Content-based image retrieval system to support the diagnosis of human papillomavirus

Article *in* Health and Technology · August 2015

DOI: 10.1007/s12553-015-0114-2

CITATIONS
3

READS
58

3 authors:

Maritza Mera-Gaona
Universidad del Cauca
12 PUBLICATIONS **86** CITATIONS
SEE PROFILE

Sandra Milena Roa Martínez
Universidad del Cauca
18 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Carolina González
Secretaría De Educación Pública
111 PUBLICATIONS **1,126** CITATIONS
SEE PROFILE

Health Technol. (2015) 5:161–165
DOI 10.1007/s12553-015-0114-2

REVIEW PAPER

Content-based image retrieval system to support the diagnosis of human papillomavirus

Maritza Mera[1] · **Sandra Milena Roa**[1] · **Carolina González**[1]

Received: 10 March 2015 / Accepted: 3 August 2015 / Published online: 18 August 2015

Abstract This article describes the development of a Content- Based Image Retrieval system of digital images of cervix biopsy infected with Human Papillomavirus. Through a query image, the system performs a search in the database of digital images of cervix biopsies. The search makes possible to recover the diagnoses of the images stored in the database that might represent lesions similar to the query image. By using feature extraction techniques of low-level, a representation of the visual content of the images is calculated. This representation is used to compare the images by means of similarity measures. The system has been built as a helping tool for medical diagnoses; the purpose is to reuse the knowledge gained by carrying out diagnoses and treatments in order to diagnose new images. The evaluation of the system produced an average precision of 80 % for the retrieval of similar images stored in the database.

Keywords Content based image retrieval (CBIR) · Medical image · Human papillomavirus (HPV)

1 Introduction

In recent years, the Content-Based Image Retrieval (CBIR) system has emerged as a solution to process and manage large amounts of visual information [1]. Nonetheless, in the medical field the CBIR systems represent a supporting mechanism for the decision-making during the diagnoses of medical images [2]. By using a medical image, it is possible to consult the diagnoses that have been performed to images with similar lesions. This represents a valuable help for physicians when they face cases difficult to diagnose, or when they are faced with external factors such as the subjectivity, inexperience, tiredness, etc. that can lead to inaccurate diagnoses. Taking into consideration the high incidence and cervical cancer mortality rates reported around the world [3], the efforts to fight this illness has been focused on prevention and early detection programs, which consequently increases the probabilities of success of the treatments. Therefore, the physicians might use different techniques and tests to detect lesions causing the cancer.

In histopathological studies which are one of the main tools used to diagnose illnesses, physicians assess the tissues of the suspicious zones through biopsies in order to determine the presence or absence of premalignant lesions produced by the Human Papillomavirus (HPV). Nevertheless, the diagnoses of lesions and abnormalities in the tissues of the patients may be affected by the poor quality of the images analyzed, the presence of difficult-to-detect lesions, and the little experience of the doctors. Therefore, the presence of false negatives or false positives in the diagnoses jeopardizes the patients' health.

This study describes the implementation of a CBIR system that contributes to diagnose digital images of cervix biopsies. The system developed allows users to search diagnoses and treatments of images with similar lesions by using a query image. Employing low level features, the images are represented throughout numerical values that describe the visual content of the image. One of the main challenges of CBIR developments is the selection of low level features that describe the content of the image since their representation is defined according to their type, origin, and information

✉ Maritza Mera
maritzag@unicauca.edu.co

Sandra Milena Roa
smroa@unicauca.edu.co

Carolina González
cgonzals@unicauca.edu.co

[1] Computational Intelligent Research Group, University of Cauca, Cauca, Colombia

available. Taking into account that the digital images of cervix biopsies have information of tissues, a characterization based on the texture analysis over the image was designed [4].

2 Methods

2.1 Database of images

The database of images that has been built, stores digital images of a set of cervix biopsies that contain information of tissues affected by the HPV. The digital images are obtained by photographing the plates of the visualized biopsies by means of a Nikon Eclipse 80i microscope, with a resolution of 40x. The resolution selected to capture the photography was set in 40x because this is the resolution used by specialists for the analysis of the cellular structures present in the plate content.

The biopsies used for the construction of the images' database belong to the Company of Pathologists of Cauca (Compañía de Patólogos del Cauca) which is a clinical laboratory that granted the access to the physical plates of the cervix biopsies and their diagnoses. The plates of the biopsies were digitalized in the Laboratory of Electron Microscopy of the University of Cauca.

