A Visual Latent Semantic Approach for Automatic Analysis and Interpretation of Anaplastic Medulloblastoma Virtual Slides

Angel Cruz-Roa¹, Fabio González¹, Joseph Galaro², Alexander R. Judkins³, David Ellison⁴, Jennifer Baccon⁵, Anant Madabhushi², and Eduardo Romero¹

¹ BioIngenium Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
 ² Rutgers, Department of Biomedical Engineering, Piscataway, NJ, USA

³ Children Hospital of L.A., Department of Pathology Lab Medicine, Los Angeles, CA, USA ⁴ St. Jude Children's Research Hospital from Memphis, TN, USA

⁵ Penn State College of Medicine. Department of Pathology. Hershey. PA. USA

Abstract. A method for automatic analysis and interpretation of histopathology images is presented. The method uses a representation of the image data set based on bag of features histograms built from visual dictionary of Haarbased patches and a novel visual latent semantic strategy for characterizing the visual content of a set of images. One important contribution of the method is the provision of an interpretability layer, which is able to explain a particular classification by visually mapping the most important visual patterns associated with such classification. The method was evaluated on a challenging problem involving automated discrimination of medulloblastoma tumors based on image derived attributes from whole slide images as anaplastic or non-anaplastic. The data set comprised 10 labeled histopathological patient studies, 5 for anaplastic and 5 for non-anaplastic, where 750 square images cropped randomly from cancerous region from whole slide per study. The experimental results show that the new method is competitive in terms of classification accuracy achieving 0.87 in average.

1 Introduction

This paper presents a new method, ViSAI, for automatic analysis and interpretation of histopathological images. The method comprises three main stages: learning of an image representation based on bag of features (BOF), characterization of the rich visual variety of a histopathological image collection using visual latent topic analysis, and connection of visual patterns with the semantics of the problem using a probabilistic classification model. The learnt probabilistic model is applied to new images, and the class posterior probability is used to determine the corresponding class. The method is applied to the classification of a type of brain cancer called medulloblastoma, which is one of the most common types of malignant brain tumors [10]. In adults, the disease is rare whereas in children the incidence amounts to a 25% of all pediatric brain tumors. Tumor classification of medulloblastoma is currently performed by microscopical examination and no quantitative image analysis and classification tools are so far available for this task. Different histologic types of medulloblastoma have different prognosis.

N. Ayache et al. (Eds.): MICCAI 2012, Part I, LNCS 7510, pp. 157-164, 2012.

[©] Springer-Verlag Berlin Heidelberg 2012

The differential diagnosis is a hard task and tends to be qualitative. Determine the subtypes of medulloblastoma are difficult to establish and subject to inter-observer variability because of the similarity between the two basic histologic subclasses: anaplastic and non-anaplastic and their similarity with a long list of differential diagnoses. The anaplastic medulloblastoma have worse prognosis and this is mostly characterized by the presence of large, irregular cells that lack organization and in some cases attempt to wrap around each other. The therapeutical management changes radically depending on the subtype of medulloblastoma so that histopathological diagnosis is useful in determining the potential outcome of the disease. Hence computerized and quantitative image analysis tools are useful in this kind of problem for better estimation of medulloblastoma subtype allowing potentially make better prognostic decisions.

Recent investigations [9] have pointed out the importance of provide computerized and automatic image analysis tools as support to diagnosis and prognosis for different types of cancer. Recent work in histologic image analysis has explored the use of automated methods for breast cancer and prostate cancer diagnosis and grading [4,8] using other images modalities. Galaro et al. [6] classified anaplastic and non-anaplastic medulloblastoma subtypes, using a BOF of Haar wavelet coefficients as local pixelbased descriptors. The authors of [6] reported on average a classification accuracy of 0.80. The above works have essentially developed tools to improve the quality of diagnosis in terms of objective support and some level of quantification. Bag of features approach is an image content representation commonly used in computer vision area which comprises three main stages: local feature extraction, dictionary construction, and image representation. This approach have been successful adapted and applied in histology images previously in image analysis and classification tasks [2,3]. On the other hand, latent topic analysis is an approach widely used in text document collection analysis to find the semantic topics related with these documents. The representative techniques are latent semantic analysis (LSA), pLSA [7] and latent Dirichlet analysis (LDA) [1]. Both pLSA and LDA suppose a generative model for documents. The main assumptions here are: first, the image content of image could be represented by an appropriate set of visual words from dictionary learnt from whole image collection represented by a good visual word representation, and second, the large visual variability in the image collection is generated from a relatively small set of visual latent factors. Under our analysis, visual latent factors correspond to high-level patterns that mix sets of visual words that co-occur with high frequency in the collection. Semantics is then linked to visual latent factors coding the relationship between the visual appearance and particular classes.

