which are distinct from the patients used to evaluate our method. We used all 115 controls to learn a codebook of SIFT features for every parcellation. Dense SIFT features were calculated at each pixel. We tested vocabulary sizes of 50, 100 and 500 and selected a vocabulary size of 50 as it resulted in higher recall and precision on the validation set. This codebook was used subsequently to estimate the histograms of SIFT features at each pixel location for all patient parcellation images in the test set.

Each parcellation image was segmented at three different scales using quick shift. We used $\sigma = \{2, 3, 4\}$ and Δ was set to 5σ. These values were chosen so that the smallest possible lesion in our validation set is over-segmented i.e., there are multiple segments that contain the lesional area. This increases the probability that a patch can be entirely
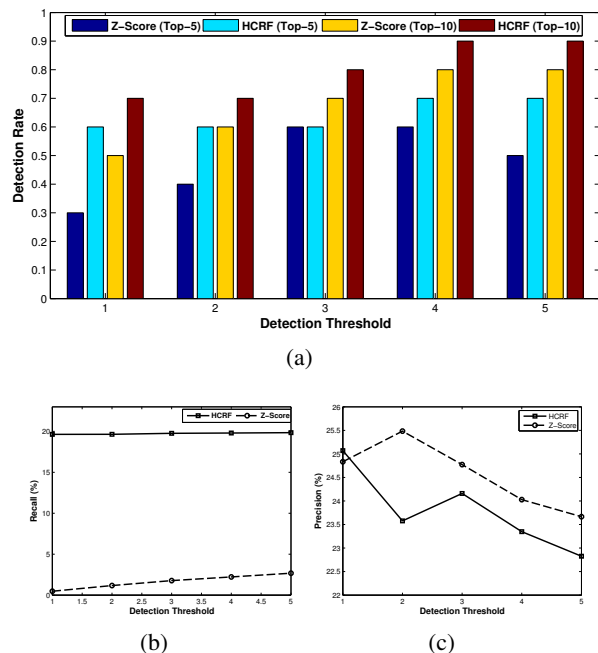

(a)

(b)   (c)

Figure 3. Comparison of detection rate (a) between then HCRF based approach and the z-score based baseline method for MRI-positive patients. The detection rate is the percentage of the patients for whom the detected clusters overlapped with the lesional area. (b) and (c) compare the recall and precision, respectively.

formed from lesion vertices, rather than having patches that partially overlap with the lesion, which would be harder to detect as outliers.

Finally, before performing outlier detection, we apply a standard dimension reduction technique on each patch using principal component analysis (PCA). Note that the PCA is done using only the control data. We retained the top m principal components that accounted for 95% of the variance in data. To set the parameters for outlier detection, based on results for the validation set, we set $k = 10$ in LoOP and γ (c.f. equation (3)) was set to 50.

Whether an image patch is an outlier depends on the set of controls used to learn the "normal" model. As cortical thickness varies both by age and gender, ideally, we could choose a customized set of controls for each patient, but currently we do not have enough controls to customize for age, but we do select controls based on the patient's gender.

4.2. Evaluation Methodology

The final detection for each subject is determined by thresholding the posterior probabilities at the leaves of the CRF, which represent the segments obtained at the finest scale. We determine the detection thresholds by dividing the last percentile of the final outlier probabilities into ten equal parts. The first threshold corresponds to the lowest proba-
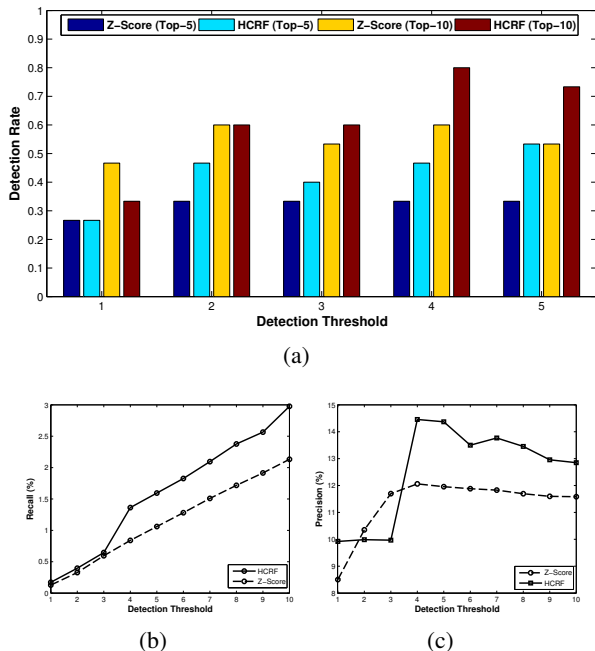
(a)

