PREDICTIVE SPARSE MORPHOMETRIC CONTEXT FOR CLASSIFICATION OF HISTOLOGY SECTIONS

Hang Chang[1,2†], *Paul T. Spellman*[4] *and Bahram Parvin*[1,2,3†]

[1] Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.

[2] Department of Electrical and Computer Engineering, University of California, Riverside, U.S.A

[3] Biomedical Engineering Department, University of Nevada, Reno, Nevada, U.S.A

[4] Center for Spatial Systems Biomedicine, Oregon Health Sciences University, Portland, Oregon, U.S.A

[†] Corresponding authors $\{hchang, b\_parvin\}@lbl.gov$

ABSTRACT

Classification of histology sections from large cohorts, in terms of distinct regions of microanatomy (e.g., tumor, stroma, normal), enables the quantification of tumor composition, and the construction of predictive models of the clinical outcome. To tackle the batch effects and biological heterogeneities that are persistent in large cohorts, sparse cellular morphometric context has recently been developed for invariant representation of the underlying properties in the data, which summarizes cellular morphometric features at various locations and scales, and leads to a system with superior performance for classification of microanatomy and histopathology. However, the sparse optimization protocol for the calculation of sparse cellular morphometric features is not scalable for large scale classification. To improve the scalability of systems, based on sparse morphometric context, we propose the predictive sparse morphometric context in place of the original implementation, which approximates the sparse cellular morphometric feature through a non-linear regressor that is jointly learned with an over-complete dictionary in an unsupervised manner. Experimental results indicates over 50 times speedup compared to our previous implementation with the help of non-linear regressor; while producing competitive performance.

Index Terms— Classification, Sparse Coding, H&E Tissue Section

1. INTRODUCTION

Tumor histology provides a detailed insight into cellular morphology, organization, and heterogeneity. For example, histology sections can be used to identify mitotic cells, cellular aneuploidy, and autoimmune responses. More importantly, if tumor morphology and architecture can be quantified in a large cohort, it will provide the basis for predictive models in a similar way that genomic techniques have identified predictive molecular subtypes.

The main barriers for tissue histology classification resides in two-folds: 1) Extremely large scale of data. Typically, for a single tumor type, there are hundreds of whole mount tissue sections, with size up to 100,000×100,000 pixels, which leads to hundreds of thousands of tissue blocks after decomposition (e.g., split the whole mount tissue sections into blocks with fixed size, such as 1000-by-1000 pixels). 2) Extremely large amount of variations in the data, which are due to sample preparation (e.g., fixation, staining) and biological heterogeneities (e.g., cell type, cell state) across histological tissue sections, especially when these tissue sections are processed and scanned at different laboratories.

Therefore, tissue histology classification shares the same threads with pattern recognition and big data in computer vision. Although, the work in [1] provides a good handler to variations in the data based on the sparse cellular morphometric context, the scalability is still an bottleneck due to the sparse optimization during feature extraction. In this paper, we propose the predictive sparse morphometric context, which approximates the sparse cellular morphometric feature through a non-linear regressor that is jointly learned with an over-complete dictionary in an unsupervised manner. And we show that our proposed approach achieves competitive results compared to [1], with significantly improved efficiency.

Organization of this paper is as follows: Section 2 reviews related works. Section 3 describes the proposed approach. Section 4 elaborates the details of our experimental setup, followed by a detailed discussion on the experimental results. Lastly, section 5 concludes the paper.

2. RELATED WORK

Several outstanding reviews for the histology sections analysis can be found in [2, 3]. From our perspective, four distinct works have defined the trends in tissue histology analysis: (i) one group of researchers proposed nuclear segmentation and organization for tumor grading and/or the prediction of tumor recurrence [4, 5, 6, 7, 8]. (ii) A second group of researchers focused on patch level analysis (e.g., small regions) based on either human engineered features [9, 10] or features from unsupervised learning [11, 12, 13], for tumor representation. (iii) A third group focused on block-level analysis to distinguish different states of tissue development using cell-graph representation [14, 15]. (iv) Finally, a fourth group has suggested detection and representation of the auto-immune response as a prognostic tool for cancer [16].

Most recently, sparse morphometric context [1] has been developed for tissue histology classification with great success. It incorporates the sparse cellular morphometric features within the spatial pyramid matching (SPM [17]) framework to capture the cellular morphometric context for invariant representation of the underlying properties in the data. Most importantly, the authors in [1] have demonstrated that system built upon sparse morphometric context is invariant to segmentation strategies, and extensible to different tumor types. However, this approach suffers from the sparse

This work was supported by NIH grant U24 CA1437991 carried out at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231.

