Combining Low-level Features for Improved Classification and Retrieval of Histology Images

Juan C. Caicedo\textsuperscript{1,2} and Ebroul Izquierdo\textsuperscript{2}

\textsuperscript{1} Universidad Nacional de Colombia, Bioingenium Research Group
\textsuperscript{2} Queen Mary University of London, Multimedia and Vision Research Group

Abstract. Feature combination for image classification and indexing is an important design aspect in modern image retrieval systems. It is particularly valuable in medical applications and specially in histology applications in which different features are extracted to estimate tissue composition and architecture. This paper presents an experimental evaluation of textural features combination for histology image classification and retrieval, following a late-fusion scheme. The main focus of this evaluation is oriented to feature normalization to guarantee fair conditions for feature comparison and integration. The experimental evaluation was carried out on a collection of histology images to evaluate the feature combination strategy. Experimental results show that it is possible to improve the system performance by appropriately considering the structure and distribution of visual features. Also, it is shown that feature combination may lead to a decreased performance due to fundamental differences between image descriptors.

1 Introduction

Histology is the field of biology that studies the architecture and composition of tissues at microscopic level. The examination of tissue slides is a fundamental tool in biology and medicine to support the decision making process in specialized tasks. In different clinical centers and research labs, these images are digitized to document procedures and to support findings, then, storing large numbers of tissue slides in centralized image repositories. Often, these image collections are being acquired in collaborative environments, and individual users are not fully aware of the potential contents of these image databases. However, these collections hide a latent source of information waiting to be exploited if the right mechanisms to access it are provided \cite{1}.

Content-based image retrieval (CBIR) systems in medical applications have been investigated during the last years, aiming to design systems with the ability to manage visual image contents and provide next generation Computer Aided Diagnosis tools, that are aware of image collection contents \cite{2}. Histology image retrieval has been the subject of several research works to model semantic similarity measures and to classify tissue slides in some semantic categories. Zhen et al. \cite{3} proposed a CBIR system based on four different visual characteristics to retrieve histology images from prostate, liver, and heart tissues. Tang et al. \cite{4}.
[4] designed a system to index histology images of the gastro-intestinal tract, splitting slides in different blocks to perform local analysis. These blocks are categorized in semantic classes and an image similarity measure is computed with respect to local findings. Naik et al. [5] designed a boosting algorithm using multiple distance measures computed on a fixed set of features to retrieve and classify breast histology slides. One common characteristic of these works is the use of multiple visual features extracted from images to achieve the underlying goal of retrieving or classifying histology images. However, strategies for feature combination are commonly underestimated, whereby the use of simple feature vector concatenations is done.

Different image characteristics often have their own structure and distribution that may be harnessed to design more effective image retrieval systems. For instance, feature histograms can be compared using a similarity measure for probability distributions while feature vectors can be matched using Euclidean metrics. In addition, even if two different features are being compared with the same metric, their scale, domain and distribution may be completely different due to the intrinsic descriptor nature. This information is critical to integrate and combine features when solving image retrieval and classification problems, and may also be useful to understand the relationships between semantic categories and visual characteristics [6].

In this paper, a late-fusion strategy is followed to combine low-level features for histology image classification and retrieval. Late-fusion indicates that each feature is processed and analyzed in an independent fashion after a similarity measure is computed between two images. Then, the combination strategy deals with processed similarity scores for each feature to produce the final result. The opposite strategy to late-fusion is early-fusion, in which feature vectors are concatenated altogether before applying any computation. In our framework, the parameters for feature normalization and integration are determined by statistical measures on distance distributions. Then, a unique image similarity measure is constructed using kernel functions to classify and retrieve images from the database. We used five different textural features in our experiments, and tested our framework in a collection of 2,828 images with various examples of the four fundamental tissues in biology. This paper is organized as follows: Section 2 presents an introduction to the image collection and the kind of histology images it contains. Section 3 presents the proposed methods and algorithms for feature extraction, feature combination and image classification and retrieval. The evaluation procedure and experimental results are presented in Section 4 and Section 5 discusses some concluding remarks.

