Acta biol. Colomb., Vol. 15 N.º 3, 2010 251 - 260

A KERNEL-BASED MULTI-FEATURE IMAGE REPRESENTATION FOR HISTOPATHOLOGY IMAGE CLASSIFICATION

Una representación multi-características de imágenes basada en kernels para clasificación de imágenes de histopatología

MORENO J.[1], Est-M.Sc.; CAICEDO J.[1], Est-Ph. D.; GONZÁLEZ F.[1], Ph. D.
[1]Grupo de Investigación Bioingenium, Facultad de Medicina-Facultad de Ingeniería, Universidad Nacional de Colombia, Bogotá, D.C, Colombia. {jgmorenofr, jccaicedoru, fagonzalezo}@unal.edu.co
Corresponding author: Jose G. Moreno. Universidad Nacional de Colombia, Carrera 30 # 45-03, Of. 207, Edif. 453 (Aulas de Ingeniería), Bogotá, D.C., Colombia. jgmorenofr@unal.edu.co

Presentado 15 de marzo de 2010, aceptado 4 de junio de 2010, correcciones 23 de junio de 2010.

ABSTRACT

This paper presents a novel strategy for building a high-dimensional feature space to represent histopathology image contents. Histogram features, related to colors, textures and edges, are combined together in a unique image representation space using kernel functions. This feature space is further enhanced by the application of Latent Semantic Analysis, to model hidden relationships among visual patterns. All that information is included in the new image representation space. Then, Support Vector Machine classifiers are used to assign semantic labels to images. Processing and classification algorithms operate on top of kernel functions, so that, the structure of the feature space is completely controlled using similarity measures and a dual representation. The proposed approach has shown a successful performance in a classification task using a dataset with 1,502 real histopathology images in 18 different classes. The results show that our approach for histological image classification obtains an improved average performance of 20.6% when compared to a conventional classification approach based on SVM directly applied to the original kernel.

Key words: Automatic image annotation, machine learning.

RESUMEN

Este trabajo presenta una estrategia nueva para la construcción de un espacio de características de gran dimensionalidad para la representación del contenido de imágenes de histopatología. Histogramas de características, relacionados con colores, texturas y bordes, son combinados para obtener una única representación de la imagen utilizando funciones de kernels. Este espacio de características es mejorado mediante la aplicación de Análisis de Semántica Latente, para modelar relaciones ocultas entre los patrones visuales. Esta información es incluida en la representación de la imagen en el nuevo es-

pacio. Luego, un clasificador de Máquinas de Vectores de Soporte es utilizado para asignar etiquetas semánticas a las imágenes. Algoritmos de procesamiento y de clasificación son utilizados en las funciones del kernel, por lo que la estructura del espacio de características es completamente controlada mediante medidas de similitud y la representación dual. El enfoque propuesto mostró un desempeño exitoso en la tarea de clasificación con un conjunto de datos de 1.502 imágenes reales de histopatología en 18 clases diferentes. Los resultados muestran que nuestro enfoque para la clasificación de imágenes histológicas obtiene una mejora promedio en el rendimiento del 20,6% en comparación con un método de clasificación convencional, basado en la aplicación de una Máquina de Vectores de Soporte sobre la función de kernel original.

Palabras clave: anotación automática de imágenes, aprendizaje máquina.

INTRODUCTION

Histology samples have been successfully used to diagnose a wide range of diseases, specially related to cancer (Lee *et al.*, 2009). Pathologists evaluate histology sections under the microscope to identify affected biological structures and determine the grade of the lesion. Observations under the microscope can be captured using a digital camera to document the case, elaborate reports and facilitate the exchange of information in research. Pathology departments in hospitals and universities can accumulate large amounts of histology images in digital format, whether they are archived in a unified information system, a common storage service or personal databases. These image collections may be used as reference material to support the decision making process in clinical services, for instance, when a physician wants to observe previously diagnosed cases, either by himself or other experts, to confirm the current findings. Also, a histopathology image collection may be used in research to follow up a disease across different patients, observing tissue mutation or recovery and characterizing visual patterns. Accessing a large collection of medical images in order to obtain useful information is a difficult task. A human being can only manage a limited amount of images in personal archives, on the order of some hundreds. In a collaborative environment, such as an information system in a large hospital, users are not aware of the complete contents in the image collection. So that, when users need some images to support a particular task, they will often review only previously known records, limiting the possibility of accessing other useful information. Automatic analysis of image contents has been proposed for a wide range of applications (Smeulders *et al*., 2000; Müller *et al*., 2004). In the area of histology images, some approaches have been proposed to design visual similarity measures among image contents (Caicedo *et al*., 2008; Caicedo *et al*., 2009; Lee *et al*., 2009) and to automatically annotate histopathology images (Caicedo *et al*., 2009). The main purpose of these approaches is to be able to identify the semantic meaning of the image content to support case-based reasoning or evidence based medicine.