2.2 Feature extraction

In pursuance of a CBIR system, a process feature extraction of the images is required. Each digital image is represented by employing an n-dimensional vector that stores its visual content. The value n represents the number of features extracted from the image.

Taking into account the origin of the database images, their content, and the relationship between the analysis of the tissues over digital images and the texture analysis, the characterization process implemented is based on the calculation of Tamura textures, Discrete Wavelets Transform, and the Co-occurrence Matrix, as described in [4].

2.3 Similarity function

According to the feature vector that describe the image visual content, the similarity function compares each feature from one image against another in order to determine the difference of the images in relation to each feature. The sum of all the differences of the features determines the similarity between the two images.

2.4 K-nearest neighbors algorithm - KNN

The algorithm KNN allows the user to search from a query image the more similar K images found in the database [5]. For this, the algorithm uses the similarity function in order to compare between the query image and the database images. The k images more similar (closer) will be displayed in the search results.

3 Results

3.1 Construction of the CBIR system

The CBIR system was built by using the elements described in the section methods. The system supports the search scheme

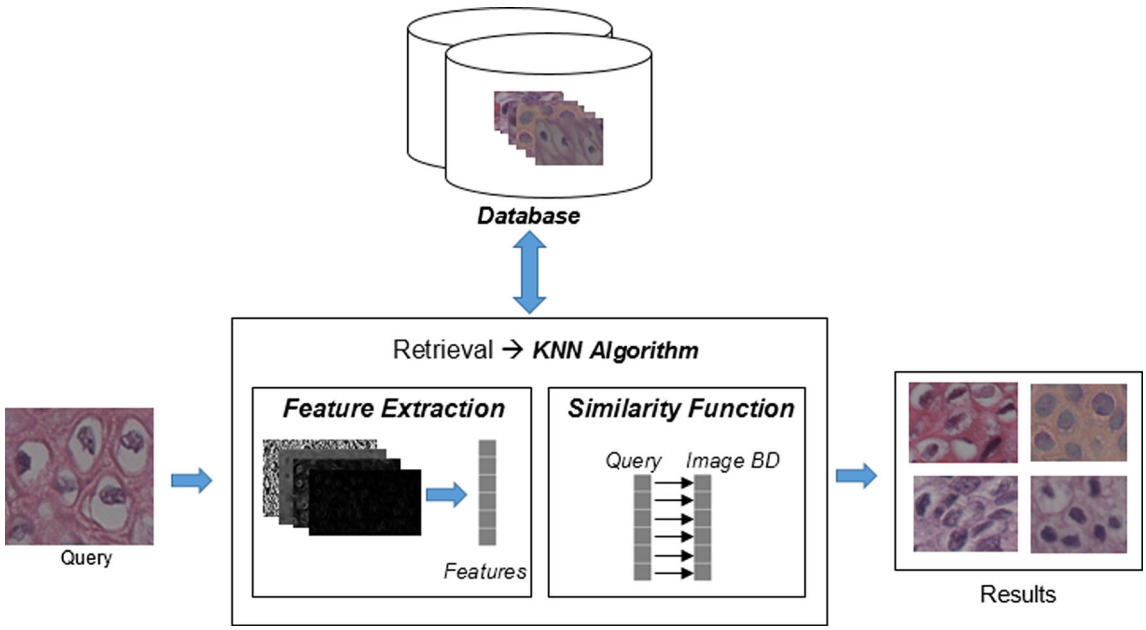

Fig. 1 Search scheme

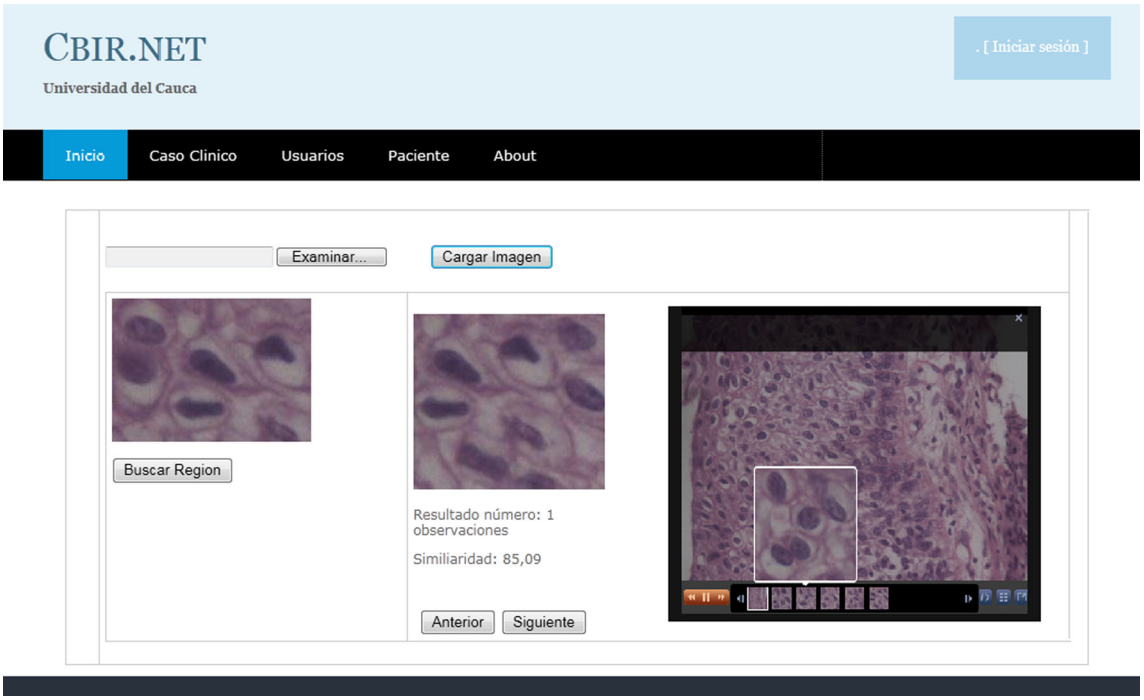

Fig. 2 Query interface

described in Fig. 1. By using a new image, the system retrieves from the database the k images that show the closest degree of similarity to the new image.

The algorithm KNN was used for the implementation of the retrieval module. This module calculates the visual features of the image searched, and compares them through the similarity function to the features of each database image. In line with the similarity results calculated between the searched image and the database images, the k images with a higher degree of similarity are selected in order to be displayed.