(b)   (c)

Figure 4. Comparison of detection rate (a) between then HCRF based approach and the z-score based baseline method for MRI-negative patients. The detection rate is the percentage of the patients for whom the detected clusters overlapped with their resection. (b) and (c) compare the recall and precision respectively.

bility in the highest 0.1% scores and so on. For the results presented in this section we determine five such thresholds to get five different possible detections. Because, this is an adaptive mechanism, it has a possible drawback that it always detects something even when the probabilities are very small. Thus we set 1×10^{-4} as the minimum probability, such that no threshold is calculated below this value.

We have chosen to evaluate and contrast the performance of the detection techniques in an information retrieval framework. We first calculate the clusters by thresholding the posterior probability at a given threshold. All the detected clusters are then ranked according to their surface area.

We compare the results of our proposed technique against a recently reported technique that also employs cortical thickness (Thesen et al., 2011) to detect FCD lesions using SBM. In this baseline approach first all the control and subject surfaces are registered to the average surface. After registration, it calculates the z-scores at each vertex for the subjects, which are then thresholded to obtain the detection results. We calculate the z-score based on gender matched controls instead of using all the controls. To facilitate comparison we calculate multiple thresholds in the exact same manner as outlined above for HCRF. We omit the last step of Thesen et al's method, which post processes the detections to eliminate "small" clusters.

4.2.1. DETERMINING DETECTION RATES

We compare the accuracy of our detection rates to the z-score method by determining the number of patients with top ranked clusters within the lesion/resection zone for each method. To this end, we calculate five thresholds based on the outlier probabilities for the HRCF method, and similarly for the z-score method. After ranking the detected clusters based on their surface area, at each threshold we consider a subject to be correctly detected if a cluster amongst the top n completely or partially overlaps with the lesion/resection. Figure 2 shows the clusters detected by the baseline and the HCRF, for an MRI-positive patient.

Figure 3(a) shows the comparison of the detection rates for MRI-positive patients. HCRF performs better than the z-score baseline across all the five thresholds, for both top five and top ten detections. Similarly, Figure 4(a) shows the same comparison for MRI-negative patients. Here, the HCRF does not dominate across all thresholds; its detection rate increases as the threshold becomes more lenient (from the third threshold onwards). HCRF detects the lesion in 12 (80%) patients, while the baseline detects only 9 (60%) subjects when considering top ten clusters.

4.2.2. DETERMINING PRECISION AND RECALL

In order to compare the quality of detections, we calculated the precision and recall for both HCRF and the z-score based method. To this end, we consider all detected clusters. We define recall as the ratio of the total surface area of all the clusters that overlap with the lesion/resection zone to the surface area of the lesion/resection zone. Similarly, we define precision as the ratio of the surface areas of clusters overlapping with the lesion/resection zone to the sum of the surface area of all the detected clusters.

HCRF produces higher recall for both the MRI-positive (Figure 3(b)) and MRI-negative (Figure 4(b)) subjects. In the case of MRI-positive subjects this difference is more pronounced and we observe that for both methods the maximum recall for MRI-positive subjects is significantly higher than for MRI-negative subjects.

Figures 3(c) and 4(c) show the precision for MRI-positive and MRI-negative subjects, respectively. Z-score performs better for the MRI-positive patients and the picture is mixed for the MRI-Negative patients. Accurately calculating precision is challenging for several reasons. A patient can have abnormalities outside the lesion/resection zone which may not be epileptogenic. Abnormal cortical thinning remote from the epileptogenic onset region has been observed in focal epilepsy (McDonald et al., 2008; Lin et al., 2007) and attributed to the destructive impact of chronic seizures on brain structure rather than developmental lesions. This might result in elevated extra-lesional false positives when

cortical thickness is used as the primary feature. Figure 5 shows such an example. Figure 5a-b shows the results of the HCRF detection on an MRI-positive patient, as can be seen there are a number of clusters, that lie outside the re-section zone and should be considered as false positives. Figure 5d shows a T1 MRI slice, where the circled area delineates a severely abnormal area (confirmed by an expert neuroradiologist) outside the resection zone, that corresponds to the largest cluster detected by HCRF. These observations led us to ask a human expert to examine these areas more closely which we describe next.

