HIERARCHICAL TASK-DRIVEN FEATURE LEARNING FOR TUMOR HISTOLOGY

Heather D. Couture[1], *J.S. Marron*[23], *Nancy E. Thomas*[34], *Charles M. Perou*[35], *Marc Niethammer*[16]

Department of Computer Science[1], Department of Statistics and Operations Research[2],
Lineberger Comprehensive Cancer Center[3], Department of Dermatology[4],
Department of Genetics[5], Biomedical Research Imaging Center[6]
University of North Carolina, Chapel Hill, NC

ABSTRACT

Through learning small and large-scale image features, we can capture the local and architectural structure of tumor tissue from histology images. This is done by learning a hierarchy of dictionaries using sparse coding, where each level captures progressively larger scale and more abstract properties. By optimizing the dictionaries further using class labels, discriminating properties of classes that are not easily visually distinguishable to pathologists are captured. We explore this hierarchical and task-driven model in classifying malignant melanoma and the genetic subtype of breast tumors from histology images. We also show how interpreting our model through visualizations can provide insight to pathologists.

Index Terms— histology, tumor, image classification, feature learning

1. INTRODUCTION

Pathologists diagnose cancer and predict prognosis by examining histology images of tumor tissue. Hematoxylin and eosin (H&E) is the most widely used set of stains and turns nuclei blue and cytoplasm pink. From this cell-level view of tumor tissue, pathologists look for signs of tumor progression including irregularly shaped nuclei and lack of cell specialization. With the further information provided by gene expression, tumors can now be grouped into clinically relevant subtypes to aid treatment decisions [1]. However, gene expression ignores the spatial arrangement of tumor tissue. It is only through histology images that we are able to analyze the cytological and architectural structure, which describe local-cell level properties and larger-scale organization, respectively.

Histological analysis presents many challenges due to variations in staining and biological heterogeneities. Each tissue type has specialized structures, making hand-crafted features developed for one type difficult to apply to another. Tumors from different genetic subtypes may also appear similar, requiring features that capture their subtle differences.

Our analysis focuses on two specific applications: diagnosis of melanoma (skin cancer) and subtyping of breast tumors. While the current standard for diagnosis involves histological review by a pathologist, breast tumor subtypes are not known to be distinguishable by pathologists from H&E images alone. We hope to determine whether these subtypes manifest morphologically and learn properties that can distinguish them.

The contributions of this work are as follows: 1) We capture biologically-relevant features by operating on the hematoxylin and eosin stain intensities extracted from histology images. 2) Task-driven dictionary learning discovers the subtle differences between tissue classes. 3) Architectural properties of tissue are captured with a hierarchical model. 4) Our visualizations provide insight into which tissue regions contribute to the overall classification of a sample.

2. BACKGROUND

Most automated analysis of histology follows a general pipeline of first segmenting nuclei, then characterizing color, texture, shape, and spatial arrangement properties of cells and nuclei [2, 3, 4]. These hand-crafted features are time-consuming to develop and do not adapt easily to new data sets. More recent work has begun to learn appropriate features directly from image patches [5, 6, 7].

Sparse coding has been shown to produce superior image classification results in comparison to other encoding methods when used in a single-level dictionary learning framework [8]. Additional modifications to improve the discrimination capability of the dictionary involve combining the reconstruction and classification error into a single objective function [9, 10, 11]. This helps to capture fine-grained differences between classes. Mairal et al. applied this to improve recognition of hand-written digits [11]. We extend it to a hierarchical dictionary learning framework for classifying large images.

Also making use of hierarchical learning, recent successes in deep learning have been shown in recognizing hand-written digits and objects [12, 13], and also applied to histology for specific tasks such as mitosis detection [14]. We expect the stronger encoding mechanism provided by task-driven sparse codes will lead to better classification and plan to provide a detailed comparison in future work.

