34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Scale Normalization of Histopathological Images for Batch Invariant Cancer Diagnostic Models

Sonal Kothari, *Student Member, IEEE*, John H. Phan, Ph.D. *Member, IEEE*, and May D. Wang*, Ph.D. *Senior* Member, *IEEE*

***Abstract*—Histopathological images acquired from different experimental set-ups often suffer from batch-effects due to color variations and scale variations. In this paper, we develop a novel scale normalization model for histopathological images based on nuclear area distributions. Results indicate that the normalization model closely fits empirical values for two renal tumor datasets. We study the effect of scale normalization on classification of renal tumor images. Scale normalization improves classification performance in most cases. However, performance decreases in a few cases. In order to understand this, we propose two methods to filter extracted image features that are sensitive to image scaling and features that are uncorrelated with scaling factor. Feature filtering improves the classification performance of cases that were initially negatively affected by scale normalization.**

I. Introduction

Histopathological analysis of biopsy specimens is essential for diagnosing and characterizing cancer. Computer-aided cancer diagnostic tools aid pathologists in making objective and timely decisions [1]. Feature extraction and data mining are the key components of diagnostic systems. For a system with a single data source, we can expect sample images to have the same spatial resolution, magnification and stain colors. Hence, a predictive model can be developed by mining image features from a training set without normalization. However, if the images are acquired from separate set-ups, images suffer from both color and scale variations. Color variations affect color-based image features and color-based segmentation, hence affecting the diagnostic performance of the model. Previous work suggests methods for color normalization and studies its effect on color segmentation accuracy with different batch images [2]. Scale variations can be caused by the variation in numerical aperture, magnification, or digitizing device of the microscope. Scale variations affect various image features such as object size, topology and texture. To the best of our knowledge, no work has been done to address this issue in histopathological images.

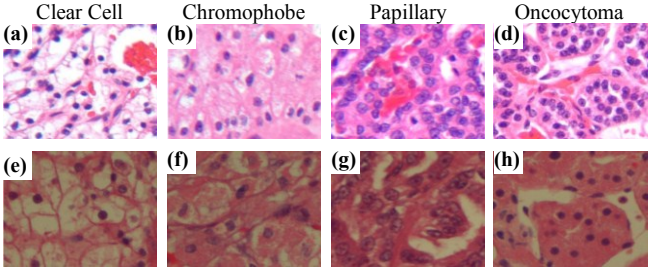

Figure 1. Sample histological tissue images (270x330 pixels) for four renal tumor subtypes from RCC1 (a-d) and RCC2 (e-h).

For natural images (e.g., photographs of buildings and landscapes), scale variations have been widely studied and researchers have suggested several scale invariant features for representing images [3]. However, in histopathological images, the size of cellular structures, such as nuclei and glands, has been reported to be very important [4]. A similar challenge is faced in organ imaging, especially when dealing with brain MRI, where the size of the brain and its structures varies among samples. Researchers have suggested methods for spatial normalization and registration of these images based on the size or volume of anatomical structures [5].

We develop a model to normalize histopathological image scale using nuclear area (measured in pixels). After scale normalization, we represent each image using a comprehensive image-feature list [6]. We then develop binary diagnostic models to classify 4 renal tumor subtypes from two image batches. Our results indicate that scale normalization improves classification performance in most cases but has no effect or negative effects for a few cases. Filtering the features based on scaling-induced variation improves prediction performance in some cases that were previously negatively affected by scale normalization.

II. Methods

A. Datasets

We perform this study on RGB images of hematoxylin and eosin (H&E) stained renal tumor tissue samples. We use two separately acquired datasets—RCC1 and RCC2 (**Figure 1**). These datasets include four prominent histological subtypes of renal tumors—chromophobe (CH), clear cell (CC), papillary (PA), and oncocytoma (ON). RCC1 contains 48 images with 12 images of each subtype. RCC2 contains 58 images with 20, 17, 16, and 5 images of CH, CC, PA, and

This research is supported by grants from NIH (P20GM072069, and CCNE U54CA119338), Georgia Cancer Coalition ((Distinguished Cancer Scholar Award to Professor MDW), Hewlett Packard, and Microsoft Research.

Sonal Kothari is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, USA (e-mail: sk9@gatech.edu).

John H. Phan is with the Department of Biomedical Engineering, Georgia Institute of Technology, USA (e-mail: jhphan@gatech.edu).