2 Histology Image Database

The image collection used in this work comes from a database of 20,000 histology images that was acquired for academic and research activities in biology. The main purpose of the collection is to allow students, professors and researchers in the biomedical field to access a wide variety of microscopy images used to study
the four fundamental tissues of living beings: connective, epithelial, muscular and nervous. Figure 1 shows some image examples of each fundamental tissue to illustrate different biological structures. Images in this collection have been acquired from tissue slides taken from different mice organs including brain, liver, hearth, lung, kidney and skin among others. All samples were drawn from healthy specimens.

Fig. 1: Image examples of the four fundamental tissues

Tissue slides were prepared using different staining methods including Hematoxylin and Eosin (H&E) and Immunohisto chemical (IC) procedures. In addition, different zoom factors were used to acquire digital images according to the structure of interest. A portion of these images have been annotated by expert biologists, describing some particular structures, organ, system and fundamental tissue. About the 60% of the images in the database remains without annotations or text descriptions, making these images inaccessible using conventional text retrieval methods. Moreover, if all images had text annotations, content-based access is still desired to retrieve similar images using the query by example paradigm, to allow users perform a visual exploration of the collection as well.

3 Content-based Histology Image Indexing

The design and evaluation of a multi-feature image retrieval system is proposed in this paper. Two different tasks are considered to allow content-based access to the histology image collection. First, the computation of a similarity measure that incorporates different visual features to retrieve images given an example query. Second, the use of the combined similarity measure to classify histology images according to its fundamental tissue. This section presents the set of visual features included in this study, the proposed merging algorithm and the classification strategy.
3.1 Low-level Feature Extraction

Textures features together with architectural features have been suggested as prominent characteristics for histology image analysis [7,8]. In this work, five textural features have been selected to describe histology image contents. These features can be grouped according to the underlying data structure as feature vectors and feature histograms as described below.

**Feature Vectors** Three texture descriptors have been considered to build up image feature vectors. Each descriptor is computed per block in a $3 \times 3$ grid, leading to an image analysis in 9 different regions. Each feature vector is constructed bounded together the values computed in each block and preserving the spatial arrangement of the processed regions [9].

1. Gabor textures: using a convolution with a Gaussian harmonic function, and 7 different frequencies to compute 7 descriptor values per block. Then, the Gabor descriptor is composed of 63 features.
2. Tamura textures: two statistics are calculated for contrast, directionality and coarseness, providing 6 descriptors in each block. The Tamura descriptor has 54 features.
3. Zernike Moments: the absolute values of the coefficients of the Zernike polynomial approximation are computed per block, providing 72 descriptors in each region. Then, the Zernike descriptor has 648 features.

These feature vectors are evaluated using the Euclidean distance in the subsequent stages, which is computed as:

$$d_2(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

**Feature histograms** In addition to the three feature vectors, two histogram features are constructed using a bag-of-features approach, that may be considered as a texture analysis as well [10]. This strategy allows to estimate the presence of local patterns in images. First, a set of local patches or blocks are extracted from training images and a local descriptor is computed for each. Then, a dictionary of patterns is constructed using a vector quantization algorithm to merge together patches with similar visual appearance. In our implementation, the k-means algorithm was used to cluster similar patches and to set cluster centroids as dictionary elements. Finally, a histogram is computed for each image, counting the occurrence of each element in the dictionary among the blocks extracted from the image. The most important parameters of this image representation are the selection of the local descriptor and the size of the dictionary [11]. Two different strategies have been followed in this work, both using a dictionary size equal to 500 elements.
1. SIFT-based dictionary: Each block in the process is represented by the rotation-invariant feature descriptor, using a histogram of 128 bins.

2. DCT dictionary: Each block is represented by the coefficients of the Discrete Cosine Transform, applied to each channel of the RGB color space. The 21 most significant coefficients per channel are preserved. In this way, the dictionary of patterns will have color information as well.

