CONCEPT DETECTION IN LONGITUDINAL BRAIN MR IMAGES USING MULTI-MODAL CUES

Jesus J. Caban[1], *Noah Lee*[2], *Shahram Ebadollahi*[3], *Andrew F. Laine*[2] *and John R. Kender*[3]

[1]University of Maryland, UMBC
[2]Biomedical Engineering, Columbia University
[3]IBM T.J. Watson Research Center

ABSTRACT

Advances in medical imaging techniques and devices has resulted in increased use of imaging in monitoring disease progression in patients. However, extracting decision-enabling information from the resulting longitudinal multi-modal image sets poses a challenge. Radiologists often have to manually identify and quantify certain regions of interest in the longitudinal image sets, which bear upon the patient's condition. As the number of patients increases, the number of longitudinal multi-modal images grows, and the manual annotation and quantification of pathological concepts quickly becomes impractical. In this paper we explore how minimal annotations provided by the user at a few time points can be effectively leveraged to automatically annotate data in the entire multi-modal longitudinal image sets. In particular, we investigate the required number of annotated images per time point and across time for obtaining reasonable results for the entire image set, and what multi-modal cues can help boost the overall annotation results.

Index Terms— Computer-Aided Diagnosis Systems, Multi-modal Images, Longitudinal Image Sets, Supervised Learning, Semi-supervised Learning, Brain Tumors

1 Introduction

MRI studies are widely accepted in brain tumor patient management as reliable indicators for obtaining prognostic information and observing the patient's response to treatment plans. Clinicians can obtain valuable insight about the disease progression and the effect of a particular treatment plan by examining the temporal images for a given patient and correlating them with other factors and biomarkers. Figure 2 shows the co-evolution pattern of the "edema" volume, as obtained from the longitudinal image studies, and the timing of a specific therapy plan.

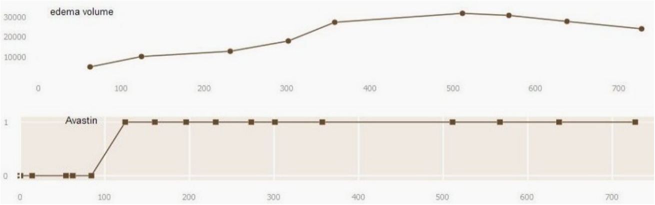

Fig. 1*:* Co-evolution of the volume of edema region and the dosage of the drug Bevacizumab (Avastin) is displayed over time for a given patient.

The challenge is how to identify and quantify the concepts of interest in the large number of imaging studies attributed to a patient. Patients with high-grade Gliomas, such as Glioblastoma Multiforme (GBM), have often about ten or more temporal studies and for each study multiple MRI protocols including T1, T2, T1 and T2 with contrast, and Fluid Attenuated Inversion Recovery (FLAIR). Figure 2 shows a subset of the longitudinal images found within a specific patient record. Clinicians simply do not have the time to manually annotate each image in the longitudinal patient records or detect patterns along multi-modal images. The process needs to be automated but at the same time guided by the experts knowledge. Accurate detection of pathological concepts and the analysis of such findings is useful in making better judgments about the effectiveness of particular cancer therapies.

Given the large amount of data and the limited number of tools to analyze multiple images simultaneously, radiologists frequently use a single MRI protocol at each time point to locate pathological concepts. For instance, concept "edema" is usually identified using FLAIR images given that such images enhance the edema regions by assigning distinct intensity values. However, noise and image artifacts can introduce a significant amount of false positives when a single image modality is used. That is why in machine annotation of pathological concepts, cues obtained from multiple protocols can often enhance the results [1].

Previously, a number of methods have been proposed to detect pathological concepts such as "edema" and "tumors" [1]. Most existing techniques rely on training a system that can learn the characteristic properties of the condition under consideration. However, information about how to minimize user input while improving accuracy has not been reported. This paper reports our approach, experiments, and results for designing a framework to automatically annotate pathological concepts within longitudinal MR images. We address some of

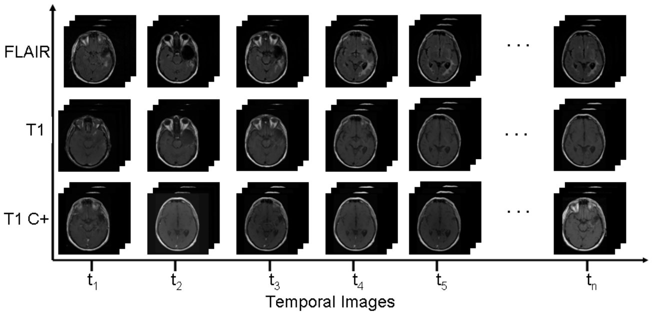

Fig. 2: Subset of the longitudinal images found within a specific patient record.

the questions that previous work did not explicitly answer including: How many 2D images should be annotated to extend the automatic detection process to the complete volumetric dataset? How many longitudinal 2D images should be annotate to broaden the detection process to temporal 3D images? Which multi-modal images should be considered when extracting features to detect the pathological concept "edema"? Which image features better capture pathological concepts such as "edema"?