* May D. Wang is the corresponding author with the joint Department of Biomedical Engineering at Georgia Institute of Technology and Emory University, USA (phone: 404-385-5059; e-mail: maywang@bme.gatech.edu).

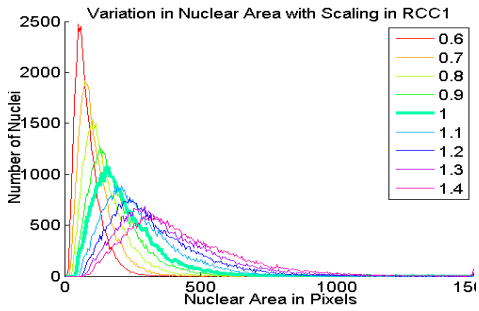

Figure 2. Nuclear area distribution at various scales for RCC1. The cyan curve corresponds to the original scale (s=1) of the dataset. Changes in image scale shift the distribution.

ON subtypes, respectively. Each sample image is about 1200x1600 pixels.

B. Scale Normalization

At a reasonable microscopic magnification, cellular nuclei are easily identified in histological tissue images. We use a concavity-based cluster segmentation method for segmenting nuclear clusters in histopathological images to extract individual nuclei [7]. In a single image, nuclear area may vary with cancer grade and cancer subtype [4]. However, if we study the distribution of all nuclei in a dataset, the distribution peaks at a specific nuclear area for the dataset. As expected, image scaling shifts the distribution of nuclear area. **Figure 2** illustrates the distribution of nuclear area as a function of scale for the RCC1 dataset (results for the RCC2 dataset are similar). The scaling factor, s, is relative to the original image size. We use the Lanczos (3-lobe) filter for scaling due to its good balance between anti-aliasing and blurring. The dataset is scaled by various scaling factors, s, ranging from 0.6 to 1.4. We observe a clear trend in the distribution of nuclear area with scaling. We mathematically formulate this relationship as follows. We know s multiplies both the x and y dimension of an image, resulting in an increase of area by s^2 (i.e., $A_s \propto s^2$, where A_s is the image area at scaling factor s). We can estimate the scale of an image dataset using the relationship $s = \sqrt{A_s / A_1}$. We quantify A_s using the median nuclear area of a dataset at scale s. Therefore, while developing the diagnostic model, we calculate A_1 (median nuclear area at original scale) for the training and testing datasets and calculate the scaling factor between them. We then scale down the dataset with larger nuclear area.

C. Feature Extraction

We extract a comprehensive set of image features from each sample [6]. This set comprises of 12 feature subsets extracted from different processed forms of the original sample images, resulting in a total of 2663 features. **Table 1** lists the 12 feature subsets and their combination set (i.e., the ***All*** set). **Figure 3** describes the flow of feature extraction, where cyan boxes represent different forms of the processed image while pink boxes represent feature subsets.

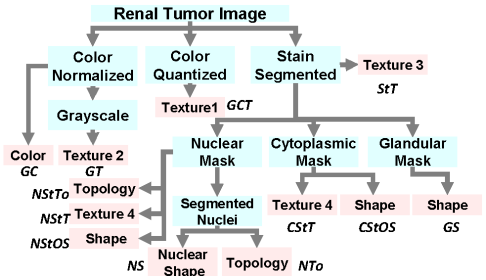

Figure 3. Flow diagram for image feature extraction. Cyan boxes: original or processed image. Pink boxes: feature subset. **Table 1** lists the number of features in each feature subset.

Table 1

Set	Count
GC	48
GT	557
GCT	493
StT	503
CStOS	249
CStT	77
GOS	249
NStOS	249
NStT	77
NStTo	56
NS	49
NTo	56
All	2663

We first describe all cyan boxes. We normalize the colors in the sample image to a standard reference, using a *color map* quantile normalization method to eliminate the color batch effects [2]. We quantize the color space of the image into 64 levels using self-organizing maps [8]. The color quantized image is similar to a grayscale image except the quantization is nonlinear. The four color stains (blue, pink, white, and red) in an H&E stained image correspond to nuclei, red-blood cells, cytoplasmic and glandular structures, respectively. We segment these colors using an automatic color segmentation method [2]. The stain segmented image is a four-level grayscale image where gray-levels of 1, 2, 3 and 4 are assigned to pixels that belong to the four structures. We then extract binary masks for nuclear, cytoplasmic and glandular structures in the image based on segmentation labels. We further segment the nuclear clusters in the nuclear mask into individual nuclei to produce the segmented nuclei [7, 9].