These feature histograms are evaluated using the Histogram Intersection measure:

\[ d_{\cap}(x, y) = \sum_{i=0}^{n} \min\{x_i, y_i\} \]

Where \( x \) and \( y \) are histograms and \( x_i \) and \( y_i \) are their corresponding \( i \)-th bins. This is a similarity measure instead of being a distance measure, i.e. the more similar two images are, the larger the score is. The histogram intersection is also known to be a valid Mercer kernel for image classification [12], that is, the histogram intersection is a symmetric positive semi-definite function, so it can be used for learning non-linear patterns with Support Vector Machines.

3.2 Feature Combination

Late feature combination for histology image classification and retrieval is proposed in this work. In order to combine the available image features, two steps are taken into account: feature normalization and feature integration. Feature normalization is performed following an automated statistical analysis, and feature integration is achieved using kernel functions.

**Feature normalization** The goal of feature normalization is to guarantee the appropriate comparison of different measurements that differ in scale and domain, while preserving the underlying characteristics of the data. In the proposed late-fusion strategy, the data to be normalized is the distance metric computed from different image features. In this Subsection we assume, without loss of generality, that the underlying metric is a distance measure that meets the properties of symmetry, non-negativity and triangular inequality. Using a training sample from the image collection, a distance or similarity matrix is computed to evaluate its distribution and composition, and also to determine the normalization parameters.

Let \( d \) be the distance value computed from a particular image feature. The statistical normalization is computed as:

\[ \hat{d} = \frac{(d - \mu)}{\sigma}, \]

where \( \hat{d} \) is the standardized distance value and \( \mu \) and \( \sigma \) are the mean and variance of the underlying distance distribution. Since the distance distribution depends on the feature structure and image contents, the normalization parameters are unknown a priori. One can assume a normal distribution of the distances
to estimate $\mu$ and $\sigma$ without any further analysis. However, the true distribution of distances might be approximated using other Probability Distribution Functions (PDF), whose parameters are estimated in different ways. To overcome this problem, we consider a set of $K$ possible PDFs, denoted by $f_k$. Using the distance samples in the computed matrix, the parameters $\Theta_k$ of each possible $f_k$ are estimated to determine whether the data is being drawn from the associated PDF or not. Finally, to select the best distribution approximation for the underlying data, the Kullback-Leibler (KL) divergence is evaluated between the histogram of actual distances and the estimated PDF.

The following steps summarize the complete procedure:

1. For each $f_k$ in the set of possible PDFs, the parameters $\Theta_k$ are estimated from the distance matrix $D$.
2. A histogram $H$ of distances with $m$ bins is built up using the samples in $D$.
3. For each bin in the histogram $H$, the value of $f_k$ is calculated in $\hat{f}_k$ to approximate the shape of $H$.
4. To select a PDF out of all possible $f_k$’s, the KL-divergence is calculated between the histogram $H$ and the approximation $\hat{f}_k$. The minimum divergence indicates the best fit.

Table 1 lists the six PDFs considered in this work to approximate distance distributions. Notice that for each PDF, the normalization parameters $\mu$ and $\sigma$ depend on the estimated parameters $\Theta$ that are computed from the sample, according to the corresponding rules.