In this paper, we propose a novel strategy to represent visual image contents to feed an automatic annotation system for histopathology images. Previous studies have shown that histopathology patterns rely on different characteristics such as color, morphology and textures (Caicedo *et al*., 2009; Caicedo *et al*., 2008). To increment the accuracy of

classification algorithms, we involve different visual features that complement each other. We propose an efficient and extensible approach to include many features in the same representation space to classify image contents. The combination of features is performed using kernel functions that map input data to a new feature space in which all visual features are combined. In addition, this combined feature space is processed using Latent Semantic Kernels to identify statistical relationships between visual patterns, and the obtained information is also included in the new image representation. Since the proposed algorithms are based on kernel functions, the input data in our framework may be easily extended to use structured objects such as graphs or trees. In fact, the proposed feature representation in this work is based on histogram structures, which are mapped to a feature space using a kernel function to exploit the geometric structure of such data.
The proposed approach results into a rich feature space to represent the visual contents of histopathology images. Afterwards, the task is to assign to each image a subset of labels from a predefined semantic vocabulary that describes histopathology image concepts. We used Support Vector Machine (SVM) classifiers to identify semantic concepts in images given a multi-feature representation. Experimental results showed an important improvement in classification performance using the proposed approach, with respect to individual visual features strategy. This paper is organized as follows: Section 2 presents the classification problem and the feature extraction, Section 3 describes the methods and models applied in the proposed method. The experimental setup and experimental evaluation is presented in Section 4 and finally, Section 5 presents some concluding remarks.

MATERIALS AND METHODS

Images used in this work have been acquired to diagnose basal cell carcinoma, the most common skin cancer in white populations (Wong *et al.*, 2003). The whole histopathology collection is composed of 5,995 images at 1,280×1,024 pixels, acquired under a Nikon microscope at the Pathology Lab. In the acquiring process a set of clinical cases was selected, related to different patients. Each patient sample was put under the microscope and after a visual inspection some images are captured at different zoom levels. A histology section is evaluated by pathologist to grade the severity of the lesion, follow up the patient condition and recommend treatments. An image collection is constantly feed at the Pathology Laboratory of the National University of Colombia with new cases and new patients, for clinical evaluations and research activities (Caicedo *et al.*, 2008). Images were stored in a common directory in JPG format for their later analysis. From this collection, researchers selected a subset of 1,502 images with different structures and lesion severities, organizing them according to 18 concepts related to carcinoma patterns or biological structures. As a result, the later process gives a dataset with images and descriptions of their related concepts, in which one image may contain several concepts associated to it. The challenge in this dataset is to identify as many correct concepts for each image from all the possible ones that we can assign to each, that is why this is not a multiclass classification problem. The list of histopathology concepts is shown on Table 1.
This dataset contains image examples of diverse carcinoma patterns and the collection also includes a set of images without any lesion. The dataset has been previously used for testing content-based image retrieval systems and classification systems as well (Caicedo *et al.*, 2009; Caicedo *et al.*, 2008).