Next, we briefly describe various feature groups in **Figure 3**. Previous work discusses these feature groups and their various parameters in detail [6]. **Color** features include frequencies of R, G, and B histograms. The **Texture1** feature set includes various commonly used texture properties such as Haralick properties from GLCM (Gray Level Co-occurrence Matrix), energy and entropy of Gabor filter responses, energy and entropy of wavelet packet sub-matrices and energy and entropy of multiwavelet decomposition sub-matrices. The **Texture2** feature set includes all features in **Texture1** as well as gray-level distribution. **Texture3** includes all features in **Texture1** as well as stain co-occurrence, capturing co-occurrence of stains similar to GLCM. **Texture4** includes Haralick properties and gray-level distribution. All of the above features capture global features.

In addition to global features, we capture object level features. The **Shape** feature set includes various shape properties of image objects including pixel area, convex hull area, solidity, perimeter, elliptical properties (area, major-minor axes lengths, eccentricity and orientation), boundary fractal, bending energy, Fourier shape descriptor reconstruction error, and object count. The **Topology** feature set captures architectural properties of images including Delaunay triangles (area and side lengths), Voronoi diagrams (area, side lengths and perimeters),

minimum spanning tree edge length and closeness. The **Nuclear Shape** feature set captures nuclear elliptical shape properties, nuclear cluster size (in number of nuclei) and nuclear count.

C. Feature Filtering

Some image features vary with image scale while others do not. Moreover, some features vary in an unpredictable manner with image scale. Consider the scatter plots in **Figure 4.** These plots capture feature variance across multiple image scales as well as feature correlation with scale. The Y-axis of this plot measures the normalized standard deviation, σ_m, of the feature m when a dataset is scaled by various scaling factors, s_j (j=1,..,S). The normalized standard deviation is an average over all samples, N, in the dataset and is normalized by the average magnitude of the feature at original scale. This normalization scales the standard deviation of all features such that they are comparable regardless of feature magnitude. σ_m is represented by the following equation:

$$\sigma_m = \sum_{i=1}^{N} \sqrt{\frac{1}{S-1}\sum_{j=1}^{S}\left(x_{m,i}^{j} - \bar{x}_{m,i}\right)^2} \Bigg/ \sum_{i=1}^{N}\left|x_{m,i}^{1}\right|, \text{ where } \bar{x}_{m,i} = \frac{1}{S}\sum_{j=1}^{S} x_{m,i}^{j}$$

$x_{m,i}^{j}$ represents the value of feature, m, of image, i, at scale, s_j. The X-axis of the plot represents the correlation r_m of feature value, $x_{m,i}^{j}$ with scale, s_j, considering all samples and scales, given by

$$r_m = \frac{\sum_{i=1}^{N}\sum_{j=1}^{S}\left(x_{m,i}^{j} - \bar{x}_m\right)\left(s_j - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{S}\left(x_{m,i}^{j} - \bar{x}_m\right)^2 \sum_{i=1}^{N}\sum_{j=1}^{S}\left(s_j - \bar{y}\right)^2}},$$

where $\bar{x}_m = \frac{1}{N}\sum_{i=1}^{N}\bar{x}_{m,i}$ and $\bar{y} = \frac{1}{S}\sum_{j=1}^{S} s_j$.

We scale the datasets using the scaling factors s_j=0.6:0.1:1.4, and propose two filtering methods based on feature variation and correlation with scale: 1) scale variant feature filtering, and 2) uncorrelated scale variant feature filtering. Feature filtering is done based on the training set, before model development. In scale variant feature filtering, we remove all features with high standard deviation, σ_m> 0.1. After this filtering we have 1297 and 1313 features for RCC1 and RCC2 respectively. In uncorrelated scale variant feature filtering, we filter features with high variation (σ_m> 0.1) and low correlation ($|r_m|$ < 0.1). This filtering method filters the high variant features only if they have variation uncorrelated to scaling. After this filtering we have 2289 and 2274 features for RCC1 and RCC2 respectively.