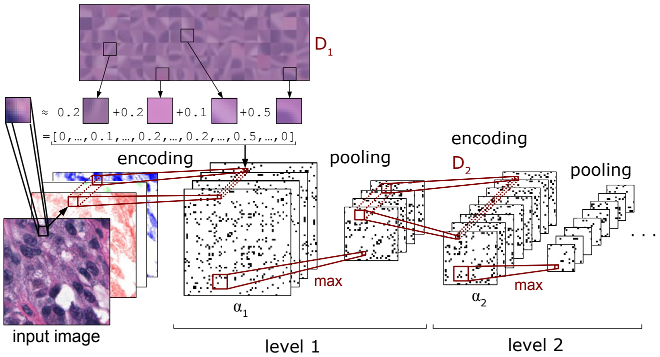

Fig. 1. Images are first color normalized and the hematoxylin, eosin, and residual stain channels extracted. Each image patch is encoded using a dictionary. Following encoding, a max pooling operation downsizes the image. By alternating encoding and pooling layers, a hierarchy of features is formed

3. APPROACH

This section outlines the steps to learn hierarchical task-driven dictionaries and apply them to encode images for classification. Fig. 1 provides an overview of image encoding.

3.1. Pre-processing

Color and intensity normalization is first applied to standardize the appearance across slides, countering effects due to different stain amounts and protocols, as well as slide fading. We use the method by Niethammer et al. that estimates the stain vectors for hematoxylin and eosin and normalizes each image [15]. The resulting stain intensity channels are then used as input to the rest of our algorithm.

The next step of learning a dictionary will operate on square patches extracted from training images. We first apply mean centering and a Zero-phase Component Analysis whitening step to reduce the redundancy of individual patches by making the features uncorrelated and to give each feature a similar variance [16]. This centering and whitening process is applied prior to encoding for each level of the hierarchy.

3.2. Unsupervised Dictionary Learning

We use sparse coding to learn a dictionary of features to represent image patches. The elastic net formulation looks for a small number of dictionary elements that, through a linear combination, can reconstruct a given image patch. This optimization is formulated as

$$\alpha^*(x, D) = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2}\|x - D\alpha\|^2 + \lambda_1\|\alpha\|_1 + \lambda_2\|\alpha\|_2^2 \quad (1)$$

for image patch x, dictionary D, and coefficients α, in which we minimize the reconstruction error while encouraging a sparse solution with an ℓ_1 norm and adding stability in the case of correlated variables with the ℓ_2 norm. Due to the computationally intensive nature of evaluating the elastic net, we perform this on a GPU.

The dictionary is computed from a set of whitened image patches by first initializing with random patches. Alternating optimizations are used to minimize the objective in (1), summed over a set of training patches. We use the online batch implementation by Mairal et al. [17].

3.3. Task-Driven Dictionary Learning

The discriminating power of the dictionary is improved by incorporating image label information into the dictionary learning framework. By minimizing the logistic loss, we learn a linear discriminant for two classes based on the sparse encodings of individual image patches. Although we focus on binary classification here, the logistic could be replaced with the softmax function to generalize to multiple classes.

We initialize the dictionary using the unsupervised dictionary learning procedure detailed in the previous section. An initial linear classifier is learned using logistic regression on the encodings $\alpha^*(x, D)$ of a set of training patches $x_1, ..., x_N$. The classifier is defined by a separating hyperplane w such that if $w^T\alpha^*(x, D) + w_0 > 0$, patch x is predicted to belong to class 2, and class 1 otherwise. The formula $1/[1 + e^{-(w^T\alpha^*(x,D)+w_0)}]$ predicts a probability indicating how likely the patch is to belong to class 2.

In improving the dictionary and classifier, the logistic loss objective function we use is as follows:

$$f(D, w) = \min_{w,D} \sum_{n=1}^{N} \log[1 + e^{-y_n(w^T\alpha^*(x_n,D)+w_0)}] + \frac{\nu}{2}\|w\|_2^2$$

where y_n is the class label (-1 or 1) associated with each patch x_n, w defines the hyperplane separating the two classes, $\alpha^*(x, D)$ is defined in (1), and parameter ν controls the regularization. We optimize this objective by stochastic gradient descent, updating D and w as

$$D \leftarrow D - \gamma\nabla_D f(D, w) \qquad w \leftarrow w - \gamma\nabla_w f(D, w)$$

where γ is the learning rate, and $\nabla_w f(D, w)$ and $\nabla_D f(D, w)$ are calculated from the logistic loss function $f(D, w)$ using $\nabla_D\alpha^*(x, D)$ derived by Mairal et al. [11].