	Baseline		Multi-Feature Representation Space						
Concepts	P	R	F	P	R	F	k	α	Performance Improvement
Pilosebaceous annexa	0.58	0.43	0.49	0.59	0.54	0.57	1,000	0.75	15.3%
Cystic change	0.85	0.62	0.72	0.88	0.71	0.79	100	0.25	10.1%
Elastosis	0.79	0.38	0.52	0.88	0.42	0.57	200	0.25	10.5%
Eccrine glands	0.70	0.21	0.32	0.65	0.35	0.46	400	0.25	43.2%
Lymphocyte infiltrate	0.57	0.33	0.42	0.53	0.41	0.46	200	0.25	10.6%
Perineural invasion	0.00	0.00	-	0.00	0.00	-	-	-	-
Lesion with fibrosis	0.83	0.55	0.66	0.82	0.61	0.70	800	0.75	6.0%
Micronodules	1.00	0.17	0.29	0.67	0.33	0.44	600	0.25	55.6%
Necrosis	1.00	0.40	0.57	1.00	0.40	0.57	200	0.75	0.0%
N-P-C. elastosis	0.75	0.50	0.60	0.88	0.58	0.70	1,200	0.25	16.7%
N-P-C. fibrosis	1.00	0.15	0.25	1.00	0.38	0.55	100	0.75	114.3%
N-P-C. infiltrate	0.68	0.29	0.41	0.59	0.42	0.49	100	0.50	21.5%
N-P-C. pilosebaceous annexa	0.33	0.09	0.14	0.33	0.09	0.14	100	0.75	0.0%
N-P-C. trabeculae	0.00	0.00	-	0.50	0.25	0.33	800	0.25	-
Morpheaform pattern	0.82	0.58	0.68	0.83	0.63	0.71	300	0.75	4.7%
Rod trabeculae	0.83	0.20	0.32	0.67	0.27	0.38	200	0.25	18.1%
Ulceration	0.50	0.20	0.29	0.50	0.20	0.29	300	0.75	0.0%
Sanguineous vessel	0.60	0.24	0.35	0.54	0.27	0.36	100	0.50	3.7%

Table 1. Precision (P), Recall (R), F-measure (F) of baseline and the proposed method for the 18 concepts. Values showed are the results for the best parameters combinations in Multi-Feature Representation Space.

Feature extraction

The purpose of this work is to include different features in the image representation. We selected six different histogram features, each capturing the probability distribution of different characteristics in the image. The first two histograms are related to color and luminance: the RGB color histogram, with 512 bins, and the Gray scale histogram. Two additional histograms related to texture have been extracted: Tamura Textures (Deselaers *et al*., 2004) and Local Binary Patterns (Berman *et al*., 1998). The last two histograms capture information of edges and invariant features: Sobel histogram and Invariant Feature histogram (Deselaers *et al*., 2004). Note that all features in this evaluation are global histograms, measuring the overall distribution of visual characteristics in histopathology images. To avoid the problem of working with histograms from different representations, histograms are normalized from frequency distributions to probability distributions, that is, adding the values in their bins, sums up to 1.

Also, we assume these distributions capturing enough global information from image content, even though similar images can have different histograms. In content-based systems, a common problem occurs when similar images have different meanings or different images have the same meaning, this is called the semantic gap and we tackle this problem using machine learning strategies to map low level features to histopathology concepts.

We did not evaluate different feature configurations to observe whether they help to improve the results because this is not part of our contribution. The feature configuration

has been selected just as has been reported in the literature (Barla *et al*., 2003), to make the process easier we use 512 bins in the histogram representations.

Multi-Feature Representation Space

Kernel functions. Since input objects in this work are histogram features, we chose the Histogram Intersection Kernel (Barla *et al*., 2003) to evaluate the similarity between a pair of features and to embed the visual characteristics in a high-dimensional vector space. This similarity measure calculates the common area between two given histograms, and this operation has been proved to be a valid Mercer's kernel (Barla *et al*., 2002). The computation of this kernel is as follows:

$$k_{\cap}\left(h_1^j, h_2^j\right) = \sum_{i=1}^{n} \min\left(h_1^{j(i)}, h_2^{j(i)}\right)$$

where n_1 and n_2 are the j-th histogram of images 1 and 2 respectively, and $h^{(i)}{}_1$, $h^{(i)}{}_2$ are the i-th bins in the j-th histogram for image 1 and image 2 respectively. This operation is evaluated between a pair of histograms of the same nature, e.g. between color histograms or texture histograms. This leads to a different feature space for each visual characteristic. In order to concatenate the corresponding feature spaces for each visual characteristic, we designed a new kernel function that corresponds to the linear combination of the histogram intersection kernel evaluated in different feature spaces:

$$k_c(I_1, I_2) = \sum_{j \in F} k_{\cap}\left(h_1^j, h_2^j\right)$$

where I_1 and I_2 are images. All features have the same weight with this formulation, i.e. all features are of the same importance. Using the kernel function k_c, images are now embed in a high dimensional feature space containing all visual characteristics. This new feature space will be named the combined feature space through this work. Note that this embed leads to a feature space that is completely different to one in which all histograms are concatenated before the evaluation of the kernel function, since the structure of each histogram is preserved in the resulting high dimensional feature space. We can build a kernel matrix that contains the values obtained with the kernel function combination between the images in the collection. This matrix is defined as:

$$K_{ij} = k_c(I_i, I_j)$$

This matrix can be rewritten like a $K = DD'$, where the D matrix is a term-document matrix and the kernel function used is a dot product.