C. Feature Selection and Classification

We consider binary endpoints comparing all pairs of classes. With 4 renal tumor subtypes, we have 6 binary endpoints per dataset, resulting in a total of 12 endpoints. In this study, we have a limited number of samples and a large number of features. As such, we use 10 iterations of 5-fold nested cross-validation (CV) to obtain an robust estimate of prediction performance [10]. We use a Bayesian classifier with the following parameters—pooled and un-pooled variance with spherical and diagonal variance matrices, resulting in four Bayesian models. We use two forms of mRMR (minimum redundancy and maximum relevance) as our feature selection methods: mRMR-d (difference) and mRMR-q (quotient) [11]. We consider 30 feature sizes ranging from 1 to 30. Thus, for each endpoint, we develop 240 (30*4*2) models. We select the optimal prediction model based on the performance estimated from internal CV. We select the simplest model within one standard deviation from the best performing model to avoid over fitting. We define the simplest model as one that prefers small feature size, pooled variance over un-pooled variance, and spherical covariance over diagonal covariance. We use average external CV classification accuracy as our performance metric.

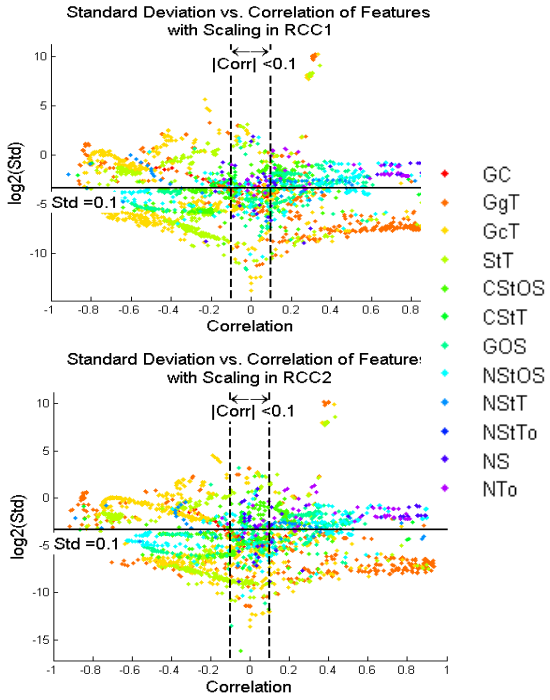

Figure 4. The relationship between feature variation and feature correlation with image scale. Feature variation is quantified as normalized standard deviation. Features above the solid horizontal line have high variance with respect to image scale. Features between the dashed lines are uncorrelated with image scale.

III. RESULTS AND DISCUSSION

A. Validation of Normalization Model

We validate our scale normalization model using empirical values. **Figure 5** compares the model to empirical values of nuclear pixel area. The cyan dashed curve and green dotted curve represent the models for RCC1 and RCC2, respectively, generated using the model $s = \sqrt{A_s / A_1}$, where A_1 is known for both datasets. The red circles and red squares represent the empirical values for nuclear area extracted from RCC1 and RCC2, respectively. We scale the datasets with the scaling factors s=0.5:0.1:2.0 and calculate their median nuclear area to obtain the empirical values.

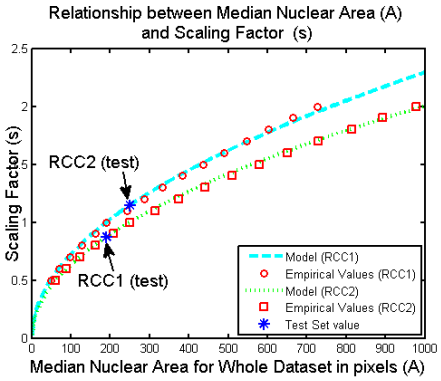

Figure 5. Comparison between scale normalization model and empirical values of nuclear area for the RCC1 and RCC2 datasets.

Empirical values for nuclear area closely support our model. The blue asterisks mark the scaling point for each dataset in relation to the other dataset. The scale of RCC2 is 1.15 times the scale of RCC1. Likewise, RCC1 is 0.87 times the scale of RCC2. Therefore, to assess classification performance, we scale the RCC2 dataset by a factor of 0.87.