3.4. Hierarchy of Features

Now that we can form dictionaries of learned features and use them to encode images, we turn to the problem of forming a feature hierarchy to capture more abstract and larger scale properties. After densely encoding every patch in an image, a max pooling operation is applied in which, for each $m \times m$ region, we take the maximum encoded value for each feature. This has the effect of providing local translation invariance and downsizing the representation to enable capture of larger-scale properties by the next level. Encoding and max pooling operations are alternated to form a feature hierarchy.

3.5. Classification

At this point, each image is represented by a set of sparse encodings of features from each level of the hierarchy and we must predict the image-level class. We can apply the logistic regression classifier to each image patch or summarize the encodings themselves and train a new classifier. We compare four image-level classification methods:

1. The mean of the patch probabilities over the image.
2. The sum of the log of patch probabilities (equivalent to multiplying probabilities).
3. A new logistic regression classifier to operate on quantile functions summarizing patch probabilities.
4. A Support Vector Machine (SVM) to operate on histograms of the patch encodings (equivalent to a mean pool of the encodings).

For the first two options, we found it to work best if a threshold to separate the two classes is learned on the training data.

3.6. Implementation Details

The procedure used for selecting parameter settings is outlined here. Patch sizes of 9×9, 5×5, and 3×3 and dictionary sizes of 128, 192, and 256 were used for the three levels, respectively, with a 3×3 max pool for each. Dictionary learning requires setting the regularization parameters λ_1 and λ_2 (Section 3.2). We selected λ_1 from 0.25, 0.5, 1.0, and 2.0 as the value that produced the best patch classification accuracy through cross-validation on the training set. We set λ_2 to $\lambda_1/10$ to add some stability to the model, while keeping the ℓ_1 norm as the main mode of regularization. The logistic loss of task-driven dictionary learning requires a regularization parameter ν (Section 3.3). We also learned this from the data as the value from 10^{-6} to 10^1 that produced the greatest patch classification accuracy. During learning, patches are randomly selected from each image and are randomly flipped and/or rotated to add more variety to the data. A learning rate γ of 10^{-5} was found to work with our data sets in combination with a batch size of $500000/N$ patches from each image, where N is the number of training images, and 60, 20, and 15 cycles through the training set for the three levels respectively.

4. EXPERIMENTS

We assess both unsupervised and task-driven dictionary learning as a hierarchy by comparing the classification accuracy on two data sets.

	Melanoma vs. nevi		Breast subtype	
	U	TD	U	TD
Level 1	55.2%	**59.0%**	50.7%	**52.0%**
Level 2	59.8%	**63.9%**	56.4%	**58.0%**
Level 3	59.0%	**70.0%**	51.1%	**54.6%**

Table 1. Patch-level classification accuracy comparing unsupervised dictionaries (U) with task-driven dictionaries (TD) for a 3-level hierarchy.

4.1. Data Set

Our melanoma data set consists of whole slide images in which a pathologist has annotated an average of eight regions containing tumor. 31 of these samples contain varying degrees of dysplastic nevi (benign), while 21 contain melanoma.

Our second data set contains breast tumor samples from a Washington University cohort of patients [1]. These take the form of a tissue microarray with two cores per patient and were imaged at the University of British Columbia. We predict the subtype of the 43 Basal and 42 Luminal A samples.

4.2. Classification Results

In order to assess the importance of both the task-driven and hierarchical components of our model, we set up experiments to measure the patch-level and patient-level classification accuracy using 5-fold cross-validation. Although prediction accuracy on patients is expected to be much greater than that on local patches, both provide a means of validation and the later is important for model interpretation in Section 4.3.

First, using the logistic regression classifiers trained during task-driven dictionary learning, we compute the patch-level classification accuracy before and after the task-driven learning process (Table 1). Both data sets show a consistent improvement of task-driven dictionaries over unsupervised ones. The melanoma data set also shows a consistent improvement from level 1 to 3, with a small decrease in the unsupervised dictionary performance for level 3. The breast subtype results show a significant drop in performance for level 3 for both methods. This data set is much more complex and poses a more challenging problem. Algorithm parameters such as patch size and dictionary size likely need to be better tuned to get better results on this data set.