### Table 1: Probability Distribution Functions in the set of possible approximations for each distance distribution.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>PDF</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal $\Theta=(\mu, \sigma)$</td>
<td>$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>Gamma $\Theta=(k, \theta)$</td>
<td>$x^{k-1} \frac{\exp(-x/\theta)}{\Gamma(k)\theta^k}$</td>
<td>$k\theta$</td>
<td>$k\theta^2$</td>
</tr>
<tr>
<td>Laplace $\Theta=(\mu, b)$</td>
<td>$\frac{1}{2b} \exp\left(-\frac{</td>
<td>x-\mu</td>
<td>}{b}\right)$</td>
</tr>
<tr>
<td>Log-norm $\Theta=(\mu, \sigma)$</td>
<td>$\frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\ln(x-\mu)^2}{2\sigma^2}\right)$</td>
<td>$e^\mu$</td>
<td>$\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}$</td>
</tr>
<tr>
<td>Rayleigh $\Theta=(\sigma)$</td>
<td>$\frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$</td>
<td>$\sigma\sqrt{\frac{\pi}{2}}$</td>
<td>$\frac{4-\pi}{2} \sigma^2$</td>
</tr>
<tr>
<td>Exponential $\Theta=(\lambda)$</td>
<td>$\lambda \exp\left(-\lambda x\right)$</td>
<td>$\frac{1}{\lambda}$</td>
<td>$\frac{1}{\lambda^2}$</td>
</tr>
</tbody>
</table>
Feature integration Feature integration is performed by the design of a kernel function that includes the information of all distances, which can be then used for image classification using Support Vector Machines (SVM), and also for image retrieval using the kernel value as ranking function. In this framework, we included feature vectors and feature histograms that are compared using different metrics, on the one hand the Euclidean distance and on the other hand the histogram intersection similarity. Although they are different in terms of behavior, one can easily combine them following some basic properties of kernel functions.

Then, the construction of the fused kernel is divided up in two parts: first, a kernel based on available Euclidean distances, and second, a kernel based on histogram intersections. The first part is done using a linear combination of the Euclidean distance on the set of feature vectors:

$$d_{\Sigma}(x, y) = \sum_{v \in F} w_v d(x_v, y_v),$$  \hspace{1cm} (1)

where $x$ and $y$ are images, $F$ is the set of feature vectors, $x_v$ and $y_v$ are the $v$-th feature vector of each image and $d$ is the Euclidean distance. Notice that this formulation allows the incorporation of factors $w_j$ to weight the relative importance of each feature in the integration process. Then, this linear combination of Euclidean distances is transformed into a kernel function using a Radial Basis Function (RBF) or Gaussian kernel as follows:

$$k_g(x, y) = \exp \left(- \frac{d_{\Sigma}(x, y)^2}{2\sigma^2} \right)$$ \hspace{1cm} (2)

The second part to construct the fused kernel function is to take advantage of the properties of the histogram intersection as kernel function and use it outside of the Radial Basis Function. In this way, the kernel $k_g$ is only defined for feature vectors and the new expression adds the remaining histogram intersection kernels as follows:

$$k_{\Sigma}(x, y) = k_g(x, y) + \sum_{h \in H} d_{\cap}(x_h, y_h)$$ \hspace{1cm} (3)

where $H$ is the set of feature histograms, $x_h$ and $y_h$ are the $h$-th histogram of images $x$ and $y$ respectively, and $d_{\cap}$ is the histogram intersection kernel. This results in a new valid kernel since the histogram intersection is a Mercer kernel and it has also been proved that linear combination of kernels is a valid kernel as well [13]. The formulation of the kernel in Equation 3 is followed in this work as the fused kernel function, setting all weights to 1, to give equal importance for all features and to rather evaluate the effects of the normalization procedure.

3.3 Image Classification and Retrieval

Image classification is performed using SVM for each category of interest. In this work, the four fundamental tissues are used as classification target to evaluate
the proposed strategy. These broad categories also serve as ground truth for image retrieval experiments, in which users search for similar images under a query by example paradigm, expecting as result relevant images in the same tissue category.

For image classification, the proposed fused kernel is used to train support vector machines after distance normalization has been done for the particular class distribution. Distance normalization according to the target class distribution may offer a performance improvement since the data is scaled with respect to the typical values of each class, thus creating an outlier effect for other class values.

In an image retrieval application under the query by example paradigm, we do not know in advance the correct class of query images. However, using the classification strategy, the system can identify the most probable category of the query and use the appropriate kernel function to rank images in the database. In this experimentation, we do not evaluate this cascade strategy for image retrieval, instead, a unified kernel measure is computed to solve all kinds of queries.