We hope that answers to such questions can provide insight on how to go about designing effective algorithms for structuring and quantifying the multi-modal longitudinal image sets.

2 Approach

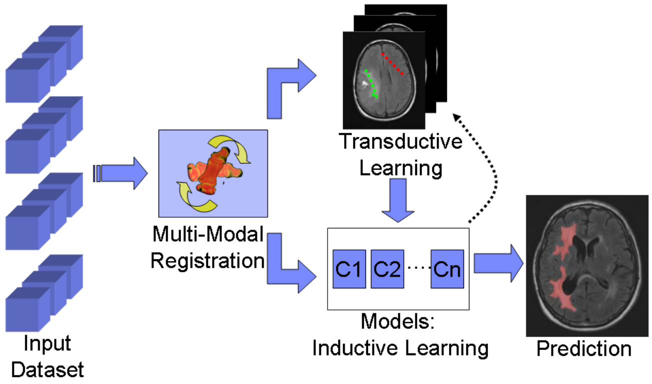

Fig. 3: Pipeline used to combine transductive and inductive learning mechanisms to accurately detect pathological concept in MR images. The transductive method provides pseudo-ground truth labels from which an inductive machine is trained.

To answer the questions stated above, we first designed a framework for automatic detection of pathological concepts in longitudinal multi-modal image sets. The system uses multi-modal cues in conjunction with transduction and inductive learning techniques to automatically extract regions corresponding to pathological concepts in MR images. In this work, we focus on the concept of "edema"; however the approach is able to generalize to other concepts of interest.

Figure 3 shows a diagram of the system. First, the multi-modal set of images belonging to every timepoint are aligned (registered) into a common coordinate system, where meaningful multi-modal image features can be extracted. A combination of linear and affine multi-resolution registration techniques is employed during this step to compensate for size, resolution, and positional differences. Second, a set of images are presented to an expert where seed positive and negative labels for the pathological concept under consideration are manually selected. These seed regions are then leveraged and propagated by the transductive inference to assign labels to all pixels in the image. In transductive learning, the data set $\mathcal{D} = [\mathcal{D}_l, \mathcal{D}_u] \in \mathcal{P}_{XY}$ consists of the labeled training set $\mathcal{D}_l = (\mathbf{x}_n, y_n)_{n=1}^{l}$ and the unlabeled set $\mathcal{D}_u = (\mathbf{x}_n, \hat{y}_n)_{n=l+1}^{l+u}$ with $\hat{y}_n$ unknown. Usually, $l \ll u$. The goal of transductive inference is to find a smooth function f in input space onto the output space, such that $f(x_i)$ is close to the associate y_i on the training set. This function could then be employed to associate labels to the elements of the unlabeled set $\mathcal{D}_u$. In [2] we proposed the *Transductive Conditional Mixture Naive Bayes* (T-CMNB) learning machine for spatial multi-modal generative classification casted as an interactive segmentation problem with minimal expert intervention. The multi-modal mixture assumption on each covariate feature dimension and spatial regularization constraints in T-CMNB allowed us to explain more complex distributions required for spatial classification in multi-modal imagery.

The resulting classification is then used as pseudo-ground truth to train an inductive model. For each training point in the transductive model, a combination of first- and second-order statistics are estimated to create a multi-dimensional descriptor. Histogram features including mean, standard deviation, and skewness are extracted from each training point in conjunction with textural features such as energy, contrast, and correlation. Those set of features are combined and used as the characteristic descriptor for each training point. SVMs [3] are then used to learn an inductive and more generic classification model. Finally, the inductive model is used to automatically identify the particular medical concept on new input data. At this step, we performed a number of experiments to answer some of the questions about how much annotation is needed to extend the automatic identification process to volumetric data and longitudinal 3D images.