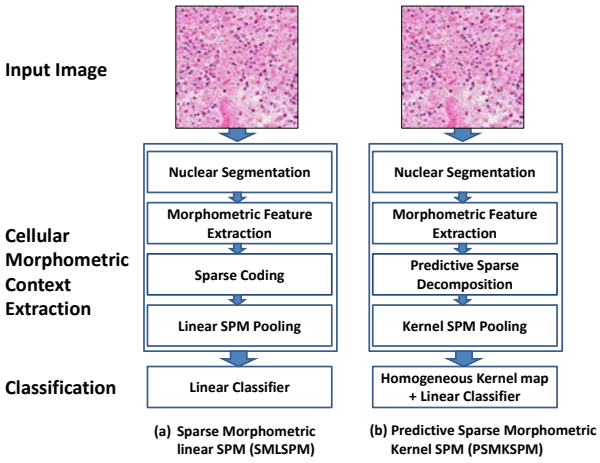

Fig. 1. Schematic comparison of the original sparse morphometric linear SPM (SMLSPM) with our proposed predictive sparse morphometric nonlinear SPM (PSMKSPM).

optimization involved in sparse cellular morphometric feature calculation, which makes it impractical for tasks with hundreds of thousands of tissue sections.

In the context of machine learning research on sparse coding, predictive sparse decomposition (PSD) [18] has been proposed for fast inference of sparse features through a non-linear regressor automatically constructed during dictionary learning. This approach has been demonstrated to be very efficient and effective through its applications to visual object recognition tasks [18].

3. APPROACH - PREDICTIVE SPARSE MORPHOMETRIC KERNEL SPM (PSMKSPM)

The workflow of our approach is shown in Figure 1(b), where PSD is employed for the fast inference of sparse cellular morphometric features. Detailed steps of our approach can be found as follows,

1. Build sparse auto encoder ($\mathbf{W}$). Given $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{m \times N}$ as a set of cellular morphometric descriptors extracted based on nuclear segmentation, we formulate the PSD optimization problem as:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{X},\mathbf{G},\mathbf{W}} \quad & \|\mathbf{Y} - \mathbf{B}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_1 + \|\mathbf{X} - \mathbf{G}\sigma(\mathbf{W}\mathbf{Y})\|_F^2 \\ \text{s.t.} \quad & \|\mathbf{b}_i\|_2^2 = 1, \forall i = 1, \ldots, h \end{aligned} \tag{1}$$

where $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_h] \in \mathbb{R}^{m \times h}$ is a set of the basis functions; $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{h \times N}$ is the sparse feature matrix; $\mathbf{W} \in \mathbb{R}^{h \times m}$ is the auto-encoder; $\sigma(\cdot)$ is the element-wise sigmoid function; $\mathbf{G} = \text{diag}(g_1, \ldots, g_h) \in \mathbb{R}^{h \times h}$ is a scaling matrix with diag being an operator aligning vector, $[g_1, \ldots, g_h]$, along the diagonal; and λ is a regularization constant. Joint minimization of Eqn. (1) w.r.t the quadruple $\langle \mathbf{B}, \mathbf{X}, \mathbf{G}, \mathbf{W} \rangle$, enforces the inference of the nonlinear regressor $\mathbf{G}\sigma(\mathbf{W}\mathbf{Y})$ to be similar to the optimal sparse codes, $\mathbf{X}$, which can reconstruct $\mathbf{Y}$ over $\mathbf{B}$ [18]. As shown in Algorithm 1, optimization of Eq. (1) is iterative, where the it terminates when either the objective function is below a preset threshold or the maximum number of iterations has been reached.