4 Experiments and Results

To evaluate the classification and retrieval performance a subset of images from the collection was selected with representative samples of the four fundamental tissues. This image data set is composed of 2,828 images, distributed in 484 for connective tissue, 804 for epithelial tissue, 514 for muscular tissue and 1026 for nervous tissue. The data set was split in two parts, using 80% for validation and training, and 20% for final tests. Three main stages are considered for experimentation, including feature combination findings, classification performance and retrieval response. In each stage of the experimental evaluation, the performance measures are presented and discussed.

4.1 Feature Combination

Feature combination is done using the normalization parameters from the statistical analysis of distances. Figure 2 present the histograms of distance measures for each feature organized by class. Observing the table by columns, one can notice that each feature tends to have a similar PDF among the different categories, which is consistent with the structure of each particular feature space. However, the particular parameters that describe the behavior of each feature in different classes may vary according to the contents and properties of the corresponding category. When the table is observed among rows, one can notice that features have a very different shape from each others. To the naked eye, only the fourth row may be assumed as a normal distribution, while the others apparently require other PDFs to be described appropriately. It is important to notice that the first 3 rows of Figure 2, show histograms distributions for distance measures while the 2 last rows show histograms distributions for similarity measures. As was mentioned before, this is an important difference since the behavior of a
Fig. 2: Probability distributions for image features. In columns each class is presented. Rows are indexed by image features in the following order from top to bottom: Tamura, Zernike, Sobel, SIFT, DCT.
distance measure grows as images become different whereas a similarity measure grows as images are more similar. This explains the distribution of histograms in the last row, which are biased to zero, indicating that most of the images are very different according to the DCT-based visual descriptor.

Fig. 3: Distribution of Gabor distances within the muscular category. Plot labels show the KL-divergence scores computed for each possible PDF. The best fit, shown in bold-continuous line, minimizes the score.

After processing distance matrices, the parameters for all possible PDFs are computed in an attempt to match the true distribution of the distances. An approximation of the distribution is then calculated for each bin of the empirical histogram using the estimated parameters. To identify the best approximation to the distribution of distances, the PDF that minimizes the KL-divergence score is selected. Figure 3 shows an example of the approximation made with different PDFs to the distribution of distances using the Gabor feature vector in the training set of muscular tissue images. The specific scores of the KL-divergence calculated between the PDF approximation and the empirical histogram of distances is presented in the Figure. Notice that the best fit has the lowest score among the different possible PDFs. This process is applied to all distance distributions and the estimated parameters of the best fit are recorded for further normalization purposes.

Table 2 shows the estimated parameters for the best fit for each feature descriptor. Notice that, when one descriptor is approximated with the same PDF along different categories, the estimated parameters that describe the distance
distributions are different, and the system can take advantage of this differences to better discriminate image categories.

Table 2: Best fit for each distance distribution with the estimated parameters.

<table>
<thead>
<tr>
<th></th>
<th>Connective</th>
<th>Epithelial</th>
<th>Muscular</th>
<th>Nervous</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gabor</strong></td>
<td>Log-norm</td>
<td>Gamma</td>
<td>Log-norm</td>
<td>Log-norm</td>
</tr>
<tr>
<td>µ = -0.697</td>
<td>σ = 0.624</td>
<td>k = 3.68</td>
<td>µ = -0.961</td>
<td>σ = 0.607</td>
</tr>
<tr>
<td><strong>Tamura</strong></td>
<td>Gamma</td>
<td>Gamma</td>
<td>Log-norm</td>
<td>Log-norm</td>
</tr>
<tr>
<td>k = 1.56</td>
<td>σ = 1.288</td>
<td>k = 1.63</td>
<td>µ = 6.957</td>
<td>µ = 7.012</td>
</tr>
<tr>
<td><strong>Zernike</strong></td>
<td>Laplace</td>
<td>Log-norm</td>
<td>Gamma</td>
<td>Log-norm</td>
</tr>
<tr>
<td>µ = 0.0762</td>
<td>σ = 0.128</td>
<td>µ = -2.513</td>
<td>k = 60.99</td>
<td>µ = -2.531</td>
</tr>
<tr>
<td>b = 0.0693</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SIFT</strong></td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>µ = 593</td>
<td>σ = 163</td>
<td>µ = 587</td>
<td>µ = 504</td>
<td>µ = 609</td>
</tr>
<tr>
<td>σ = 203</td>
<td></td>
<td></td>
<td>σ = 184</td>
<td></td>
</tr>
<tr>
<td><strong>DCT</strong></td>
<td>Exponential</td>
<td>Exponential</td>
<td>Exponential</td>
<td>Exponential</td>
</tr>
<tr>
<td>λ = 0.0024</td>
<td></td>
<td>λ = 0.0024</td>
<td>λ = 0.0029</td>
<td>λ = 0.0026</td>
</tr>
</tbody>
</table>

