MULTI-CLASS SINGLE-LABEL CLASSIFICATION OF HISTOPATHOLOGICAL WHOLE-SLIDE IMAGES

Daniel Bug\textsuperscript{1}, Julia Schüler\textsuperscript{3}, Friedrich Feuerhake\textsuperscript{2}, Dorit Merhof\textsuperscript{4}

\textsuperscript{1} Institute of Imaging and Computer Vision, RWTH Aachen University, Germany
\textsuperscript{2} Institute for Pathology, Hannover Medical School, Germany
\textsuperscript{3} Oncotest GmbH, Freiburg, Germany

ABSTRACT

Digitizing histopathological slides into whole-slide images enables the use of image analysis techniques for a comprehensive study of diseases at tissue level. In this work, we investigate possible configurations of classifiers and features to classify entire tissue slides. Feature candidates are first evaluated individually and then combined to form a strong classifier. For evaluation of this nine-class problem we use one sparsely and one densely extracted data set to obtain a conservative and an optimistic estimate of the performance. To this end, the best results in terms of overall accuracy (86.6\%) and F1-score (67.8\%) are achieved by a Random Forest classifier with a Color-Histogram feature.

1. INTRODUCTION

In digital pathology, slides of stained histopathological tissue are scanned using whole-slide imaging systems, resulting in a digital multi-scale representation of the slide. Computer-aided analysis of these whole-slide images (WSI) allows to calculate objective measures facilitating diagnosis and medical research. Recent work on immunotherapies evaluates the influence of certain characteristics in the tumor microenvironment (TME) as an indicator for tumor progression and therapeutic success [1, 2, 3]. A first step towards an automated analysis of the TME is the exhaustive classification of the composing tissue. An initial approach of tissue classification [4] analyzes the tissue composition by solving a segmentation problem with three classes using morphology and texture features. However, information contained in the stain is not considered. More recent proposals to classify histopathological images can be found in [5, 6], where Bag-of-(Visual-)Features is applied for categorization. These algorithms extract visual patterns at key points of an image and bin them into a histogram representation. An SVM is used to map the histograms to their respective class. Furthermore, [7] introduces a multi-class solution for automated Gleason-grading of prostate cancer. A cascaded classification pipeline, based on graph-features of the nuclei positions, labels six classes which denote the different tumor grades and three additional tissue types. Finally, a review on texture-based system is given in [8] showing cases of successful applications of color and texture representations in Gleason-grading of prostate cancer. The paper covers various subcategories of features and reviews methods very similar to this application. Image categorization and prostate cancer grading are use-cases for a region-of-interest classification approach and work with relatively large selected image patches. However, the required region size for a comprehensive description in their respective representation imposes limitations on the use of these approaches in a spatially dense classification scenario.

In this work, we demonstrate how the multi-class problem of comprehensive tissue classification can be addressed using classical machine-learning techniques and features, including an evaluation of the performance. Comprehensive herein refers to a grid-wise classification of potentially overlapping patches that is as dense as possible and as sparse as necessary in terms of computational overhead. Our work targets medical research and assumes a scenario in which the staining protocol can be controlled.

2. MATERIAL

Our database consists of whole-slide images (WSI) showing 20× magnified hematoxylin & eosin stained human lung-tumor xenografts grown in immunocompromized mice. A biomedical expert labeled regions in each WSI in terms of nine distinct classes: tumor (44\%), mouse-stroma (14\%), necrosis (14\%), muscle (2\%), vacuole (2.2\%), blood-vessel (0.2\%), connective tissue (2\%), and the two non-tissue classes technical artifact (1.6\%) and background (20\%), the latter including lumen from larger blood-vessels, mechanical disturbances of the tissue and other "white spaces". As denoted in percent, the class frequencies are quite imbalanced. From this image data, approximately 160,000 patches are extracted in a 32 × 32px sampling grid at the following patch sizes: 32 × 32px, 64 × 64px, and 80 × 80px. The resulting larger patches partly overlap and hence some of the information correlates. Therefore, we consider estimates from this data...
set (SET-1) as optimistic. For the extraction of a second set (SET-2), the sampling grid is chosen as 80 × 80px which yields approximately 30,000 non-overlapping patches with the same patch sizes. Due to the low number of training samples these estimates can be considered as too conservative. In both sets, a clear separation between training and validation data is given.

