Logo and text removal for medical image retrieval

Henning Müller, Joris Heuberger, Antoine Geissbuhler

University and University Hospitals of Geneva,
24, Rue Micheli–du–Crest, 1211 Geneva 14, Switzerland
henning.mueller@sim.hcuge.ch

Abstract. The amount of visual medical information being produced in large hospitals is exploding. Most University hospitals produce Millions of images per year (Geneva Radiology: 20'000 images per day). Currently, the access to these images is most often limited to an access by patient identification. Sometimes search by text in the radiology report or in the DICOM headers is possible. Still, all the implicit information potentially available through the image and the accompanying case report is discarded in this case. Content–based visual data access is based on direct visual properties of the images that are extracted automatically from all images in a database. This delivers objective features for searching images but the features are commonly on a very low semantic level (colour histograms, simple texture analysis such as wavelet filter responses). Another problem that especially occurs in medical teaching files but also in routine images is text and logos around the main object in the image. For retrieval this is mainly noise that can have a negative influence on retrieval quality. In our approach, we extract the main object from the image by removing logos that are added to the images as well as frames around the images and text fields or other elements that are not needed. This is mainly based on properties of the text that occurs on the images, and especially of the logo of the university hospitals of Geneva. Frames around the images are removed reliably. First results show that the retrieval quality can be augmented well with such an approach. Especially queries with relevance feedback deliver much better results as the query is more focused. Proper, quantitative evaluation on a large data set is still missing but will be performed shortly.

1 Introduction

Content–based image retrieval has been one of the most active research areas in computer vision over the past 15 years [1]. The goal is to be able to retrieve images based on a visual description, for example by submitting example images as queries. In the medical field, content–based data access has been proposed several times as an important tool to help in the increasingly visual diagnostic process [2,3,4]. Tools will be needed to use the rising amount of visual data up to its full potential. Several projects on medical image retrieval exist such as *IRMA*[1]

[1] `http://www.irma-project.org/`

[5] and *medGIFT*[2] [6]. Content–based retrieval has already been used successfully as a diagnostic aid [7] and is generally proposed in domains such as case–based reasoning or evidence–based medicine. Currently, most techniques rely on analysing the global image content when using varied PACS–like databases that can contain images of any modality and anatomic region as well as photographs [5,6]. Only in specialised domains, segmentation techniques can be used to extract the main image object reliably. Even more often, regions of interest are marked manually by the practitioner which is prohibitively expensive for large databases such as the 65,000–image teaching file that we are working on.

Similar algorithms for the detection of text are often used on video sequences to identify boxes of text in video and also to decipher the text shown on screen and in images to augment retrieval quality [8]. In this case, a character recognition often follows the step of text identification. In our case, we only want to remove text areas as much as possible, so accurate finding is not necessary.

2 Removal of logos, text, and other problems

We needed to develop a solution that runs completely automatic on the 65,000 images without any manual intervention. Main problems identified in the images of our teaching file are large regions around the object in the images that were mainly used by system parameters, scales and the logo of the university hospitals but also black frames as can be seen in Figure 1.

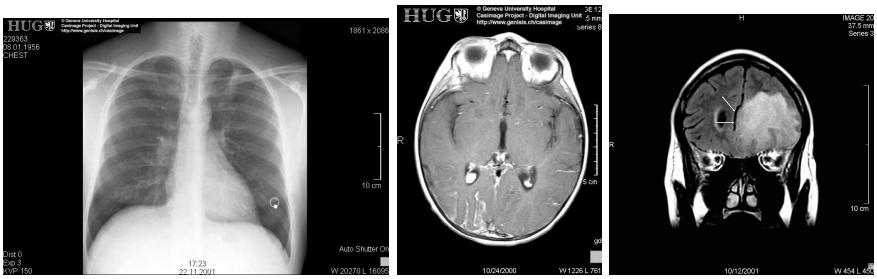

Fig. 1. Images before the removal of logos and text.

Our algorithm uses the properties of the text and logo part, which both contain several small, non–connected components and also the fact that these parts occur most often in certain parts of the image. Text and logos are also most often in white or a very light grey level. If a logo was detected in the upper region, the logo part was filled with black pixels. The removal of further structures such as text was done through smoothing followed by an edge detection, and

[2] `http://www.sim.hcuge.ch/medgift/`

thresholding to remove structures of low intensity. Then, parts smaller than a certain size (including most of the text) can be removed and a bounding box can be put around the remaining object for indexing and retrieval. Besides separately treating the logo we also had to treat grey squares separately that occur in the bottom right part of several images. Otherwise, these structures were too large to be removed by the automatic algorithm.

This delivered very good results in a set of 500 images that we controlled manually. 204 of the 500 had significant parts around the objects removed accurately. In 185 images, no removal was necessary. This means that in 80% of the cases the result was satisfying. In 105 images, not everything but only part was removed. This is often due to large structure such as letters indicating left and right on x–rays. Still, these images were not worse than beforehand, with respect to indexing but often better. In only 6 images, too much was removed, which is below 1%, and in general nothing of importance was removed. The only problem occurs on chest x–rays where part of the bottom was removed, which did actually not contain any information relevant for the diagnosis. No other images had significant parts being removed by error. This low rate of erroneous removal was one of the important goals of the project. The errors are due to the very slow changes in these images in the bottom area. For the diagnosis, these problems do not make any difference as the bottom part does in this case not contain any relevant information. For all the simple image modifications we use the Insight toolkit (itk[3]) performing the following steps:

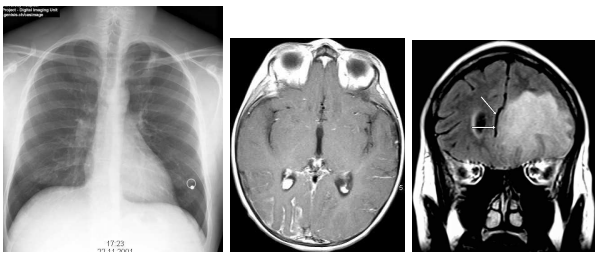

Fig. 2. Images after the removal of logos and text.