3 Experiments and Results

To study how minimum user input can be used to quickly learn and classify new data, our framework was tested with a collection of multi-modal and temporal MRI studies of patients with high-grade glioma brain tumors.

3.1 Multi-modal Image Features

Radiologists often analyze the gray-level values of FLAIR images to determine the *edema* progression. Our first experiments were to determine which image features and modalities improve the identification of the pathological concept "edema". We extracted first- and second-order statistics from each training point. By training and testing a large set of images in a round-robin fashion, we found that even when FLAIR images are used, textural features improved the accuracy of the classification by at least 1%. In addition we found that always when the image features were extracted using FLAIR together with T1 and T1 with contrast, the accuracy of detecting the pathological concept "edema" increased above that of FLAIR alone. This shows that despite FLAIR being the primary image modality used by radiologists to identify "edema", multi-modal protocols appear to have hidden cues which can improve the automatic identification process of edema regions. In addition, although image intensity levels in FLAIR are the primary image features used by radiologists to determine edema regions, our results show that the combination of histogram statistics with textural properties always performed better than just using intensity-based image features.

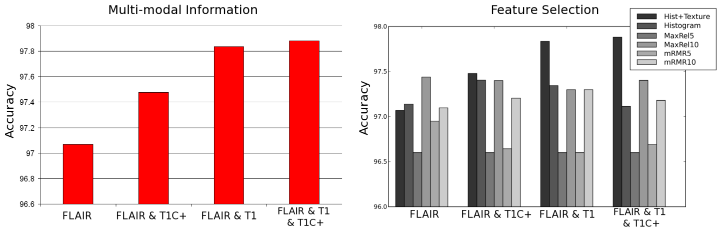

Fig. 4: (left) Benefits of using the multi-modal cues to automatically detect edema regions. When the image features were extracted using FLAIR, T1, and T1 with contrast, the accuracy of detecting the pathological concept "edema" always increased. (right) After training a set of inductive models for each multi-modal combination, we found that the combination of histogram- and texture-based features always performs better than histogram-based features alone.

To further answer the question regarding which image features to extract, two feature selection techniques were applied to the complete feature vector: *maximum relevance* and *minimum Redundancy Maximum Relevance (mRMR)*[4]. We found that when only FLAIR images were used, the maximum relevance features were able to improve the automatic detection process; however when using any combination of multi-modal features, the aggregate feature vector always performed better than any subset of features. Figure 4(left) shows the results of training ten models for each protocol combination, note that multi-modal features always improve classification. Figure 4(right) shows the results of training ten models with different set of image features. From the plot we can see that with any protocol combination, histogram- and textural-based image features always performed better than intensity-based features and that with multi-modal features the aggregate feature vector always outperforms any subset of features.

3.2 Number of Training Images

Obtaining insight regarding the number of images that need to be annotated is an important question faced during the design of any CAD system. First, an effective concept detection system should use minimum user annotation to learn anatomical properties. Second, since the MR parameters can change significantly between scans and/or between different scanners, learning within a single image can introduce a significant amount of mis-classification.

Given a 3D MRI image, we would like to determine how many 2D images (slices) need to be annotated to infer the rest of the volume (30+ slices). For this experiment, 20 datasets of multi-modal MRI images were used. After testing over 225 different combinations, we found the threshold number to be three (3). That is, we found that on average the complete volume can be accurately classified based on the annotation of three images. Figure 5(left) shows a stacked plot with the average accuracy and variability with different number of annotated images. Note that when only one or two images are annotated, the classification results are not that accurate and include a significant amount of variability. In addition, when more than three images are used, the improvements of the classification are not that significant.

A question that arises from these experiments is which three images should be annotated. We performed several experiments annotating images from different part of the volume and did not find any specific pattern of the effects of picking a set of images versus randomly annotating some of the images. Therefore, we can conclude that any three different images can be annotated when accurate prediction of the complete volume is needed.

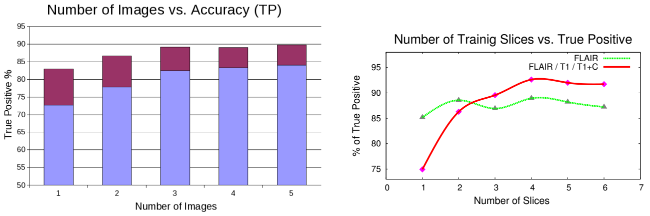

Fig. 5: (left) We found that when multi-modal image features are used, annotating at least three images provides enough information to learn the characteristic properties of the concept under consideration. (right) When a single image is annotated, the prediction of other 2D images is better determined if features from a single image modality are used. However, when more than two images are annotated, multi-modal images features always perform better than a single image modality.