The implementation of the system was developed by using .NET framework and the library of artificial vision OpenCV. Throughout the .NET framework, a web application designed to support the search and management of digital images of cervix was developed, as shown in Fig. 2. The algorithms associated to the feature extraction of images were implemented in C++ language with the intention of making use of the speed of the language, and of facilitating the access to the library OpenCV functionalities. All the implementations carried out in C++ language were exported to a Dynamic-Link Library (DLL) with the purpose of integrating them into the web application built in C Sharp language. The architecture that describes the main components of the system is shown in Fig. 3.

The system architecture is presented in Fig. 3.

3.2 Evaluation of the CBIR system

A collection of 12 test images was selected in order to evaluate the average precision of the system recovery. For each test image there are five similar images in the database. The evaluation was carried out asking the system for each test image and then comparing the first five recovered images carried out by the system with the ideal results. The results of the evaluation were documented in Table 1 as a classification problem. Each test image as well as its ideal answer were

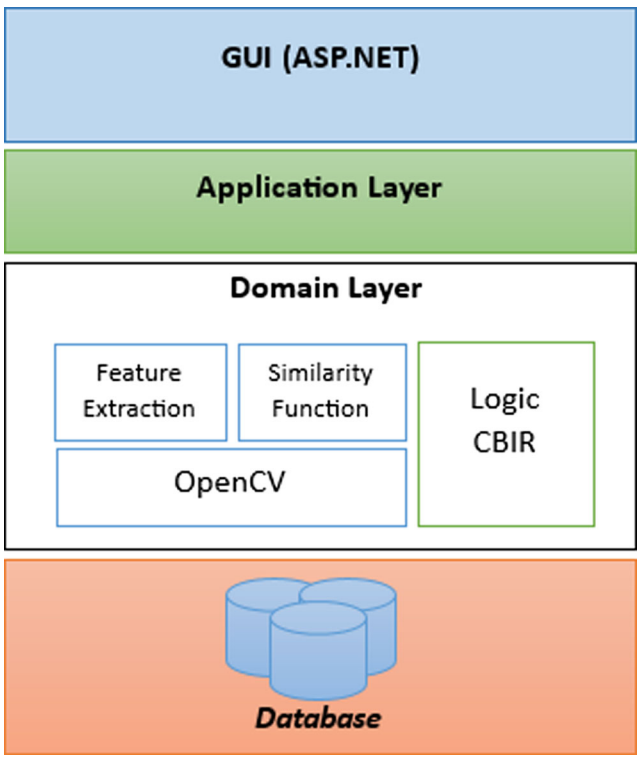

Fig. 3 System architecture

Table 1 Evaluation results

Class	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	%
C1	4	1											40 %
C2		2								1		2	40 %
C3			4						1				80 %
C4				4	1								80 %
C5					5								100 %
C6						5							100 %
C7						1	4						80 %
C8								4		1			80 %
C9									5				100 %
C10										4	1		80 %
C11										2	3		60 %
C12										1		4	80 %
												Average	80 %