- removal of grey squares in the bottom right part;
- detection and removal of university logo in the upper part (thresholding, white pixel count before and after an erosion, filling of region with black if value in a certain range);
- smoothing with a median filter;
- edge detection, thresholding, dilatation, removal of small objects, erosion;
- bounding box and cropping of the background.

[3] `http://www.itk.org/`

The results of this process can be seen in Figure 2 showing the same as Figure 1.

3 Influence on the retrieval quality

When removing logos and text fields from the images before indexation for retrieval, we can focus the search by image content much more on the important structures for retrieval. Background information might in some case lead to seemingly good results because images of the exactly same machine have very similar background information. Still, this is not the part the we would like to retrieve visually and is rather retrieval by chance.

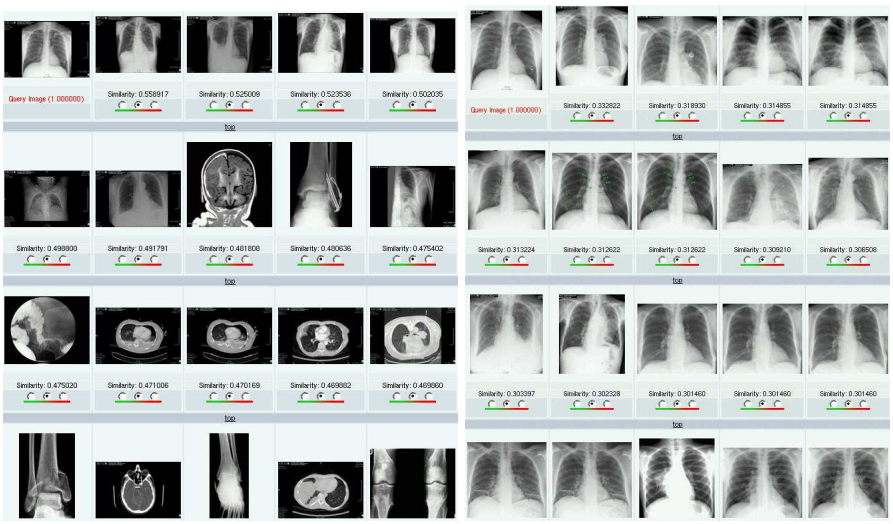

Fig. 3. Retrieval with an image before and after logo and text removal.

Figure 3 shows an example query with the original image and the treated image. We can see that the original leads to five good retrieval results, which have the same black background frame. The treated image leads to much better retrieval results as all retrieved 20 images are chest x–rays. For the retrieval of the images we use medGIFT that is based on the GNU Image Finding Tool (GIFT[4]). The database that we indexed is that of the imageCLEF[5] image retrieval competition.

[4] `http://www.gnu.org/software/gift/`
[5] `http://ir.shef.ac.uk/imageclef2004/`

4 Conclusion

Image retrieval on varied databases of unrestricted PACS data or medical teaching files, which are too varied for general object segmentation, can profit from an image pre–treatment such as background removal when problems of noise in the background are identified. Manual intervention is prohibitively expensive when analysing thousands of images so a fully automatic algorithm needs to be employed. Our algorithm is simple and robust. Not all texts and logos are entirely removed but we only have a very small number of images where too much of the image was removed, which is important. The retrieval results were rarely degraded but delivered generally much better results. Especially when employing relevance feedback, the retrieval quality became better. A quantitative analysis of retrieval results is still needed to identify images where retrieval quality became better and task where the quality degraded. We are also working on optimising the algorithm to improve quality of removal for the images with remaining text parts. The entire software and the image database are available as open source (itk, GIFT) so results can easily be reproduced.

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content–based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22 No 12** (2000) 1349–1380
2. Lowe, H.J., Antipov, I., Hersh, W., Arnott Smith, C.: Towards knowledge–based retrieval of medical images. the role of semantic indexing, image content representation and knowledge–based retrieval. In: Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA), Nashville, TN, USA (1998) 882–886
3. Tagare, H.D., Jaffe, C., Duncan, J.: Medical image databases: A content–based retrieval approach. Journal of the American Medical Informatics Association **4** (1997) 184–198
4. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content–based image retrieval systems in medicine – clinical benefits and future directions. International Journal of Medical Informatics **73** (2004) 1–23
5. Lehmann, T., Güld, M.O., Thies, C., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., Wein, B.B.: Content–based image retrieval in medical applications. Methods of Information in Medicine **43** (2004) 354–361
6. Müller, H., Rosset, A., Vallée, J.P., Geissbuhler, A.: Integrating content–based visual access methods into a medical case database. In: Proceedings of the Medical Informatics Europe Conference (MIE 2003), St. Malo, France (2003)
7. Aisen, A.M., Broderick, L.S., Winer-Muram, H., Brodley, C.E., Kak, A.C., Pavlopoulou, C., Dy, J., Shyu, C.R., Marchiori, A.: Automated storage and retrieval of thin–section CT images to assist diagnosis: System description and preliminary assessment. **228** (2003) 265–270
8. Chen, D., Odobez, J.M., Thiran, J.P.: A localization/verification scheme for finding text in images and video frames based on contrats independent features and machine learning methods. Signal Processing: Image Communication **19** (2004) 205–217