We also measure the patient-level classification accuracy using each of the methods detailed in Section 3.5 (Table 2). This shows a fairly consistent improvement from level 1 to 3 for the first three methods that summarize the image using the patch classifier. However, the breast subtype results are not as consistent as those for melanoma, likely due to the reasons already mentioned for the patch-level results. The task-driven dictionary method outperforms the unsupervised dictionary on the melanoma data set, but only in some settings on the breast subtype data set. The SVM method on feature

	Melanoma vs. nevi		Breast subtype	
	U	TD	U	TD
1. Mean of patch probabilities				
Level 1	65.5%	53.6%	61.5%	59.3%
Level 2	82.9%	84.4%	64.9%	64.6%
Level 3	84.5%	88.5%	70.1%	62.1%
2. Sum of log of patch probabilities				
Level 1	63.3%	74.7%	64.6%	64.2%
Level 2	84.7%	86.5%	62.4%	63.4%
Level 3	82.7%	88.4%	67.5%	58.6%
3. Logistic regression on quantile of patch probabilities				
Level 1	59.6%	67.5%	72.9%	66.4%
Level 2	79.1%	78.5%	65.7%	63.7%
Level 3	81.1%	82.4%	63.5%	65.6%
4. Linear SVM on histogram of features				
Level 1	86.5%	84.7%	69.8%	71.3%
Level 2	84.7%	84.5%	70.6%	65.4%
Level 3	82.9%	84.4%	68.3%	70.2%

Table 2. Patient-level classification accuracy comparing unsupervised dictionaries (U) with task-driven dictionaries (TD) for a 3-level hierarchy using the four different methods described in Section 3.5.

histograms performs well across the different levels, but does not show an improvement from higher levels.

For comparison, we also tested the set of hand-crafted features developed by Miedema et al. that capture the size, shape, stain intensity, texture, and local spatial arrangement of cells and nuclei [2]. We summarized these measures as the mean and standard deviation of each across all cells in the image and measured the 5-fold cross-validation accuracy using a linear SVM. On the melanoma data set, these hand-crafted features achieved a classification accuracy of 89.9%, only slightly higher than our best feature learning results. For the breast subtype data set, they achieved 69.9% accuracy, only slightly lower than our best feature learning results.

4.3. Model Interpretation

We now turn to the problem of identifying which regions of an image are most associated with each class. Using the logistic regression classifier trained on patches, we can predict the probability that an individual patch belongs to each class (Section 3.3). We form a colormap in which blue indicates class 1 has a higher probability, red indicates class 2, and white is neutral. This is shown for a melanoma image in Fig. 2 and compares the results from unsupervised and task-driven dictionaries for a 3-level hierarchy. These results show that the task-driven dictionary produces a slightly higher confidence in classification for levels 1 and 2, as indicated by slighty more red coloring and less blue. The confidence in melanoma also increases up the levels; however, level 3 shows a decrease in confidence for the unsupervised dictionary.

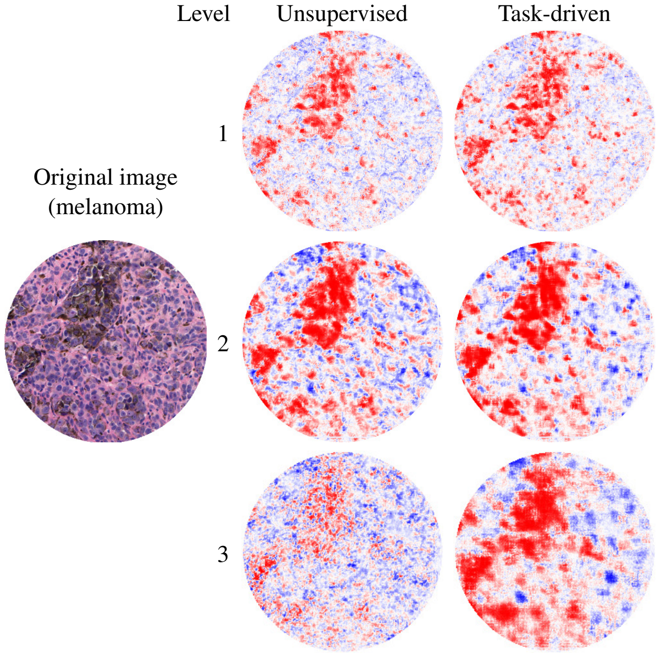

Fig. 2. Relevance maps for a sample image: red indicates features associated with melanoma; blue indicates benign nevi.