3. METHOD

A central idea of this work is to base the classification on both color and texture features. Color features are extracted in the RGB, HSV and LAB space in the form of per-patch histograms. All three channels are used, but for reasons of efficiency each channel is linearly reduced to 85 bins resulting in a feature vector of length 255. As texture features we extract: Local-Binary-Patterns (LBP) [9], Histogram of Oriented Gradients (HOG) [10], Daisy (DSY) [11], Hu-Moments (HUM), Local-Phase-Quantization (LPQ) [12] and adaptations of the Greylevel Co-Occurance Matrix (GLCM) [13] as well as a novel fractal dimensions estimates (FDE) feature based on [14]. In the following, GLCM and FDE are explained:

1. The GLCM is a statistic of neighboring gray values in a texture and usually used to compute meta features. All classifiers used in this work are capable of abstracting statistical measures during the learning process, which is why the costly computation of the statistical meta features is omitted. Instead, we blur and subsample the matrix to use a distribution estimate as feature. In detail, the GLCM is computed for four directions and two distances, which yields 256x256x4x2 values. We reduce the dimension of direction by a maximum projection to robustify the feature against rotations. The distances are remapped into an image space of 512x256, blurred and subsampled, resulting in 32x16 (or vectorized 512) features. The subsampling factor of 16 is generally an arbitrary choice providing a trade-off between precision of the GLCM estimate and length of the feature vector, the same holds for the number of pixel distances at which to compute the feature.

2. As a new texture feature, we introduce a vector of fractal dimension estimates (FDE). Many fractals, e.g. the Koch curve or the Sierpinski triangle, are composed by an iterator replacing a structure with $N$ versions of itself, down-scaled by factor $s$. For such fractals, a dimension measure can be defined as \[ f_D = \frac{\log(N)}{\log(s^{-1})}. \] (1)

The measure $f_D$ can be estimated using the box-counting algorithm, where the box size is a representation of scale and the count measures the number of repetitions in a shape. In the classical feature, a texture is coded as the slope of a line-fit on the $f_D$ values computed from multiple box sizes, i.e. a single scalar number. We found that considerable information is contained in the deviations from this fit, which motivates the use of the full vector of $f_D$ estimates as feature, computed from multiple shapes. These are generated by binarizing the grayscale image at different intervals, which are taken from overlapping percentiles of the (grayscale) foreground in a WSI. Using percentiles, we intend to provide robustness against varying lighting conditions in different slides. For each interval, all complete overlaps of boxes and shapes are counted utilizing 2D convolutions between a binary image and a uniform box-filter followed by thresholding and summation. The box sizes follow the Fibonacci series 1, 2, 3, 5, 8, ... up to the first number larger than one-third of the patch size. The scale $s$ of a box is computed as the ratio of box size over patch size. Finally, applying Equation 1 yields a vector with all combinations of binarizations and sizes, which characterizes the texture by its $f_D$ estimates (FDE).

For our purpose, nine largely overlapping binarization intervals turned out to be a good choice. The non-uniform variation of the box size was mainly introduced for speed-up,
4. EXPERIMENTS

Each feature was evaluated standalone with a 4-layer Multi-layer-Perceptron (MLP), an AdaBoost with linear base-classifier (ADA) and a Random-Forest classifier (RFO) on the patch sets described in Sec. 2. Additionally, a Support-Vector Machine (SVM) classifier was included in the evaluation for SET-2 (since the computational complexity of SVMs during training exceeds \(O(n^2)\) in the number of samples, they had to be omitted in the evaluations on SET-1). In all cases, anti-proportional class weights were used, to balance weak accuracy, ADA did not leave out as many classes as MLP and RFO learned 400 base-classifiers each. The results of a 5-fold cross-validation for SET-1 are shown in the upper half of Table 1. Finally, the top-performing features in the color-space and texture categories were evaluated in combination using 10-fold cross-validation as shown in the lower half of Table 1. The number of base-classifiers was raised to 800 to compensate for the larger feature space. Equivalently, Table 2 shows the results for the reduced non-overlapping SET-2. A confusion matrix for the winning setup is visualized in Figure 2 and statistical measures are given in Table 3. In the middle section of Figure 1, a classification result for a WSI is shown.

Table 1. Performance for classifier and feature combinations at different patch sizes in SET-1. ACC is the overall accuracy and F1 the average F1-score of the 9 classes. Note that for each feature only the best performing classifier is listed. *Negative numbers denote how many classes were not found for each feature only the best performing classifier is listed. Bold numbers highlight promising configurations.