4.2 Image Classification

For image classification we evaluated first each feature separately and the proposed combined kernel as well. For this experiments, the training set was used to determine configuration parameters for the classifier, including both, the complexity of the SVM and the parameter σ of the RBF kernel. The parameter tuning task was conducted applying 10-fold cross validation. After the training algorithms were applied, the test set is used to determine the real ability of classifiers to discriminate between histology image categories. The performance measure used in this evaluation is F-measure, that is the harmonic mean between precision and recall, so the higher their values, the better the performance.

Table 3 shows the F-measure for all categories discriminated by feature. The individual feature that performed the best was the DCT-based bag-of-features, which also got the best general performance. The Table shows that feature vectors (Gabor, Tamura, Zernike) have in general a poor performance with respect to feature histograms (SIFT and DCT bag-of-features). Hence, we investigated the combination of features in three different levels as is shown in Table 3: feature vectors combination, feature histograms combination and all features combination. The feature vectors combination showed a successful improvement with respect to individual feature vectors, with an average F-measure increase of 9.5% with respect to the best individual feature (Tamura). For all classes, the feature vectors combination has outperformed individual features.

On the other hand, feature histograms combination does not showed an improvement over the best individual feature histogram, suggesting that some
Table 3: Image classification performance. F-measure.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Connective Epithelial</th>
<th>Muscular</th>
<th>Nervous</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor</td>
<td>0.36</td>
<td>0.33</td>
<td>0.32</td>
<td>0.49</td>
</tr>
<tr>
<td>Tamura</td>
<td>0.39</td>
<td>0.48</td>
<td>0.29</td>
<td>0.53</td>
</tr>
<tr>
<td>Zernike</td>
<td>0.33</td>
<td>0.39</td>
<td>0.26</td>
<td>0.40</td>
</tr>
<tr>
<td>Vector comb.</td>
<td>0.42</td>
<td>0.50</td>
<td>0.36</td>
<td>0.55</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.64</td>
<td>0.83</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>DCT</td>
<td><strong>0.84</strong></td>
<td><strong>0.88</strong></td>
<td><strong>0.78</strong></td>
<td><strong>0.88</strong></td>
</tr>
<tr>
<td>Histogram comb.</td>
<td>0.71</td>
<td>0.86</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td>All features comb.</td>
<td>0.68</td>
<td>0.71</td>
<td>0.60</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Formal loss during the combination process. In particular case with two feature histograms, the performance of DCT is clearly better than SIFT for all categories. Even though the difference is not as large as that observed with feature vectors, the combination of both histograms does not lead to a better classification performance. This may be partially explained by the radical difference in the distribution of the similarity measures, that, according to Table 2, has been set to Exponential for DCT and Normal for SIFT. Despite statistical normalization, the resulting combined feature space is not able to preserve the intrinsic structure and discrimination power of the DCT feature space.

Finally, the combination of all features in a unique kernel function results in an improvement with respect to the feature vectors but not at all with respect to feature histograms.

4.3 Image Retrieval

For image retrieval experiments, the data set of 2,828 images was split in a different random partition to use as query images a 10% of the collection, that is, 280 images were chosen to serve as queries. Then, the evaluation consisted on taking each query image and ask the system to retrieve the most similar ones according to the underlying strategy. To determine if the results are relevant or not with respect to the query, the category information is used as ground-truth.