5. DISCUSSION

We have shown the application of hierarchical task-driven dictionary learning in predicting the diagnosis of melanoma and the subtype of breast tumors. Our method achieved classification accuracies comparable to that using hand-crafted cell morphology features. The patch-level classification results indicate that the task-driven method has great promise in learning subtle features that distinguish classes. It is not clear to us yet which of the four image-level classification methods is best suited for our task, and so we will continue to refine these methods. We also have plans to compare performance with a convolutional neural network.

Our method for identifying regions of an image most associated with a particular class produced a visualization that highlights important areas of the image. Since interpreting our models in the context of pathology is so important in the application area of medicine, we will continue to investigate other methods for visualization and interpretation of features.

6. ACKNOWLEDGMENTS

This work was supported, in part, by NIH 2P41EB002025 and the University Cancer Research Fund, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, CA112243, CA11243-05S109, and P30ES010126. Data collection for the breast subtype data set was funded by the Strategic Partnering to Evaluate Cancer Signatures. The Tesla K40 GPU was donated by the NVIDIA Corporation.

7. REFERENCES

[1] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, et al., "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of clinical oncology*, vol. 27, no. 8, pp. 1160–1167, 2009.

[2] Jayson Miedema, James Stephen Marron, Marc Niethammer, David Borland, John Woosley, Jason Coposky, Susan Wei, Howard Reisner, and Nancy E Thomas, "Image and statistical analysis of melanocytic histology," *Histopathology*, vol. 61, no. 3, pp. 436–44, Sept. 2012.

[3] Lee A D Cooper, Jun Kong, David A Gutman, Fusheng Wang, Jingjing Gao, Christina Appin, Sharath Cholleti, Tony Pan, Ashish Sharma, Lisa Scarpace, Tom Mikkelsen, Tahsin Kurc, Carlos S Moreno, Daniel J Brat, and Joel H Saltz, "Integrated morphologic analysis for the identification and characterization of disease subtypes," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 317–23, Jan. 2012.

[4] Hang Chang, Gerald V Fontenay, Ju Han, Ge Cong, Frederick L Baehner, Joe W Gray, Paul T Spellman, and Bahram Parvin, "Morphometic analysis of TCGA glioblastoma multiforme," *BMC Bioinformatics*, vol. 12, no. 1, pp. 484, Jan. 2011.

[5] Yin Zhou, Hang Chang, Kenneth Barner, Paul Spellman, and Bahram Parvin, "Classification of histology sections via multispectral convolutional sparse coding," in *Proc. CVPR*, 2014.

[6] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto Osorio Gonzalez, "A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection," in *Proc. MICCAI*, 2013.

[7] Ju Han, Hang Chang, Leandro Loss, Kai Zhang, Fredrick L Baehner, Joe W Gray, Paul Spellman, and Bahram Parvin, "Comparison of sparse coding and kernel methods for histopathological classification of gliobastoma multiforme," in *Proc. ISBI*, Mar. 2011, pp. 711–714.

[8] Adam Coates and AY Ng, "The importance of encoding versus training with sparse coding and vector quantization," in *Proc. ICML*, 2011.

[9] Marc' Aurelio Ranzato and Martin Szummer, "Semi-supervised learning of compact document representations with deep networks," in *Proc. ICML*, July 2008, pp. 792–799.

[10] Zhuolin Jiang, Zhe Lin, and Larry S Davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition.," *IEEE PAMI*, vol. 35, no. 11, pp. 2651–64, Nov. 2013.

[11] Julien Mairal, Francis Bach, and Jean Ponce, "Task-driven dictionary learning.," *IEEE PAMI*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[12] Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean, and Andrew Y. Ng, "Building high-level features using large scale unsupervised learning," in *Proc. ICML*, 2012.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1106–1114.

[14] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Proc. MICCAI*, 2013.

[15] M. Niethammer, D. Borland, J.S. Marron, J. Woolsey, and N.E. Thomas, "Appearance normalization of histology slides," in *Proc. MICCAI, International Workshop on Machine Learning in Medical Imaging*, 2010.

[16] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, 2000.

[17] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *Proc. ICML*, 2009.