<table>
<thead>
<tr>
<th>SET-1</th>
<th>32x32</th>
<th>64x64</th>
<th>80x80</th>
</tr>
</thead>
<tbody>
<tr>
<td>160,000 patches</td>
<td>ACC% F1%</td>
<td>ACC% F1%</td>
<td>ACC% F1%</td>
</tr>
<tr>
<td>RFO+RGB</td>
<td>79.3</td>
<td>60.1</td>
<td>85.0</td>
</tr>
<tr>
<td>RFO+HSV</td>
<td>81.4</td>
<td>63.5</td>
<td>85.8</td>
</tr>
<tr>
<td>RFO+LAB</td>
<td>81.3</td>
<td>63.1</td>
<td>86.0</td>
</tr>
<tr>
<td>RFO+LBP</td>
<td>68.5</td>
<td>-2*</td>
<td>72.0</td>
</tr>
<tr>
<td>MLP+HOG</td>
<td>65.6</td>
<td>-6*</td>
<td>71.1</td>
</tr>
<tr>
<td>MLP+DSY</td>
<td>65.8</td>
<td>-5*</td>
<td>75.6</td>
</tr>
<tr>
<td>RFO+HUM</td>
<td>70.4</td>
<td>-1*</td>
<td>73.7</td>
</tr>
<tr>
<td>RFO+LPQ</td>
<td>69.2</td>
<td>-2*</td>
<td>74.6</td>
</tr>
<tr>
<td>RFO+FDE</td>
<td>76.1</td>
<td>45.3</td>
<td>81.4</td>
</tr>
<tr>
<td>RFO+GLCM</td>
<td>77.9</td>
<td>49.8</td>
<td>84.1</td>
</tr>
<tr>
<td>RFO+HSV+FDE</td>
<td>–</td>
<td>–</td>
<td>81.8</td>
</tr>
<tr>
<td>RFO+LAB+GLCM</td>
<td>–</td>
<td>–</td>
<td>84.6</td>
</tr>
<tr>
<td>RFO+LAB+GLCM+FDE</td>
<td>–</td>
<td>–</td>
<td>81.8</td>
</tr>
</tbody>
</table>

Table 2. Performance for classifier and feature combinations at different patch sizes in SET-2. ACC is the overall accuracy and F1 the average F1-score of the 9 classes. Note that for each feature only the best performing classifier is listed. *see Table 1.

<table>
<thead>
<tr>
<th>Classes:</th>
<th>TU</th>
<th>MST</th>
<th>NEC</th>
<th>MUS</th>
<th>VAC</th>
<th>BV</th>
<th>CT</th>
<th>TA</th>
<th>BG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>85.9</td>
<td>85.6</td>
<td>75.5</td>
<td>97.6</td>
<td>59.3</td>
<td>52.3</td>
<td>60.6</td>
<td>84.3</td>
<td>99.7</td>
</tr>
<tr>
<td>Recall</td>
<td>95.6</td>
<td>72.3</td>
<td>67.2</td>
<td>92.5</td>
<td>43.9</td>
<td>08.7</td>
<td>52.3</td>
<td>39.5</td>
<td>99.8</td>
</tr>
<tr>
<td>F1 Score</td>
<td>90.5</td>
<td>78.4</td>
<td>71.1</td>
<td>95.0</td>
<td>50.5</td>
<td>14.8</td>
<td>56.1</td>
<td>53.8</td>
<td>99.8</td>
</tr>
</tbody>
</table>

Table 3. Per class measures for the strongest classifier in the experiment. The classes are: tumor (TU), mouse-stroma (MST), necrosis (NEC), muscle (MUS), vacuole (VAC), blood-vessel (BV), connective tissue (CT), technical artifact (TA) and background (BG).

5. DISCUSSION

5.1. Classifier characteristics

The performance, measured in terms of overall accuracy and F1-score, was nearly identical for MLP and RFO, usually with a difference below 2% and slight advantages for the RFO. Furthermore, the trends for the evaluated features were the same for each classifier, meaning that a strong feature with RFO would be equally strong with ADA compared to other features. However, ADA performed approximately 10% worse than MLP and RFO throughout the experiments, which explains its lack of appearance in Table 1 and 2. Despite the weak accuracy, ADA did not leave out as many classes as MLP and RFO thanks to its learning strategy of re-weighting previously misclassified samples. Due to the computational complexity of Support-Vector Machines, an evaluation was only feasible for the smaller data set SET-2. In case of color-space features, the SVM performance is about 3% below the performance of RFO and MLP, but similar to ADA, the optimization process manages to predict all tissue classes. For texture features, the SVM performance drops as low as 31.9% in the case of HOG, while the best performing texture feature remains GLCM with an accuracy of 71.9%. 