Finding visual pattern relationships

Additionally to concatenate all feature spaces, we want to explicitly include information of the correlation among different visual characteristics. Latent Semantic Analysis (LSA) is a technique used in text document (Deerwester *et al*., 1999) for addressing two commons problems: polysemy and synonymity. Singular Value Decomposition (SVD) is applied to the term-document matrix in this technique, and using only the Eigenvectors with the k largest Eigenvalues, it is possible to build a new space to represent the documents. Semantics relationships are found in the new space because the terms are

mapped to concepts. The use of LSA is herein proposed to discover latent associations between visual patterns in the combined feature space (Monay *et al*., 2003). Imagine that histopathology images frequently reveal texture patterns with a purple color, or some pink edges. Since feature spaces have been modeled in an independent way, these relationships are not quite explicit in the combined feature space. We would like to include this information in the image representation as well, so that, in addition to the coordinates that are indexed by colors, textures and edges, we want to index additional coordinates with those feature associations that reveal latent patterns in histopathology images.
The algorithm to tackle this problem in a kernel-induced feature space is Latent Semantic Kernels (LSK; Cristianini *et al*., 2002), that allows us to perform Latent Semantic Analysis without having explicit feature vectors for the image representation. So far, the combined feature space is an image representation space induced by the linear combination of different visual characteristics evaluated with the histogram intersection kernel. Using a training sample of images in the combined feature space, LSK identifies an orthogonal basis for the training data, sorted according to the directions of maximum variance. This is done by decomposing the kernel matrix for a training dataset in Eigenvectors and Eigenvalues.
We want to model the directions of maximal variance of the data in the combined feature space, so that, a dual representation of the vectors in the feature space is required. Assuming *D* as the matrix with the feature vectors in the combined feature space, we have:

$$D' = U\Sigma V' \longrightarrow K = V\Sigma U' U\Sigma V' = V\Sigma^2 V' = V\Lambda V',$$

where *U* is the matrix with the Eigenvectors of the co-occurrence matrix and the Eigenvectors of the kernel matrix. We want to project any given feature vector in the first components of to reorganize the dataset in terms of the directions of maximal variance. This can be done as follows (Cristianini *et al*., 2002):

$$\phi(I)U_k = \left(\lambda_i^{-\frac{1}{2}} \sum_{j=1}^{l} (v_i)_j k_c(I_j, I) \right)_{i=1}^{k},$$

where $\phi(I)$ is the image representation in the combined feature space, (λ_i, v_i) is the i-th Eigenvalue, Eigenvector pair of the kernel matrix, and k_c is the kernel function in the combined feature space. This new representation of the data only contains the information of correlated patterns, and this will be named the latent semantic space in this paper. The kernel in the latent semantic space can be calculated using:

$$k_l(I_1, I_2) = \phi(I_1)U_k \; U_k' \; \phi(I_2)',$$

Note that all the computations can be reduced to evaluations of the kernel function in the combined feature space, k_c. By projecting the image representation into a latent semantic space with reduced dimensions, an implicit feature fusion leads to index the coordinates of this new space with the most significant correlations between visual patterns in the combined feature space.

Multi-Feature Representation Space

So far, we have two different representation spaces for image contents: the combined feature space and the latent semantic space. These spaces offer different advantages,

the combined feature include all information without information discard, but on the other hand, the visual feature relationships are present in the latent semantic space. To complete the multi-feature representation space, a convex combination of both feature spaces is designed:

$$k_m(I_1,I_2) = \alpha\, k_c(I_1,I_2) + (1-\alpha)k_l(I_1,I_2),$$

where $\alpha\varepsilon[0,1]$. This combination embeds the image representation in a feature space with the information of the combined feature space concatenated with the latent semantic space. The parameter α allows controlling the importance of each coordinate in the final image representation. The complete image representation in the multi-feature representation space is calculated by evaluating the new kernel function in a pair of histopathology images.

Experimental setup

An experiment with a set of 1,502 histopathology images was performed. These images were labeled by experts with 18 different labels, where each one may be present or not in each image. Low level features (colors, textures, edges) were taken into account for the representation of the images. The set was split into two subsets, 1,201 images for training to build latent semantic space and tune the parameters of the SVM, and the other 301 images to evaluate the performance of the proposed strategy. Three different values of α were evaluated and the results obtained with the original kernel were used as the baseline. Note that when $\alpha = 0$ we have the combined features space, and the latent semantic space is obtained when $\alpha = 1$. With $\alpha = 0.5$, the combined features space is as weighted as latent semantic space, but when $\alpha = 0.25$ and $\alpha = 0.75$ we assigned different weights for the spaces. We test all these values, but only the last three values for α are related to the multi-features representation space and the first one is the baseline.