In situations where prediction of the complete volumetric data is needed but less than two images are annotated, the use of a single image protocol (i.e. FLAIR) gives better results than using multi-modal image features. However, once more than two images are annotated, multi-modal image features always improve the automatic detection process. Figure

5(right) shows some of our results.

We believe that the reason behind the threshold value three is due to the high anisotropic sampled imagery frequently found within MRI studies and the uncertainty introduced by the registration. Since there is not a one-to-one mapping between different image modalities due to the large spacing between slices, multi-modal image features from a single image might not capture the complete characteristic properties of the concept under consideration, however, if more than two images are annotated, the benefits of multi-modal features then become clear.

		Number of Temporal Images			
		1T	*2T*	*3T*	*4T*
Number of Slices per Volume	*1S*	71.22	79.77	82.68	83.46
	2S	71.85	81.46	84.32	85.56
	3S	72.99	82.02	86.46	86.73

Table 1*:* Results of annotating different number of slices and 2D temporal images. From the results we can see that to extend the automatic detection process to longitudinal dataset, annotating temporal data is crucial. In particular, we can see that once more than a single temporal dataset is annotated, there is a significant improvement on the overall accuracy. We also found that annotating three images from two different timepoints (i.e. six images) had about the same accuracy than annotating a single image from three different timepoints. In addition, by annotating more training points and temporal images the accuracy of the classification can increase, however after three timepoints, the rate at which the improvement occur becomes minimal. Thus, we can conclude that three is good threshold that can be considered during the design of CAD systems.

Given the significant differences commonly found between consecutive MR scans, longitudinal MR images present a great amount of intensity differences and noise. If a pattern is learned from a single timestep, the model frequently cannot be extended to other temporal 3D images or the results will be highly dominated by mis-classification. How many temporal images are required to create an accurate detection system to classify volumetric data over time? We found that to guarantee an 80% or better classification and detection of the pathological concept edema over longitudinal data, image from three different timepoints should be annotated. Table 1 presents our results.

Figure 6 shows our results with a particular temporal dataset. Our results highlight the importance of training a system with images of at least three different timepoints. Note the significant amount of variability that can occur when images from a single timepoint are annotated. However, when three different timepoints are used, the accuracy increases and the results are more uniform. On average, by annotating three images of three different timepoints, we were able to automatically identifying the edema volume over time with 88.6% true positives.

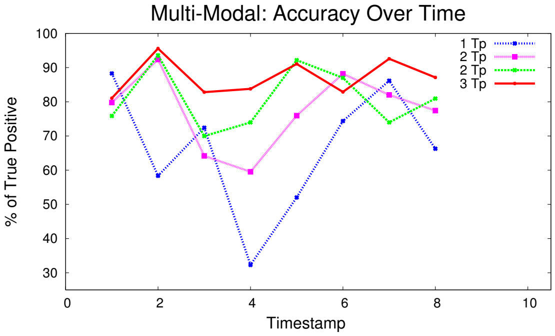

Fig. 6*:* Comparison of the effects of training a system with a different number of temporal images. Note that once three temporal images are used for training, the accuracy of the longitudinal classification is mostly uniform.

4 Discussion and Conclusion

This paper shows how the combination of transductive and inductive learning techniques can enable the development of flexible concept detection systems for longitudinal multi-modal imagery with minimum user interaction. In addition, it shows how minimum user input can be used to effectively leverage entire multi-modal longitudinal image sets. In particular, we showed that, multi-modal image features improve the overall classification results when the annotations comes from more than two images. We also present results about the number of annotated images that are needed to guarantee accurate classification results within a volume and/or with temporal 3D images. We believe that such insight will help with the design of flexible CAD systems and effective algorithms for quantifying multi-modal longitudinal datasets.

5 References

[1] E.D. Angelini, J. Atif, J. Delon, E. Mandonnet, H. Duffau, and L. Capelle, "Detection of glioma evolution on longitudinal mri studies," *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pp. 49–52, April 2007.

[2] Noah Lee, Jesus J. Caban, Shahram Ebadollahi, and Andrew F. Laine, "Interactive segmentation in multi-modal medical imagery using a bayesian transductive learning approach," *SPIE Medical Imaging*, vol. 7260, 2009.

[3] Vladimir N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[4] Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.