The new method in this paper addresses the problem of automated classification of histological slides. The method provides higher accuracy coupled with an interpretation of classification results for the expert in terms of semantic domain, rather than being a black-box approach. In particular, the main contributions of the present work are:

- A strategy to characterize the rich visual content of histology images combining an image representation, based on BOF, and a texture local descriptor for image patches based on a Haar wavelet transform.
- A visual latent-topic analysis model for finding higher-level visual patterns that combine multiple visual words.



Fig.1. Overall scheme for the visual latent semantic analysis for automatic classification and interpretation method

• A method that is able, not only to globally classify a virtual slide, but also provides interpretability determining the most important visual patterns associated with such classification, and identifying their location in the original image.

2 Visual Latent Semantic Analysis of Medulloblastoma Images

ViSAI method performs an implicit semantic identification of the visual patterns that characterize each class (anaplastic and non-anaplastic). The strategy is built upon an image representation, based on BOF, and a visual latent semantics analysis, based on probabilistic latent semantic analysis (pLSA). The overall scheme is depicted in Figure 1.

2.1 Image Representation

Step 1: Local Features Extraction. square patches of 50 pixels, that is the minimum spatial resolution that covers a cellular unit (nucleus and cytoplasm), are extracted using a dense grid strategy with an overlap of 80%. Each patch is represented by two different local descriptors, raw-blocks (block) and Haar wavelet coefficients (haar). The raw-block descriptor corresponds to luminance values of pixels in the patch, whereas the Haar descriptor corresponds to the filter responses of two scales of a Haar-wavelet-based wavelet transform. Luminance differences are reduced by applying, before feature extraction, a mean-zero-variance-one normalization.

Step 2: Dictionary Construction. the dictionary is built by applying a conventional k-means clustering strategy to a sample of image patches from the training set. The number of clusters, *k*, corresponds to the dictionary size, and each centroid corresponds

to a visual concept. An important difference with a previous texton-based approach [6] is that our dictionary is constructed from images randomly selected from both classes, regardless of whether it belongs to the class anaplastic or not, by which the visual concepts are mixed up within a very heterogeneous dictionary.

Step 3: Novel Image Representation. each image is then represented by a *k*-bin histogram, capturing the frequency of each visual word in the image given the local descriptor and the visual dictionary. A collection of images can be represented then by a matrix, *X*, where each column corresponds to an image and each row to a visual word. If each column is normalized (L1 normalization), each position in the matrix, $X_{i,j}$, corresponds to the conditional probability of finding a visual word w_i in an image d_j , i.e., $p(w_i|d_j)$.

2.2 Visual Latent Semantic Image Analysis

Step 1: Visual Latent Factors Analysis. The new method uses an approach similar to pLSA that assumes that the presence of different terms (visual words) in a set of documents (images) can be explained by the presence of a reduced set of hidden variables called latent factors. Specifically, the conditional probability $P(w_i|d_j)$ can be represented as:

$$P(w_i|d_j) \approx \sum_k P(w_i|z_k) P(z_k|d_j) \tag{1}$$

where $P(w_i|z_k)$ is the conditional probability of visual word w_i given latent factor z_k and $P(z_k|d_j)$ is the conditional probability of a latent factor z_k given the image d_j . The latent factors can be found by solving an optimization problem that looks for P(W|Z)and P(Z|D) that minimizes the Kullback-Leibler divergence between the left and right sides of Equation 1 [5]. In our case, we solve the optimization problem by modeling it as a matrix factorization problem: P(W|D) = P(W|Z)P(Z|D), where P(W|D) = X is the histogram representation of the image collection discussed in Subsection 2.1, P(W|Z)contains the latent factors represented in terms of visual words, and P(Z|D) contains the representation of the images in the collection, in terms of latent factors.