These experiments give insights on how the method performs, however a more systematic evaluation of the impact of this parameter has to be done. This is part of our future work. The measures used to evaluate the results are Precision, Recall and F-measure which are defined as follows:

Where *tp* are the number of images with the correct histopathology concept assigned, *fp* are the number of images that the system assigned incorrectly assigned the histopathology

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F = 2\frac{Precision * Recall}{Precision + Recall}$$

concept, and f^n is the number of the images that contained the histopathology concept but the system did not find. We used F-measure to tune the parameters because it provides a unified measure that includes precision and recall simultaneously. Also, specific Precision and Recall measures are reported for the final test.

RESULTS

We used as baseline the values of Precision, Recall and F-measure obtained to each concept with a classifier that uses only one feature set. Our method did not compare with other methods, because these experiments were made to compare classifiers that use a multi-feature strategy against a classifier that use only the adding features to determine if using our latent multiple features lead to better results.

Our reported results are based on experiments conducted on a training set with 1,200 images and a test set with 301 images. 10-folding cross validation was used for tuning the parameters with the training set, and the measure used to select the better SVM parameters was F-measure. Then, the trained SVM model is applied to the test set; these results used the same parameters found in the 10-fold cross validation process. It is not necessary to repeat over the test set because there are not variances.

Table 1 shows in the three first columns Precision, Recall and F-measure obtained for the baseline and the next columns show the best values of Precision, Recall and F-measure obtained by the proposed method for each of the concepts in the histopathology images with the respective values of α and k. The last column shows the Performance Improvement obtained with respect to F-measure results of the baseline. The obtained results show that the proposed model provides better performance with respect to the baseline and the average improvement obtained for all histopathology concepts is 21%. For N-P-C. fibrosis the improvement can be easily observed, however, in some cases we do not get improvement at all or it is very low. In three cases the method does not show an improvement, this is the case of Necrosis, N-P-C. infiltrate and Perineural invasion concepts. To Perineural invasion, the F-measure is not reported because it cannot be calculated, since the Recall and Precision values are zero. To N-P-C. trabeculae, the baseline do not get any good results and the proposed method assigned some rights labels but the Performance Improvement cannot be calculated, because the F-measure value cannot be calculated to the baseline as in Perineural invasion. The largest improvement was present in N.P.C. Fibrosis and the results show an increase in the Recall value obtained for our method but the Precision value is the same. This result was common in almost all histopathology concepts, the Precision value do not increase or only slightly and the Recall value is increased.

Figure 1 illustrates the influence of α on the classification results. Specifically, it shows the F-measure average obtained with the model for the histopathology concepts. To combine the F-measure results of all these concepts, the values were weighted with the test size of each one of the concept classes.

In figure 2 the size of the latent semantic space is studied. This shows the average performance for the latent semantic space, the combined features space and the multi-features representation space with best α values obtained ($\alpha = 0.75$). In most *k* values evaluated, we see that the average performance of the multi-features representation space outperform the other models and the baseline, and the latent semantic space showed a lower performance, showing that this text document strategy cannot be directly applied in the histopathology image classification problem.

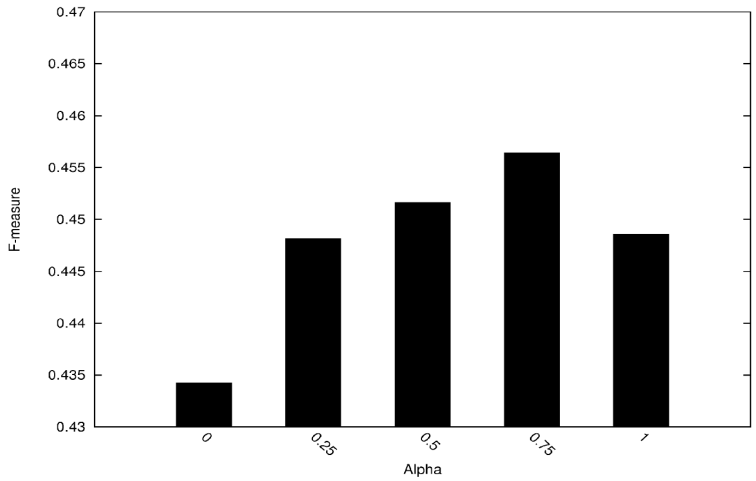

Figure 1. Average performance obtained for different values of the parameter.