4.2.3. COMPARISON TO HUMAN EXPERT

We compared the results of HCRF detection, to the senior neuroradiologist, at one of the worlds leading tertiary epilepsy treatment centers. The radiologist was first presented with the anonymized clinical scans of each subject, this included T1 and T2 weighted MRI along with FLAIR. He was then asked to identify any possible abnormality, being totally blind to any post-surgical data and the HCRF detection results. We compared his findings to the clusters found by HCRF at the most lenient (fifth) threshold.

For this experiment, we require patients that had a full pre-surgical MRI dataset (e.g., T2-weighted images, FLAIR). Note that pre-surgical clinical MRI sequences were not obtainable for many subjects due to their having been referred from external centers. This set of patients included six of the MRI-positive patients and only 3 of the MRI-negative patients. We included three MRI-negative patients who did not have an Engel outcome of 1 (i.e., they were not seizure free) because we had their clinical scans.

For the MRI-positive patients, the radiologist detected abnormalities in all six cases, that correlated with the HCRF detections for five subjects considering the top five clusters, and with all six in the top ten clusters. For three of the MRI-positive subjects his findings also identified abnormal regions that overlapped with extra-lesional clusters detected by HCRF that ranked amongst the top ten. An example is depicted in Figure 5. In the case of MRI-negative subjects, the radiologist was unable to identify any visible abnormality in all six subjects. HCRF on other hand identified three subjects out of a total of five that had outcomes of Engel class 1-3. For one MRI-negative subject with an Engel-4 outcome, HCRF found no cluster among the top ten that overlapped with his resection zone.

The preliminary results on the MRI-negative patients, albeit on a small sample, are promising because the HCRF method is able to identify high ranked clusters within the resection zones of MRI-negative patients who have complete to partial seizure freedom after surgery. Overall, the results indicate that the HCRF approach has a higher sensitivity to histopathologically-confirmed lesions that are not
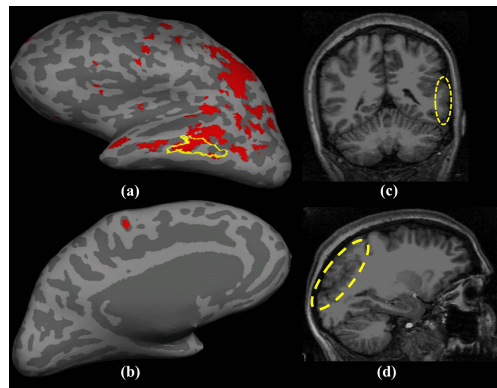

Figure 5. An MRI-positive for whom the clusters detected outside the resection are also abnormal. Detected clusters using HCRF (a)-(b). The resected area (c), and the area corresponding to the largest cluster outside the resection (d) shown on a T1 MRI slice.

visible to an expert radiologist, even when a full set of clinical MRI sequences are available for review.

5. Conclusion

In this work we have extended the multiscale HCRF framework traditionally used for supervised image segmentation and object localization to perform semi-supervised outlier detection. We have applied this proposed method to the challenging task of detecting cortical lesions in patients suffering from treatment resistant epilepsy, using their surface-based MRI morphometric data. Our results show that for MRI-positive subjects, we are able to correctly localize lesions in 90% of the subjects. As compared to the z-score approach we not only correctly localize the lesion in a higher number of patients, but we do so with a substantially higher recall. Similarly, in the case of MRI-negative patients we are able to achieve a detection rate of 80%, as compared to the baseline z-score detection rate of 60% and the human expert rate of 0%. For the MRI-negative patients, the HCRF approach achieves both higher recall and precision (at higher thresholds). Our next steps are to incorporate HCRF results into the weekly meeting of radiologist and neurosurgeons to help in IEEG placement for MRI-negative patients who may be candidates for surgery. To improve on our current approach we have two initial avenues. We will assess whether performance improves if we include other features such as gray-white contrast, and cortical curvature. Second, we want to further extend the proposed HCRF framework to incorporate the resection zones as "soft labels" to learn its parameters jointly.