Algorithm 1 Construction of Sparse Auto Encoder: $\mathbf{W}$

Input: Training set $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{m \times N}$
Output: Sparse auto encoder $\mathbf{W} \in \mathbb{R}^{h \times m}$

1: **Initialize:** Randomly initialize $\mathbf{B}$, $\mathbf{W}$, and $\mathbf{G}$
2: **repeat**
3: Fixing $\mathbf{B}$, $\mathbf{W}$ and $\mathbf{G}$, minimize Eq. (1) w.r.t $\mathbf{X}$, where $\mathbf{X}$ can be either solved as a ℓ_1-minimization problem [19] or equivalently solved by greedy algorithms, e.g., Orthogonal Matching Pursuit (OMP) [20].
4: Fixing $\mathbf{B}$, $\mathbf{W}$ and $\mathbf{X}$, solve for $\mathbf{G}$, which is a simple least-square problem with analytic solution.
5: Fixing $\mathbf{X}$ and $\mathbf{G}$, update $\mathbf{B}$ and $\mathbf{W}$, respectively, using the stochastic gradient descent algorithm.
6: **until** Convergence (maximum iterations reached or objective function $\leq$ threshold)

2. Build dictionary ($\mathbf{D}$) of sparse cellular morphometric types, where $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_K]^\top$ are the K sparse cellular morphometric types to be learned by the following optimization:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{Z}} \sum_{m=1}^{M} ||\mathbf{x}_m - \mathbf{z}_m\mathbf{D}||^2 \\ &\text{subject to } card(\mathbf{z}_m) = 1, |\mathbf{z}_m| = 1, \mathbf{z}_m \succeq 0, \forall m \end{aligned} \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_M]^\top$ is a set of sparse codes generated through the nonlinear regressor ($\mathbf{X} = \mathbf{G}\sigma(\mathbf{W}\mathbf{Y})$); $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_M]^\top$ indicates the assignment of the feature type, $card(\mathbf{z}_m)$ is a cardinality constraint enforcing only one nonzero element of $\mathbf{z}_m$, $\mathbf{z}_m \succeq 0$ is a non-negative constraint on the elements of $\mathbf{z}_m$, and $|\mathbf{z}_m|$ is the ℓ_1-norm of $\mathbf{z}_m$. During training, Equation 2 is optimized w.r.t both $\mathbf{Z}$ and $\mathbf{D}$; In the coding phase, for a new set of $\mathbf{X}$, the learned $\mathbf{D}$ is applied, and Equation 2 is optimized w.r.t $\mathbf{Z}$ only.

3. Build spatial pyramid representation [17] for the characterization of sparse morphometric context, which is a concatenation of histograms of sparse cellular morphometric types at different scales and locations.

4. Build multi-class linear SVM for classification [21]. In this step, a homogenous kernel map was applied [22] to approximate the χ^2 kernel prior to the construction of linear classifier due to fact that χ^2 kernel is one of the most suitable kernels for histogram representations as suggested in [23].

4. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental comparison with the previous state-of-art (SML-SPM) was carried out on both (i) Glioblastoma Multiforme (GBM) and (ii) Kidney Renal Clear Cell Carcinoma (KIRC) datasets, which contain images (mostly 1000×1000 pixels) curated from TCGA [1] whole slide tissue sections, scanned with 20X objective (0.502 micron/pixel) and 40X objective (0.252 micron/pixel), respectively. Examples of GBM and KIRC datasets are shown in Figure 2 and Figure 3, respectively, and more details about these two datasets can be found in [1]. The implementation details of PSMKSPM are listed as follows,

1. Cellular morphometric features (same as listed in Table 2 in [1]) were extracted via MRGC [8] and normalized independently with zero mean and unit variance;

[1] http://cancergenome.nih.gov/

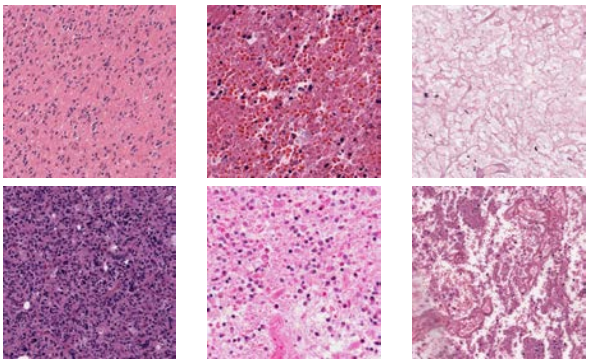

Fig. 2. GBM Examples. First column: Tumor; Second column: Transition to necrosis; Third column: Necrosis.

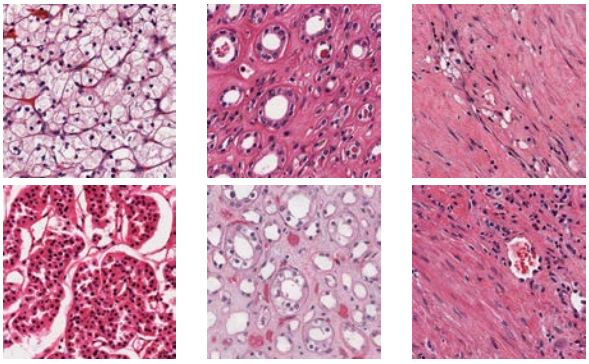

Fig. 3. KIRC examples. First column: Tumor; Second column: Normal; Third column: Stromal.

