Using Crowdsourcing for Multi–label Biomedical Compound Figure Annotation

Alba G. Seco de Herrera[1], Roger Schaer[2], Sameer Antani[1], Henning Müller[2]

[1] Lister Hill National Center for Biomedical Communications,
National Library of Medicine,Bethesda, USA;
[2] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland;
Email: albagarcia@nih.gov

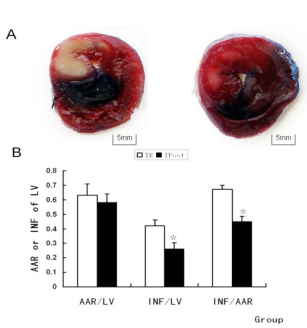
Abstract. Information analysis or retrieval for images in the biomedical literature needs to deal with a large amount of compound figures (figures containing several subfigures), as they constitute probably more than half of all images in repositories such as PubMed Central, which was the data set used for the task. The ImageCLEFmed benchmark proposed among other tasks in 2015 and 2016 a multi–label classification task, which aims at evaluating the automatic classification of figures into 30 image types. This task was based on compound figures and thus the figures were distributed to participants as compound figures but also in a separated form. Therefore, the generation of a gold standard was required, so that algorithms of participants can be evaluated and compared. This work presents the process carried out to generate the multi–labels of $\sim$ 2650 compound figures using a crowdsourcing approach. Automatic algorithms to separate compound figures into subfigures were used and the results were then validated or corrected via crowdsourcing. The image types (MR, CT, X–ray, ...) were also annotated by crowdsourcing including detailed quality control. Quality control is necessary to insure quality of the annotated data as much as possible. $\sim$ 625 hours were invested with a cost of $\sim$ 870$.

Keywords: Multi–label annotation, compound figures, crowdsourcing

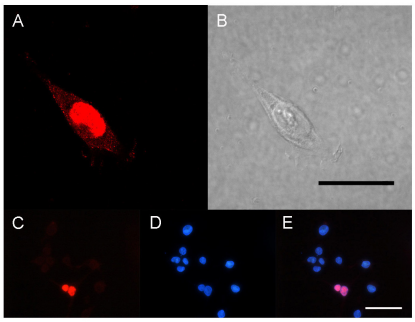
1 Introduction

Probably more than 50% of the figures in the biomedical literature in PubMed Central (PMC)[3] are compound figures (figures consisting of several subfigures) based on estimations of analysing a subset of the data [11]. In total, PMC in 2016 contains over 4 million images, so the extent of the knowledge stored in compound figures is important. A few simple examples of compound figures are shown in Figure 1 but not all images are as easy to separate. Information indexing and information retrieval (IR) systems for images should be capable of distinguishing the parts of compound figures that are relevant to a given query to deliver focused retrieval results. Identifying the image types of subfigures can

[3] http://www.ncbi. nlm.nih.gov/pmc/



(a) Mixed modalities in a single figure with 3 subfigures labeled as 2 (A and B).



(b) Mixed modalities in a single figure with no visual gaps between most subfigures.

Fig. 1. Examples of compound figures in the biomedical literature.

help to characterize compound figures, either by using the subfigures separately or the entire compound figure. In addition, image modality is an important piece of information that can be integrated into any retrieval system to enhance or filter its results [12, 17]. Therefore, the ImageCLEFmed[4] image classification and retrieval benchmark proposed in 2015 and 2016 a multi–label task aiming at labeling all compound figures with each of the modalities of the subfigures contained without knowing the subfigure separations that are contained in the image [10, 11]. It provides a useful scenario to compare the effectiveness of systems to access the detailed content of compound figures. This article presents the work carried out to generate a high quality ground truth for the evaluations in the task.

Image sharing sites like Flickr[5] offer a large number of images often with several tags describing the images added by the user, even though the quality can vary. Sometimes the content of the images is described but sometimes also what the image is about or what the image evokes, for example in terms of feelings. Some studies [14, 15] have shown the great potential of crowdsourcing in the context of medical imaging. However, in the medical open access literature almost no meta-data exist for figures and subfigures besides the free text captions.

Work has been done for multi–label annotation in the past. In NUS-WIDE [4], a small set of images from Flickr is manually annotated with 81 concepts. Wang et al. [18] encode each image into a vector and then a sparse label coding based on subspaces is applied to harness multi–label information. Nowak et al. [16] assessed ground truth of 99 multi–label images by using experts and mechanical turk. However, to the best of our knowledge, no previous work deals with

[4] http://imageclef.org/
[5] https://www.flickr.com/

multi–label annotation of compound figures or similar images from the medical literature.

This paper presents the methodology followed to annotated the collection created for the 2016 ImageCLEFmed task. The remainder of the article is organized as follows. Section 2 describes the database and methods used. Results obtained are presented in Section 3. The article concludes in Section 4.

2 Methods

This section describes the methods used to multi–label compound figures. The Crowdflower[6] platform was used for the crowdsourcing [5].

