Crowdsourcing opportunities in medical imaging

Antonio Foncubierta-Rodríguez and Henning Müller
University of Applied Sciences Western Switzerland (HES-SO)
{antonio.foncubierta, henning.mueller}@hevs.ch

1. Introduction

The production of medical images in clinical practice has been growing quickly over the past decades [1] [2]. Medical images amount for 30% of the global storage in 2010 according to [3] and mammographies in the US alone to over 2.5 Petabytes in 2009.

In addition to medical images stored in clinical Picture Archiving and Communication Systems (PACS), medical images are frequently used in the biomedical literature, carrying much information in connection with the associated text.

Despite the valuable amount of information stored, images are seldom used more than once in clinical practice and image content is very rarely analyzed to facilitate reuse. Improving accessibility to medical images both in clinical and research environments can provide clinicians, trainees and researchers with additional and valuable tools in their daily work.

Computer assisted indexing, classification and retrieval can successfully improve the accessibility of relevant images from the medical literature and reuse of clinical images for clinical decision support. However, these tools often require large training sets to deliver a convincing performance. For optimal, generalizable results, the size of these sets needs to be large and representative of the actual class distribution found in real-world data. Ground truth generation of large datasets is a tedious, repetitive task that is costly and time-consuming, and might require specialists to perform the annotation.

Crowdsourcing has received much attention in the past years, and has been proposed several times to solve challenging problems by using the so-called wisdom of the crowd [4].

In this paper, a medical image crowdsourcing-based ground truth generation method that reduces the manual interaction as much as possible is discussed, together with several ideas for crowdsourcing in medical imaging.

2. Crowdsourcing-based ground truth generation multimedia analysis

Although crowdsourcing has previously been used for obtaining definite solutions to challenging scientific problems [4], it has also been proposed as a way of obtaining reliable, comprehensive ground truth in a time-effective way with limited costs.

One of the first examples of ground truth generation was the ESP Game (Edd Smith's People), developed by von Ahn [5] and later acquired by Google to improve their image search engine. The game consisted of presenting two players with an image and requiring them to assign labels or tags to it. If the two players agreed on a label, the label would be assigned to the image. The same author developed later a game-based crowdsourcing platform, called Game With A Purpose (GWAP[1]), that proposed other image-related games, including classification, manual segmentation, labeling, etc. The term GWAP has also become popular in the game community [6].

Similar games have also been deployed, such as LabelMe[2] [7], which makes it possible to use a labeling and segmentation framework on user-uploaded datasets within crowdsourcing platforms such as Amazon Mechanical Turk, as shown in Figure 1.

Figure 1 LabelMe interface for image labeling and annotation.

When no game-like application is developed, crowdsourcing platforms like Amazon Mechanical Turk, Crowdflower, Zoombucks, etc. offer contributors the possibility to earn money for their work. The amounts paid are decided by the task creators according to the amount of units produced, correct units produced or even bonuses for extraordinary work (in terms of quality or quantity).

With Amazon Mechanical Turk being the most popular platform, most of the ground-truth work is based on their infrastructure [8]. However, the requirement of a US-based credit card might be blocking some research groups from using crowdsourcing.

When the tasks are better suited for a group of people with specific knowledge, crowdsourcing contributors can be filtered-out within a selection phase to control

[1] http://www.gwap.com/
[2] http://labelme.csail.mit.edu/

high quality of responses before the actual task begins.

3. Opportunities of crowdsourcing for the medical image domain

Although medical images are an important area of interest for the information retrieval community, there has been little effort in using crowdsourcing to develop ground truth for medical-related datasets. Special knowledge and the requirement to get high quality annotations also make crowdsourcing difficult in this domain.

In [9] we propose that ground truth generation can be quickly performed by non-experts after an initial training phase and with good explications. The study consisted of classifying a dataset of 3415 images of the biomedical literature into a modality hierarchy of 38 categories [10] as shown in Figure 2. The intent of classifying these images was to use them as ground truth for the medical task in the ImageCLEF [3] benchmarking event. In order to evaluate the quality and speed of the annotations, three user groups where compared: a) a medical doctor manually classified all 3415 images, providing a gold standard for the crowdsourcing experiments, b) a set of 18 known experts with experience in medical imaging classified the dataset and c) 2470 contributors from open crowdsourcing. The study also included an iterative approach, where images from a smaller training set were used to train an automatic system that was later evaluated using crowdsourcing, asking the contributors to accept or reject the automatic classification.

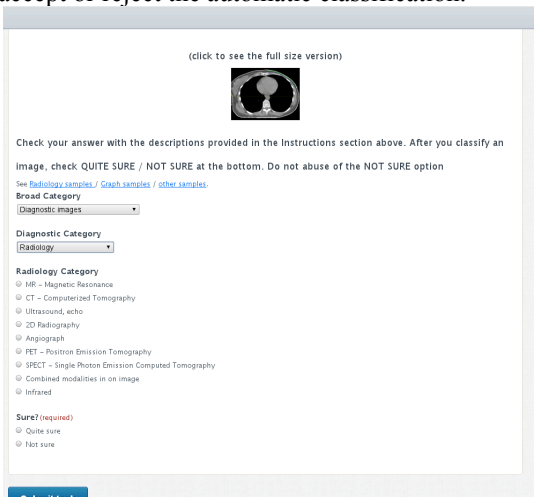

Figure 2 Screenshot of the crowdsourcing interface for modality classification described in [9].