2. The number of basis functions ($\mathbf{B}$) was fixed to be 128, and the SPAMS optimization toolbox [24] was adopted for efficient implementation of OMP to compute the sparse code, $\mathbf{X}$, with sparsity prior set to 30;
3. The level of pyramid was fixed to be 3;
4. Linear SVM was used for classification.

For fair comparison, we follow the experimental setup as in [1], where all the evaluations were repeated 10 times with randomly selected training/testing images, and the final results were reported as the mean and standard deviation of the classification rates. The detailed comparisons are shown in Table 1 and Table 2, respectively, where the results on SMLSPM were directly cited from [1].

4.1. Discussion

1. PSMKSPM is competitive with SMLSPM in terms of performance. As shown in Tables 1 and 2, the performance of the proposed approach (PSMKSPM) is competitive with the performance of SMLSPM on both GBM and KIRC datasets, with various configurations.
2. PSMKSPM is superior to SMLSPM in terms of scalability. During sparse cellular feature calculation, PSMKSPM involves only element-wise nonlinearity and matrix multiplication, as a result, it is much more efficient than SMLSPM, which needs further optimization. Table 3 gives an intuitive example, showing the significant speed-up achieved by PSMKSPM.

Dataset	**GBM**	**KIRC**
Average #cells/image	606	302
SMLSPM (sec/image)	2.6	2.0
PSMKSPM (sec/image)	0.05	0.03
Speed-Up	52X	66X

Table 3. Comparison of computational time for sparse cellular morphometric feature extraction on both GBM and KIRC datasets, where the dictionary size for both SMLSPM and PSMKSPM were fixed to be 256.

5. CONCLUSIONS

In this paper, we proposed a spatial pyramid matching approach based on predictive sparse cellular morphometric feature, for tissue histology classification. By utilizing non-linear regressor, jointly learned with an over-complete dictionary in an unsupervised manner, for the inference of sparse cellular morphometric features, the proposed approach achieves significant speed up compared to the state-of-art [1], without loss of performance. As a result, it enables the automatic analysis of extremely large histological dataset in an efficient and effective way. Our future work will focus on processing entire tumor banks for the construction of predictive models of clinical outcome.

6. REFERENCES

[1] Chang, H., Borowsky, A., Spellman, P., Parvin, B.: Classification of tumor histology via morphometric context. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. (2013) 2203–2210 1, 2, 3, 4

[2] Demir, C., Yener, B.: Automated cancer diagnosis based on histopathological images: A systematic survey. Technical Report, Rensselaer Polytechnic Institute, Department of Computer Science. (2009) 1

[3] Gurcan, M., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., Bulent, Y.: Histopathological image analysis: a review. IEEE Transactions on Biomedical Engineering **2** (2009) 147–171 1

[4] Axelrod, D., Miller, N., Lickley, H., Qian, J., Christens-Barry, W., Yuan, Y., Fu, Y., Chapman, J.: Effect of quantitative nuclear features on recurrence of ductal carcinoma in situ (DCIS) of breast. Cancer Informatics **4** (2008) 99–109 1

[5] Datar, M., Padfield, D., Cline, H.: Color and texture based segmentation of molecular pathology images using HSOMs. In: ISBI. (2008) 292–295 1

[6] Basavanhally, A., Xu, J., Madabhushu, A., Ganesan, S.: Computer-aided prognosis of ER+ breast cancer histopathology and correlating survival outcome with oncotype DX assay. In: ISBI. (2009) 851–854 1

[7] Doyle, S., Feldman, M., Tomaszewski, J., Shih, N., Madabhushu, A.: Cascaded multi-class pairwise classifier (CASCAMPA) for normal, cancerous, and cancer confounder classes in prostate histology. In: ISBI. (2011) 715–718 1

[8] Chang, H., Han, J., Borowsky, A., Loss, L.A., Gray, J.W., Spellman, P.T., Parvin, B.: Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association. IEEE Trans. Med. Imaging **32**(4) (2013) 670–682 1, 2, 4