considered as elements of the same type. The element $T_{i,j}$ of the table describes the quantity of elements of the class i that were classified by the system as elements of class j.

The results obtained in the evaluation carried out describe an average precision of 80 % for the retrieval. This result is acceptable considering that the system has as its main objective to suggest to physicians similar cases instead of automating the diagnosis of an image.

4 Discussion

Taking into consideration the high incidence and cervical cancer mortality rates [3] and the difficulty to diagnose the lesions causing the cancer produced by the Human Papillomavirus (HPV), a computational tool to support the diagnoses carried out by the physicians is required, so they can increase the reliability of the diagnoses. This will contribute to reduce the false negatives or false positives diagnosed to the patients.

Considering the literature review, different content based image retrieval systems have been developed to support the diagnoses of medical images resulting from radiology [6], hyspathological analysis [7], scans, tomography [8], among others [9]. However, the visual content of the medical images is different in regard of the origin and acquisition means. Thus, the systems mentioned above cannot be used to diagnose the HPV. The system developed represents a contribution to the development of the CBIR systems for the clinical field.

In addition to this, the use of the system developed in a clinical environment will contribute to decrease the time spent by the physician to diagnose the biopsy images since the system allows the user to search diagnoses and treatments carried out to images with similar lesions. Then, the reliability of the diagnoses could be improved taking into account that during the diagnosing process the physicians can reuse the experience of tested clinical cases already proved.

5 Conclusions

The difficulty to diagnose digital images of cervix biopsies is associated with a bad quality of the images, the subjectivity of the physicians, and the difficulty to detect lesions displayed in the images. This situation risks the reliability of the diagnoses carried out and consequently the patients' health.

The development of the CBIR system represents a significant contribution to diagnose lesions causing cervical cancer, making it possible to identify lesions difficult to diagnose. In order to evaluate the reliability of the system, a set of test images was used to request a search. The images retrieved in each search were evaluated on the bases of their similarity with the images under revision. The evaluation of the system produced an average precision of 80 % for the retrieval of similar images stored in the database. As a further study, we suggest to evaluate the use of the system in order to determine if the system can reduce the time spent to diagnose an image, and if it is possible to improve the quality of the diagnoses carried out in a clinical context.

Compliance with Ethical Standards All procedures performed with digital images of cervical biopsies were in accordance with the ethical standards of the Company of Pathologists of Cauca. In addition, the images used in this work do not associate information that allows to identify the patient or their origin.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Muller H. A review of content based image retrieval systems in medical applications clinical benefits and future directions. Int J Med Inform. 2004;73.

2. Muller H, Zhou X, Depeursinge A. Medical visual information retrieval: state of the art and challenges ahead. Multimed Expo IEEE Int Conf. 2007;683–686.
3. Jemal A, Bray F, Center M, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: Cancer J Clin. 2011;61:69–90.
4. Roa S, Mera M. Content-based retrieval of cervical microscopic images infected with human papillomavirus using textural features. Rev Ing Bioméd. 2013;7:69–80.
5. Kozma L. K Nearest Neighbors algorithm (kNN), Special Course in Computer and Information Science. Helsinki University of Technology; 2008.
6. Akgül C, Rubin D, Napel S, Beaulieu C, Greenspan H, Burak A. Content-based image retrieval in radiology: current status and future directions. J Digit Imaging. 2011;24:208–22.
7. Caicedo JC. A Prototype System to Archive and Retrieve Histopathology Images by Content. Bogotá: National University of Colombia; 2008.
8. Lam M, Disney T, Pham M, Raicu D, Susomboon R and Furst J. Content-based image retrieval for pulmonary computed tomography nodule images. Proc. SPIE 6516, Medical Imaging 2007: PACS and Imaging Informatics. 2007.
9. Müller H, Deserno T. Content-based medical image retrieval. Biol Med Phys Biomed Eng. 2011; 471–494.