The study concluded that non-experts can quickly build the ground truth at a limited cost of quality, which strongly varies among categories when comparing the annotations of the crowd to those from a set of known experts. The study resulted in several outcomes

- strict quality control and thus a good gold standard is necessary to obtain meaningful results (in our case 50% of the proposed images had a known ground truth and judges below 80% quality were automatically removed);
-very good explications and a tutorial are necessary to make the experiment work. Users need to know precisely what is expected from them;
- to increase the speed of annotations, many users are required, but users that contribute more to the system become more familiar and therefore provide better annotations, so frequent users need to be favored;
- complex tasks still take much time and demotivate users; simplifying tasks produces better (higher inter-rater agreement) and faster results (yes/no questions were answered twice as fast as a simple classification);
- a multi-step approach where system output can be validated or refused by *crowdsourcers* could make it scalable to big data.

One of the advantages of crowdsourcing is that it allows a very large number of users to annotate data (in our case more than 2400 persons tested the system in two weeks). For instance, image retrieval often relies on the concept of visual similarity, which is related to many aspects and can be defined in various ways (subjectiveness). A large-scale experiment with *crowdsourcers* allows inferring a visual similarity model from several users' understanding and creating a solid ground truth for visual similarity evaluation.

Other initiatives like the VISCERAL project[4] propose the creation of a silver corpus, where the results with high variance can be further analyzed manually, potentially using crowd sourcing or specialist annotators.

4. Conclusion

Crowdsourcing is a promising tool for scientific research, specifically in multimedia environments. However, there is the risk of abandoning machine learning techniques in favor of crowdsourcing in some cases. This should be carefully analyzed in order to obtain the best from both worlds.

Medical imaging has been underusing crowdsourcing techniques, but this might change in the future when models for strict quality control and high-quality tutorials for the tasks are delivering convincing results.

It is difficult to predict the background of the contributors, as well as the outcome of their work in crowdsourcing environments. There has been much effort in getting rid of untrusted or *incorrect* annotations, however, rather than blocking their participation in the system, stronger methods to learn

[3] http://www.imageclef.org/

[4] http://www.visceral.eu/

from them need to be defined, because a consistent trend of wrongly classified images from a certain group of contributors is showing a pattern of how things can be perceived, and can be the key to redefining concepts that have been understood as static for a long time.
Crowdsoursing can become an important tool in many domains but it is not a one shot thing but an iterative process. Analyzing first results, leaning from them and then redefining tasks are important. Quality control is equally essential for obtain good outcomes to develop many of tomorrow computer-based tools based on human intelligence.

References

[1] K. P. Andriole, J. M. Wolfe and R. Khorasani, "Optimizing Analysis, Visualization and Navigation of Large Image Data Sets: One 5000-Section CT Scan can ruin your whole day," *Radiology,* vol. 259, no. 2, pp. 346-362, 2011.

[2] H. D. Tagare, C. Jaffe and J. Duncan, "Medical Image Databases: A Content--Based Retrieval Approach," *JAMIA,* vol. 4, no. 3, pp. 184-198, 1997.

[3] "Riding the wave: How Europe can gain from the rising tide of scientific data," Submission to the European Comission. 2010. [Online]. Available: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

[4] F. Khatib, F. DiMaio, F. C. Group, F. V. C. Group, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, M. Jaskolski and D. Baker, "Crystal structure of a monomeric retroviral protease solved by protein folding game players," *Nature Structural & Molecular Biology,* no. 18, pp. 1175-1177, 2011.

[5] L. von Ahn and L. Dabbish, "Labelling images with a computer game," in *SIGCHI conference on Human factors in Computing Systems (CHI'04)*, New York, USA, 2004.

[6] B. Steinmayr, C. Wieser, F. Kneißl and F. Bry, "Karido: A GWAP for Telling Artworks Apart," in *The 16th International Conference on Computer Games (CGAMES2011)*, Kentucky, USA, 2011.

[7] B. C. Russel, A. Torralba, K. P. Murphy and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International Journal of Computer Vision,* vol. 77, no. 1-3, pp. 157-173, 2008.

[8] O. Alonso and S. Mizzaro, "Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment," in *SIGIR 2009 Workshop on The Future of IR Evaluation*, Boston, USA, 2009.

[9] A. Foncubierta-Rodríguez and H. Müller, "Ground Truth Generation in Medical Imaging: A Crowdsourcing-based Iterative Approach," in *ACM multimedia 2012 workshop on Crowdsourcing for multimedia (CrowdMM'12)*, Nara, Japan, 2012.

[10] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman and S. Antani, "Creating a classification of image types in the medical literature for visual categorization," in *SPIE Medical Imaging*, San Diego, USA, 2012.

Antonio Foncubierta-Rodríguez received the M.Eng. degree in telecommunication engineering at the University of Seville, Spain in 2009. In 2007, he worked part-time as a researcher for the Department of Communications and Signal Processing in the University of Seville. His research was related to video compression and transmission over mobile networks, leading to a master's thesis. In 2008 he worked in a project on medical image retrieval for the University Hospitals Virgen del Rocío in Seville.
Currently, as a PhD Student at the University of Geneva, he is a research assistant at University of Applied Sciences Western Switzerland in Sierre, where he works on several Swiss national and EU projects.

Henning Müller studied medical informatics at the University of Heidelberg from 1992–1997. After a diploma in telemedicine he worked at Daimler-Benz research and technology North America in Portland, OR. In 2002, he received the Ph.D. degree on content-based image retrieval at the University of Geneva with a research stay at Monash University in Melbourne, Australia, in 2001. Since 2002 he has been working at the medical faculty of the University of Geneva and since 2007 he has been professor at the HES-SO. Henning is currently coordinator of the EU-project Khresmoi and scientific coordinator of the VISCERAL project. He has published over 300 research articles and organized the benchmarking event ImageCLEF for ten years.