	Method	DictionarySize=256	DictionarySize=512	DictionarySize=1024
160 training	PSMKSPM	92.28 ± 0.81	92.18 ± 0.97	**92.35 ± 0.80**
	SMLSPM [1]	92.35 ± 0.83	92.57 ± 0.91	**92.91 ± 0.84**
80 training	PSMKSPM	90.89 ± 0.89	**90.93 ± 0.57**	90.62 ± 0.49
	SMLSPM [1]	90.82 ± 1.28	90.29 ± 0.68	**91.08 ± 0.69**
40 training	PSMKSPM	88.71 ± 0.36	**89.03 ± 0.78**	88.41 ± 1.03
	SMLSPM [1]	88.05 ± 1.38	87.88 ± 1.04	**88.54 ± 1.42**

Table 1. Performance on the GBM dataset, where PSMKSPM and SMLSPM were evaluated based on the segmentation method: MRGC [8].

	Method	DictionarySize=256	DictionarySize=512	DictionarySize=1024
280 training	PSMKSPM	98.26 ± 0.34	98.30 ± 0.35	**98.41 ± 0.39**
	SMLSPM [1]	98.15 ± 0.46	**98.50 ± 0.42**	98.21 ± 0.44
140 training	PSMKSPM	97.73 ± 0.35	**97.77 ± 0.38**	97.73 ± 0.38
	SMLSPM [1]	97.40 ± 0.50	**97.98 ± 0.35**	97.35 ± 0.48
70 training	PSMKSPM	96.58 ± 1.06	**96.66 ± 0.88**	96.43 ± 0.90
	SMLSPM [1]	96.20 ± 0.85	**96.37 ± 0.85**	96.19 ± 0.62

Table 2. Performance on the KIRC dataset, where PSMKSPM and SMLSPM were evaluated based on the segmentation method: MRGC [8].

[9] Bhagavatula, R., McCann, M.T., Fickus, M., Castro, C.A., Ozolek, J.A., Kovacevic", J.: A vocabulary for the identification and delineation of teratoma tissue components in hematoxylin and eosin-stained samples. J Pathol Inform. **5**(19) (2014) 1

[10] Kong, J., Cooper, L., Sharma, A., Kurk, T., Brat, D., Saltz, J.: Texture based image recognition in microscopy images of diffuse gliomas with multi-class gentle boosting mechanism. In: ICASSAP. (2010) 457–460 1

[11] Chang, H., Nayak, N., Spellman, P., Parvin, B.: Characterization of tissue histopathology via predictive sparse decomposition and spatial pyramid matching. Medical image computing and computed-assisted intervention–MICCAI (2013) 1

[12] Chang, H., Zhou, Y., Borowsky, A., Barner, K., Spellman, P., Parvin, B.: Stacked predictive sparse decomposition for classification of histology sections. International Journal of Computer Vision (2014) 1–16 1

[13] Zhou, Y., Chang, H., Barner, K., Spellman, P., Parvin, B.: Classification of histology sections via multispectral convolutional sparse coding. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. (June 2014) 3081–3088 1

[14] Acar, E., Plopper, G., Yener, B.: Coupled analysis of in vitro and histology samples to quantify structure-function relationships. PLoS One **7**(3) (2012) e32227 1

[15] Bilgin, C., Ray, S., Baydil, B., Daley, W., Larsen, M., Yener, B.: Multiscale feature analysis of salivary gland branching morphogenesis. PLoS One **7**(3) (2012) e32906 1

[16] Fatakdawala, H., Xu, J., Basavanhally, A., Bhanot, G., Ganesan, S., Feldman, F., Tomaszewski, J., Madabhushi, A.: Expectation-maximization-driven geodesic active contours with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on breast cancer histopathology. IEEE Transactions on Biomedical Engineering **57**(7) (2010) 1676–1690 1

[17] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. (2006) 2169–2178 1, 2

[18] Kavukcuoglu, K., Ranzato, M., LeCun, Y.: Fast inference in sparse coding algorithms with applications to object recognition. Technical Report CBLL-TR-2008-12-01, Computational and Biological Learning Lab, Courant Institute, NYU (2008) 2

[19] Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: In NIPS, NIPS (2007) 801–808 2

[20] Tropp, J., Gilbert, A.: Signal recovery from random measurements via orthogonal matching pursuit. Information Theory, IEEE Transactions on **53**(12) (2007) 4655–4666 2

[21] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874 2

[22] Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(3) (2012) 480–492 2

[23] Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. (2009) 1794–1801 2

[24] Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. **11** (March 2010) 19–60 3