Step 2: Semantic Latent Factors. Given a set of annotated training images, each image, d_i , is associated with one of two classes: anaplastic ($C_1 = A$) or non-anaplastic ($C_2 = NA$). If, for instance, the image is anaplastic then $P(C_1|d_i) = 1$ and $P(C_2|d_i) = 0$. Following the same reasoning as in previous subsection this probability can be expressed in terms of the latent factors as follows:

$$P(C_i|d_j) \cong \sum_k P(C_i|z_k) p(z_k|d_j)$$
⁽²⁾

The probability $P(C_i|z_k)$ effectively connects each latent factor with the semantics of the problem represented in terms of the biological concept associated with each class. These probabilities can be found applying a matrix factorization algorithm [5] that fixes P(Z|D) from the previous step and looks for P(C|Z) such that: P(C|D) = P(C|Z)P(Z|D).

Step 3: New Image Classification. A new image, d^* , is first represented using the strategy discussed in Subsection 2.1. This produces a normalized histogram $P(W|d^*)$.

The image is then represented in terms of latent factors (z_k) finding $P(Z|d^*)$ such that: $P(W|d^*) \cong P(W|Z)P(Z|d^*)$, where P(W|Z) was previously found in step 1 solving Equation 1. Then, the posterior class probability of the image $P(C|d^*)$ is calculated using: $P(C|d^*) \cong P(C|Z)P(Z|d^*)$, where P(C|Z) was previously calculated in step 2 solving Equation 2 and obtained $P(Z|d^*)$. Finally, the class assigned to the new image is the one with the maximum class posterior probability.

3 Experimental Results and Discussion

3.1 Experimental Design

The dataset comprises 10 labeled histopathological cases from St. Jude Children's Research Hospital, which 5 are anaplastic and 5 are non-anaplastic. Each slide is a whole virtual slide of 80000×80000 pixels with one or more cancerous regions with a large tumoral variability, manually annotated by a neuro-pathologist. For every slide, 750 individual images of 200×200 pixels non-overlapping where extracted uniformly at random from these cancerous regions, resulting in a database of 7500 different images: half of them anaplastic. The local feature extraction is carried out as was described in Subsection 2.1 for two local descriptors *block* and *haar*. The dictionary size was tested with different sizes, 10, 20, 40, 80, 160 and 320 visual words. Finally the number of latent factors was fixed to the dictionary size since this amounts to a maximum number of elemental building blocks (visual words) to be represented by visual semantic concepts (latent factors). The probabilistic analysis described in Subsection 2.2 provides a probability value associated with the semantic importance of each of the initial visual features so that these can be grouped together (into latent factors) by similar levels of relevance per class (Equations 1 and 2).



Fig. 2. Dictionary size vs average accuracy for both local features (block and haar)

In this sense, three different experiments were proposed to evaluate each parameter of the proposed method. First, determine the visual words appearance from dictionaries and impact of dictionary size for each local descriptor (block or haar) evaluation average accuracy in test data set over multiple trials of crossvalidation. Second, evaluate the classification performance of the proposed image representation (Subsection 2.1) for each local descriptor using the proposed method (Subsection 2.2) compared with a k-NN classifier. Third, a visual map-

ping of high-level concepts to identify spatial regions associated by each one (anaplastic and non-anaplastic).

3.2 Visual Dictionary Construction

Figure 2 shows the impact of the dictionary size on the validation average accuracy. Overall, haar-based dictionary outperforms the raw-block representation by about 7%.

For both representations the best performances were obtained with the largest dictionary size obtaining 0.77 (blocks) and 0.86 (haar) in average accuracy. Visual words can be related to a particular class calculating the posterior class probability. Table 1 shows the 20 visual words with highest posterior class probability for both classes, for the two types of representation. In all the cases the visual dictionary is composed basically of smooth samples of nuclei and cells in different orientations and arrangements, a result that matches perfectly with a very basic principle in pathology analysis which introduces the cell as the basic unit of information. Also some differences between the visual words associated to each class can be observed, for both (block and haar), non-anaplastic patterns are more homogeneous. This is consequent with the biological definition of anaplasia, where cells are undifferentiated and nuclei shapes and sizes have high variability.

Table 1. A sample of the visual dictionaries obtained using different local features. Second column shows the 20 visual words with highest probability by class (A:anaplastic, NA:non-anaplastic) for each kind of local feature in a dictionary size of 160.



3.3 Classification Performance

For each of multiple trials of cross-validation, we used 4 anaplastic and 4 non-anaplastic slides for training the visual semantic model and 1 anaplastic and 1 non-anaplastic slides for testing. For comparison, the *k*-NN classifier was employed. An optimal value for *k*, 10, was found by cross validation such as was suggested in [6]. Different dictionary sizes and patch representations were evaluated to determine the best configuration for this classification task in terms of classifier accuracy, specificity, and sensitivity. The results are presented in Table 2, which shows the performance obtained with each image representation strategy and classification algorithm. Clearly the haar-based representation outperforms the block-based representation, independent of classifier choice. The new classification method is competitive with respect to the *k*-NN, a classifier used in [6], the improvement was 3.6% in accuracy (haar), 12.7% in sensitivity (block) and 10.1% in specificity (haar).