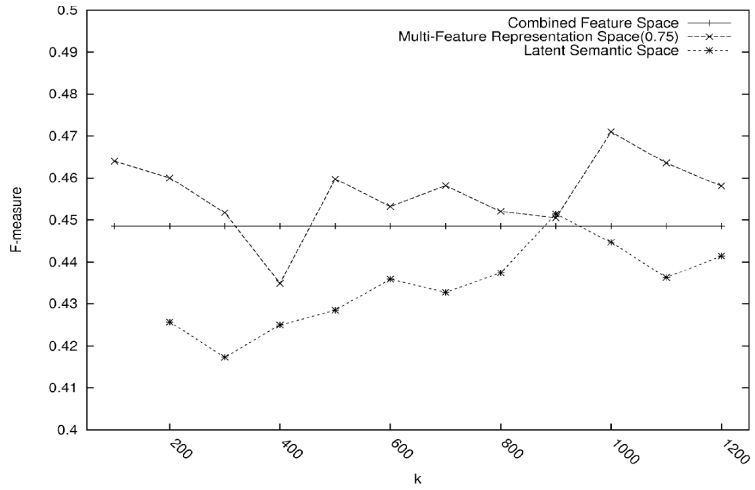

Figure 2. Average performance dependence on the semantic space size (parameter k) for , latent semantic space and combined feature space.

CONCLUSIONS

In this paper an approach for histopathology image automatic annotation was introduced, using a strategy based on the decomposition of the kernel matrix to exploit the latent relationship between different low-level visual features. This method builds a new image representation using a convex combination of two different spaces, the combined feature space and the latent semantic space. This generalizes the representation and makes possible to use only the original space, only the semantic space or the multi-feature representation space. The results show that this representation outperform the baseline method, the combined feature space and the latent semantic space, showing

that visual patterns found in the multi-feature representation space can improve the results obtained and help in the annotation task for almost all the concepts evaluated.

ACKNOWLEDGMENTS

To Universidad Nacional de Colombia.

BIBLIOGRAPHY

BARLA A, ODONE F, VERRI A. Histogram intersection kernel for image classification. Proc Int Conf Image Proc. 2003;2:513-516.

BARLA ANNALISA, *et al.* Image Kernels. Pattern Recognition with Support Vector Machines. 2002:617-628.

BERMAN AP, SHAPIRO LG. A flexible image database system for content-based retrieval. Int Conf Pattern Recognit. 1998;1:894-898.

CAICEDO J, CRUZ A, GONZALEZ F. Histopathology Image Classification Using Bag of Features and Kernel Functions. Artif Intell Med. 2009:126-135.

CAICEDO J, GONZALEZ F, ROMERO E. Content-Based Medical Image Retrieval Using Low-Level Visual Features and Modality Identification. Advances in Multilingual and Multimodal Information Retrieval. 2008:615-622.

CRISTIANINI N, SHAWE-TAYLOR J, LODHI H. Latent Semantic Kernels. J Intell Inf Syst. 2002:127-152.

DEERWESTER S, *et al.* Indexing by latent semantic analysis. J Am Soc Inf Sci Technol. 1999;6(41):391-407.

DESELAERS T, KEYSERS D, NEY H. Features for Image Retrieval: A Quantitative Comparison. Pattern Recognit. 2004:228-236.

LEE G, *et al.* A knowledge representation framework for integration, classification of multi-scale imaging and non-imaging data: Preliminary results in predicting prostate cancer recurrence by fusing mass spectrometry and histology. Proc IEEE Int Symp Biomed Imaging. 2009:77-80.

MONAY F, GATICA-PEREZ D. On image auto-annotation with latent space models. Proceedings of the eleventh ACM international conference on Multimedia. 2003:275-278.

MÜLLER HENNING, *et al.* A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. Int J Med Inform. 2004;1(73):1-23.

SMEULDERS AWM, *et al.* Content-Based Image Retrieval at the End of the Early Years. IEEE Trans Pattern Anal Mach Intell. 2000;12(22):1349-1380.

WONG CSM, STRANGE RC, LEAR JT. Basal cell carcinoma. Br Med J. 2003:794-798.