Acknowledgments

References

Awasthi, P., Gagrani, A., and Ravindran, B. Image modelling using tree structured conditional random fields. In *IJCAI*, pp. 2060–2065, 2007.

Bell, M.L., Rao, S., and So, E.L. , et al. Epilepsy surgery outcomes in temporal lobe epilepsy with a normal MRI. *Epilepsia*, 50(9):2053–2060, 2009.

Benbadis, S.R., Heriaud, L., Tatum IV, W.O., and Vale, F.L. Epilepsy surgery, delays and referral patternsare all your epilepsy patients controlled? *Seizure*, 12(3):167 – 170, 2003.

Bernasconi, A., Bernasconi, N., Bernhardt, B.C., and Schrader, D. Advances in mri for 'cryptogenic' epilepsies. *Nat Rev Neurol.*, 7(2):99–108, 2011.

Besson, P., Bernasconi, N., and Colliot, O., et al. Surface-based texture and morphological analysis detects subtle cortical dysplasia. In *MICCAI*, pp. 645–652, 2008.

Blumcke, I., Thom, M., and Aronica, E.,et al. The clinico-pathologic spectrum of focal cortical dysplasias: A consensus classification proposed by an ad hoc task force of the ILAE diagnostic methods commission. *Epilepsia*, 52 (1):158–174, 2011.

Breunig, M., Kriegel, Hans-Peter, Ng, R.T., and Sander, J. LOF: Identifying Density-Based Local Outliers. In *ACM SIGMOD ICMD*, pp. 93–104. ACM, 2000.

Comaniciu, D. and Meer, P. Mean shift: a robust approach toward feature space analysis. *PAMI*, 24(5):603–619, 2002.

Dale, A.M., Fischl, B., and Sereno, M.I. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

Desikan, R.S., Sgonne, F., and Fischl, B., et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006.

Fischl, B. and Dale, A.M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *PNAS*, 97(20):11050–11055, 2000.

Fischl, B., Salat, D.H., and Busa, E. , et al. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

Kriegel, Hans-Peter, Kröger, P., Schubert, E., and Zimek, A. LoOP: Local Outlier Probabilities. In *ACM CIKM*, pp. 1649–1652, 2009.

Lin, J.J., Salamon, N., and Lee, A.D.,et al. Reduced neocortical thickness and complexity mapped in mesial temporal lobe epilepsy with hippocampal sclerosis. *Cereb. Cortex*, 17(9):2007–2018, 2007.

Lowe, D.G. Object recognition from local scale-invariant features. In *ICCV*, pp. 1150–1157, 1999.

McDonald, C.R., Jr, D.J.H., and Ahmadi, M.E.,et al. Regional neocortical thinning in mesial temporal lobe epilepsy. *Epilepsia*, 49(5):794–803, 2008.

Nordahl, C.W., Dierker, D., and Mostafavi, I., et al. Cortical folding abnormalities in autism revealed by surface-based morphometry. *J Neurosci.*, 27(43):11725–11735, 2007.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988. ISBN 1558604790.

Plath, N., Toussaint, M., and Nakajima, S. Multi-class image segmentation using conditional random fields and global classification. In *ICML*, pp. 817–824, 2009.

Reynolds, J. and Murphy, K. Figure-ground segmentation using a hierarchical conditional random field. In *CRV*, pp. 175–182, 2007.

Rimol, L.M., Nesvg, R., and Hagler Jr., D.J., et al. Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. *Biological Psychiatry*, 71(6):552–560, 2012.

Salat, D.H., Buckner, R.L., and Snyder, A.Z.,et al. Thinning of the cerebral cortex in aging. *Cerebral Cortex*, 14 (7):721–730, 2004.

Sutton, C. and McCallum, A. An Introduction to Conditional Random Fields, 2010. eprint arXiv:1011.4088.

Thesen, T., Quinn, B.T., and Carlson, C., et al. Detection of epileptogenic cortical malformations with surface-based MRI morphometry. *PLoS ONE*, 6(2):e16430, 2011.

Tllez-Zenteno, J.F., Dhar, R., and Wiebe, S. Long-term seizure outcomes following epilepsy surgery: a systematic review and meta-analysis. *Brain*, 128(5):1188–1198, 2005.

Vedaldi, A. and Fulkerson, B. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008a.

Vedaldi, A. and Soatto, S. Quick shift and kernel methods for mode seeking. In *ECCV*, pp. 705–718, 2008b.