B. Comparison of Diagnostic Performance

We illustrate the effect of scale normalization on classification performance by developing diagnostic models for 12 endpoints using four methods. First, we develop models with no normalization and use this as our baseline (**Table 2**, **M1**). Second, we normalize datasets using scale normalization and develop diagnostic models using all features (**M2**). Third, we normalize datasets, filter the features using scale variant feature filtering and develop diagnostic models with the filtered feature set (**M3**). Fourth, we normalize datasets and develop diagnostic models using features after filtering uncorrelated scale variant features (**M4**). **Table 2** compares the results for **M2**, **M3** and **M4** to the baseline, **M1**. We have highlighted the performance in green or pink if there is an increase or decrease compared to M1. It can be seen that **M2** improves or has no effect on performance in all but three cases: **CC vs. PA** with RCC1 as training, and **CH vs. ON** and **CC vs. PA** with RCC2 as training. This decrease is possibly due to selection of features that have high and unpredictable variation with scale. It can be observed that out of these three endpoints, the performance of the first two—**CC vs. PA** with RCC1 as training and **CH vs. ON** with RCC2 as training—can be improved by both filtering methods. However, filtering methods may remove some useful features for other diagnostic endpoints. Overall **M4**, with uncorrelated scale invariant feature filtering, performs better compared to **M3** with scale invariant feature filtering.

IV. Conclusion

We have described a novel scale normalization method for histological images based on nuclear area. We verified the scale normalization model using empirical values and illustrated the effect of scale normalization on classification performance using 12 renal tumor subtype endpoints. Scale normalization improves the predictive performance of 6 endpoints and decreases the performance of 3 endpoints. We proposed two feature filtering methods to improve prediction performance and observed that feature filtering improves the performance of two out of three endpoints. We have considered datasets with only slight scale variations (~15%) and still observed improvements in performance. We hypothesize that scale normalization could greatly improve histological image classification performance in cases with larger scale batch effects.

Table 2: Diagnostic performance of models with *All* features.

Endpoints		M1	M2	M3	M4
Train=RCC1, Test=RCC2	CH vs. CC	0.50	0.58	0.67	0.60
	CH vs. ON	0.54	0.56	0.54	0.42
	CH vs. PA	0.92	0.92	0.92	0.88
	CC vs. ON	0.92	0.92	0.88	0.88
	CC vs. PA	0.75	0.63	0.79	0.92
	ON vs. PA	0.94	0.96	0.92	1.00
Train=RCC2, Test=RCC1	CH vs. CC	0.51	0.51	0.51	0.54
	CH vs. ON	0.60	0.68	0.48	0.36
	CH vs. PA	0.81	0.47	0.92	0.86
	CC vs. ON	0.82	0.95	0.82	0.86
	CC vs. PA	0.79	0.48	0.48	0.48
	ON vs. PA	0.86	0.90	0.62	0.90

Note: Values highlighted in green and pink have increased and decreased performance compared to M1, respectively.

M1: No normalization, **M2**: Scale normalization, **M3**: Scale normalization and filtering of scale variant features, **M4**: Scale normalization and filtering of uncorrelated scale variant features

Acknowledgment

We thank Dr. Todd Stokes for his valuable comments and suggestions.

References

[1] M. N. Gurcan, *et al.*, "Histopathological image analysis: a review," *IEEE Rev Biomed Eng,* vol. 2, pp. 147-171, 2009.

[2] S. Kothari, *et al.*, "Automatic batch-invariant color segmentation of histological cancer images," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 2011, pp. 657-660.

[3] J. Wood, "Invariant pattern recognition: A review* 1," *Pattern Recognition,* vol. 29, pp. 1-17, 1996.

[4] J. Eble, *et al.*, *Pathology and genetics of tumours of the urinary system and male genital organs*: IARC press Lyon, 2004.

[5] J. Lancaster, *et al.*, "Anatomical Global Spatial Normalization," *Neuroinformatics,* vol. 8, pp. 171-182, 2010.

[6] S. Kothari, *et al.*, "Histological Image Feature Mining Reveals Emergent Diagnostic Properties for Renal Cancer," in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, 2011, pp. 422-425.

[7] S. Kothari, *et al.*, "Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques," in *Biomedical Imaging: From Nano to Macro, 2009 IEEE International Symposium on*, 2009, pp. 795-798.

[8] J. Vesanto, *et al.*, "Self-organizing map in Matlab: the SOM toolbox," in *Proceedings of the Matlab DSP Conference*, 1999, pp. 16-17.

[9] S. Kothari, *et al.*, "Extraction of informative cell features by segmentation of densely clustered tissue images," in *Engineering in Medicine and Biology Society, 2009. Annual International Conference of the IEEE*, 2009, pp. 6706-6709.

[10] R. M. Parry, *et al.*, "k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction," *Pharmacogenomics J,* vol. 10, pp. 292-309, 2010.

[11] H. Peng, *et al.*, "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, pp. 1226-1238, 2005.