 Table 2. Classification performance in terms of accuracy, specificity and sensitivity for dictionary size of 320

	ViSAI			k-NN		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
block	0.78	0.89	0.67	0.80	0.79	0.81
haar	0.87	0.86	0.87	0.84	0.88	0.79



Fig. 3. Visual semantic maps for test images from a cross validation trial. Images in the first column were correctly classified (A-anaplastic and NA-non-anaplastic). See Subsection 3.4 for description.

3.4 Visual Pattern Mapping and Interpretation

The methodology presented in this paper allows for new unknown images to determine the spatial probability maps for each class using conditional probabilities. Figure 3 illustrates the semantic visual maps generated for two unknown test images in a particular trial of cross validation. Rows one and two show images of the same size used in training (200×200) , whereas rows three and four show a larger field of view (2000×2000) . The center column shows the corresponding BOF histogram representation for each image (i.e. occurrence frequency of each visual word of the dictionary in the image). Bin (visual words) has been sorted according to the posterior class probabilities: visual words with a higher class probability of being anaplastic are at the left, and visual words with a higher class probability for non-anaplastic are at the right. The third column shows the posterior class probability. The fourth column shows a map that indicates how likely a particular region of the image is to be related to one of the semantic classes. This is depicted by the log-odds of the posterior class probabilities to emphasize the differences. A positive value (*red*) indicates a higher probability for class anaplastic, a negative value (blue) indicates a higher probability for class non-anaplastic. An advantage of the BOF image representation is that new method is able to scale towards larger image size as shown in the rows three and four of Figure 3, where the correct prediction of classes and spatial location of semantic regions is more challenging.

4 Concluding Remarks

The new method ViSAI was successfully applied to the challenging problem of automatic discrimination between two subtypes of medulloblastoma using computer derived image features extracted from whole slides. The goal of our method was not just to improve classification accuracy, but to provide interpretable results using latent semantic analysis and BOF image representation. The method provides an interpretation layer that helps out the pathologist to determine the type of patterns present in the sample under examination with a potential improvement of diagnostic significance. This attribute is particularly useful in problems relating to stratification of the disease where the distinction between disease sub-classes might reside in very subtle visual cues. The experimental results are promising and indicate that visual latent semantic analysis has potential as a tool for analyzing the complex visual patterns exhibited by histopathology images.

References

- Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Mach. Learn. 3, 993–1022 (2003)
- Cruz-Roa, A., Caicedo, J., González, F.: Visual pattern mining in histology image collections using bag of features. Artif. Intell. Med. 52(2), 91–106 (2011)
- 3. Cruz-Roa, A., Díaz, G., Romero, E., González, F.: Automatic Annotation of Histopathological Images Using a Latent Topic Model Based on Non-negative Matrix Factorization. J. Path Inform. 2(1), 4 (2011)
- Dalle, J.R., Leow, W.K., Racoceanu, D., Tutac, A.E., Putti, T.C.: Automatic breast cancer grading of histopathological images. In: EMBS 2008, pp. 3052–3055 (2008)
- 5. Ding, C., Li, T., Peng, W.: On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. Comput. Stat. Data An. 52(8), 3913–3927 (2008)
- Galaro, J., Judkins, A., Ellison, D., Baccon, J., Madabhushi, A.: An integrated texton and bag of words classifier for identifying anaplastic medulloblastomas. In: EMBC 2011, pp. 3443–3446 (2011)
- 7. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42, 177–196 (2001)
- Kwak, J.T., Hewitt, S.M., Sinha, S., Bhargava, R.: Multimodal microscopy for automated histologic analysis of prostate cancer. BMC Cancer 11(1), 62 (2011)
- Madabhushi, A., Agner, S., Basavanhally, A., Doyle, S., Lee, G.: Computer-aided prognosis: Predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. Comput. Med. Imag. Grap. 35, 506–514 (2011)
- Roberts, R.O., Lynch, C.F., Jones, M.P., Hart, M.N.: Medulloblastoma: A population-based study of 532 cases. J. Neuropath. Exp. Neur. 50(2), 134–144 (1991)