5.2. Feature importance

The performance of the best classifier was strongly influenced by the feature set. As expected, color features performed best, followed by texture features. The combination of color and texture features resulted in the highest accuracy, while combinations of color features yielded the best F1-score. However, the performance of the classifiers was highly variable, with some features performing well in certain classifiers but poorly in others. The best performing features included GLCM, HOG, and RFO, while others such as HUM and DSY performed poorly in all classifiers. The choice of classifier and feature set had a significant impact on the performance of the classifier, with MLP and ADA performing better than RFO in terms of overall accuracy and F1-score, respectively.

5.3. Conclusion

In conclusion, the experiments showed that the choice of classifier and feature set had a significant impact on the performance of the classifier. The best performing features included GLCM, HOG, and RFO, while others such as HUM and DSY performed poorly in all classifiers. The choice of classifier and feature set had a significant impact on the performance of the classifier, with MLP and ADA performing better than RFO in terms of overall accuracy and F1-score, respectively. The results demonstrate the importance of selecting appropriate classifiers and features for the classification of breast tissue in WSI.
Fig. 2. Visualization of the confusion matrix for the RFO classifier with LAB feature for $80 \times 80$px patches. Note that a log scaling is used to keep small values visible.

5.2. Data sets and patch sizes

The main differences between SET-1 and SET-2 are the number of samples and the presence or absence overlap. Both evaluations consistently show increasing classification performance with increasing patch size. While there is a notable improvement from $32 \times 32$px to $64 \times 64$px, the next step to $80 \times 80$px seems relatively small. Such a performance saturation can actually be expected from histogram-based features, since larger patches are more likely composed of mixed tissues and therefore distinctive histogram properties used for classification are concealed. Another observation is the consistent performance of the features in both evaluations relative to each other. However, the higher number of training samples in SET-1 decreased the number of unidentified classes for many texture features, compared to SET-2. In contrast to SET-2, where the combination of color and texture features results in slight improvements, the accuracy and F1-score tend towards the lower rates of the respective texture feature in the combination in case of SET-1. Moreover, regarding features and patch sizes, we would like to point out that HOG and DSY which are densely computed in image space will result in at least quadratically growing feature spaces, implying long computing times in both training and application. In contrast, FDE grows more slowly along with the patch size which makes it comparatively efficient to compute. Yet, it has the benefit of adapting the length of the feature vector to the patch size.

5.3. Features and Classes

A central goal of this work was to design an image description that would distinguish the various classes with a combination of texture and color features. The results indicate that based on the evaluated set of features the major discriminant characteristics are learned from the color space. In this subset, the converted color spaces, HSV in particular, perform better than the native RGB space. The LAB space represents the data close to human vision, while HSV has one channel specifically encoding the hue invariant to illumination changes, which both proved to be good representations for stained images of tissue. Since we assume a medical research scenario with control over the staining protocol, the reduced variance is another plausible explanation for the success of color features and might impose limitations regarding the transfer to clinical (diagnostic) application. Surprisingly, most of the established texture features are incapable of representing the full set of classes in the tissue data. LBP and HUM are both unsuited to detect blood-vessels (BV) (by finding blood-cells or according patterns in the image) and even with the much better performing GLCM and FDE features, BV remain the most challenging class. One reason might be the very low class frequency (0.2%), another reason the scales at which BV appear, reaching from a fraction of the patch (only indicated by a few blood cells) to multiple patches (large vessel). Even with GLCM and FDE, the recall of BV clearly remains the lowest of all classes. Vacuoles are mostly confused with necrosis, because of similarities in areas where the necrosis has destroyed the cells entirely, leaving nothing but plasma. Without the context of the surrounding tissue, the patches look very much alike in these regions, even for experts. Another occasionally occurring but less important confusion happens between mouse-stroma and connective tissue, two classes with similar visual appearance and biological function. All other classes achieve satisfying results, usually with higher precision than recall.

In this work, an increase in context was introduced exclusively through an increase in patch size, implying an upper limit to the use of histogram-based features in terms of discriminative capacity and computing time. Alternatively, the multi-scale representation could be utilized to compute features at a compact patch size, but for multiple levels of resolution.

6. CONCLUSION

This work evaluates multiple configurations of classifiers and suitable features in order to solve the multi-class problem of comprehensive tissue classification. The evaluation on two data sets indicates an advantage of color features over texture features and a preference for converted color-spaces. While for a small data set, a combination of color- and texture-features improved the classification, this tendency was not seen for a second, more densely extracted data set. A combination of a Random Forest Classifier and LAB-space Histogram feature achieved the best results in terms of overall accuracy (86.6%).
7. REFERENCES