The experimental evaluation started again observing the response of individual low-level features, as is shown in Table 4. The performance measures presented in this Table are Mean Average Precision (MAP); R-prec point, that is, the point in which precision and recall get the same value; and Precision after the first 20 results, i.e. the proportion of relevant images in the first page of results.

In this evaluation, feature vectors do not present a good enough performance with respect to feature histograms. The large difference in performance between feature vectors and feature histograms makes it difficult to obtain an overall improvement when combining all features in a unique similarity measure. However, the combination of feature vectors is able to produce improved results with respect to individual similarity measures. In the case of feature histogram combination, the result is also improved and remains comparable to the performance of
Table 4: Image retrieval performance

<table>
<thead>
<tr>
<th>Feature</th>
<th>MAP</th>
<th>R-prec</th>
<th>Prec 20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random Search</strong></td>
<td>0.201</td>
<td>0.212</td>
<td>0.197</td>
</tr>
<tr>
<td>Gabor</td>
<td>0.267</td>
<td>0.255</td>
<td>0.244</td>
</tr>
<tr>
<td>Tamura</td>
<td>0.259</td>
<td>0.262</td>
<td>0.250</td>
</tr>
<tr>
<td>Zernike</td>
<td>0.255</td>
<td>0.256</td>
<td>0.253</td>
</tr>
<tr>
<td><strong>Vector comb.</strong></td>
<td>0.260</td>
<td>0.262</td>
<td>0.250</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.327</td>
<td>0.308</td>
<td>0.597</td>
</tr>
<tr>
<td>DCT</td>
<td>0.323</td>
<td>0.301</td>
<td>0.663</td>
</tr>
<tr>
<td><strong>Histogram comb.</strong></td>
<td>0.328</td>
<td>0.308</td>
<td>0.604</td>
</tr>
<tr>
<td><strong>All features comb.</strong></td>
<td>0.286</td>
<td>0.286</td>
<td>0.336</td>
</tr>
</tbody>
</table>

the best individual feature histogram. Figure 4 presents the recall-precision plot for the three combination strategies. Notice that even though the performance of the all-features-combination is improved with respect to the feature-vectors-combination, the behavior and tendency of all-features-combination is closer to the lower bound rather than the upper bound.

![Recall-Precision Plot](image_url)

Fig. 4: Recall-Precision plot for feature image retrieval response using three different combination strategies: feature vectors combination, feature histograms combination and all features combination.

These experimental results have shown that combining visual features may lead to an improved response when features have a similar behavior in terms of statistical distribution and performance. In addition, it has been shown that
combining features with different properties not always necessarily lead to obtain better performance.

5 Conclusions and Future Work

This paper has presented and extensive experimental evaluation of a strategy for feature combination in histology image classification and retrieval. This strategy is based on late fusion of distance measures as opposed to the traditional concatenation of feature vectors. Experimental results showed that features with similar statistical behavior and similar performance may be effectively combined to achieve an improved system response. On the other hand, when the underlying features have different nature, different performance response or a radical difference in statistical distribution, the combined representation does not necessarily follow an increased performance.

This experimental evaluation supports other studies in statistical learning theory [14], that have shown that features with similar performance may be combined together to effectively improve the response of a learning machine, while features with different behavior may decrease the final result. In that sense, feature combination for image retrieval should be strongly guided by a careful analysis of the feature structure and distribution to guarantee the best response in real production systems.

Another important aspect of feature combination strategies is the modeling of weighting schemes to adjust the relative importance of different features according to the underlying classification task. In this paper, it was mentioned that these parameters may be easily included to adapt the image representation according to high level criteria. This may significantly improve the response of image classification and retrieval systems as has been suggested in other image retrieval studies [15]. This kind of strategies to automatically adjust the image representation for solving problems in histology applications will be part of our future work.

References