2.1 Dataset

The database used is a subset of 231,000 images from PMC that contained over 4,200,000 images in 2016. Figure 2 shows that hierarchy of images classes that was used [10, 11] to classify all subfigures into types.

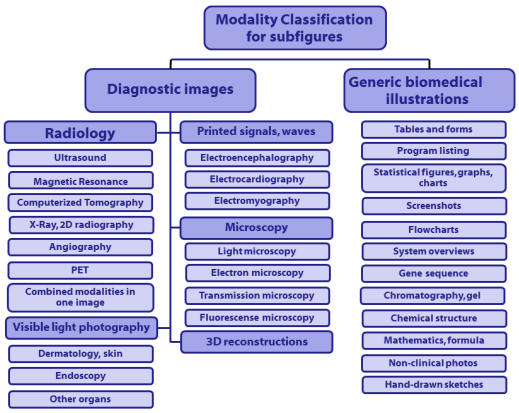


Fig. 2. The image class hierarchy proposed by ImageFmore inmed.

2.2 Overview

To simplify the evaluation of the multi-label annotation of compound figures and optimize the knowledge gained, the task was divided into several subtasks. The following tasks were carried out to evaluate all steps of the process of analysing content in compound figures:

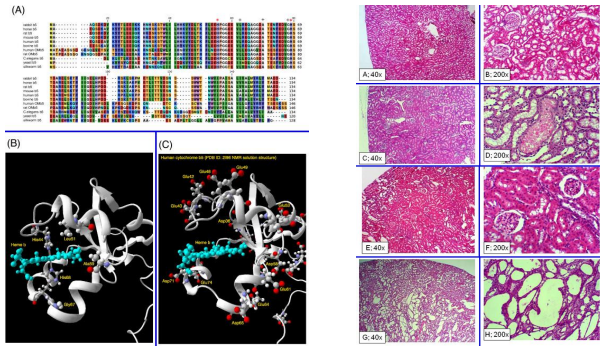
[6] http://www.crowdflower.com/

1. automatic compound figure detection (decide whether a figure is a compound or non–compound figure);
2. automatic compound figure separation (find the lines that cut compound figures into their parts);
3. manual compound figure separation verification (check whether images were correctly separated);
4. automatic subfigure classification (automatic determination of the type of image in a subfigure);
5. manual subfigure classification verification (validate the results of the previous step);
6. manual subfigure classification (manually classify the images incorrectly classified automatically);
7. manual class balancing (assure that all classes are represented);
8. compound figure multi–label assignment.

Details on each of the steps are given below.

Automatic Compound Figure Detection The procedure described in [6] was used to automatically classify the figures into image types including a 'compound or multipane figure' class. Figures classified as 'compound or multipane figure' were then randomly selected for the next steps in the classification to be able to take as many figures as possible into account.

Automatic Compound Figure Separation Compound figures were automatically separated into subfigures using the approach proposed by Chhatkuli et al. [3]. Figure 3 shows two compound figures automatically separated into subfigures using this approach. However, not all selected compound figures were correctly

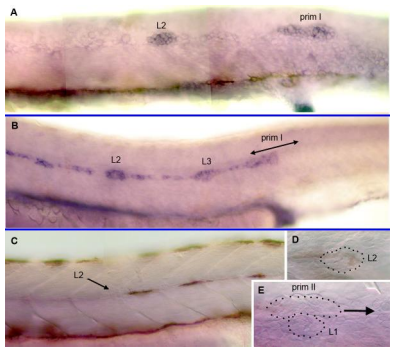


(a) Compound figure containing 3 subfigures.

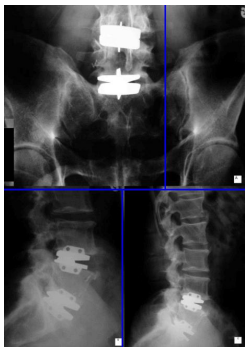
(b) Compound figure containing 8 subfigures.

Fig. 3. Examples of compound figures correctly separated into subfigures automatically. The blue lines show the detected separators.

separated into subfigures (see Figure 4 for examples that were incorrectly separated). Both missing lines occured and additional lines within single subfigures. Therefore, a verification step was implemented to identify incorrect separations and then correct them.



(a) Compound figure containing 5 subfigures but separated only into 3 subfigures.



(b) Compound figure containing 3 subfigures but separated into 4 subfigures.

Fig. 4. Examples of compound figures incorrectly separated into subfigures automatically.

Manual Compound Figure Separation Verification In this step, a crowdsourcing task was run where the following simple question was proposed:

- Is the compound figure correctly separated?:
  - Yes;
  - No.

The figures marked as correctly separated were used for the following step. Incorrectly separated figures were manually separated in a subsequent step.

Automatic Subfigure Classification The subfigures obtained using the automatic separation from the previous step were automatically classified into image types using an approach based on k–Nearest Neighbors (k–NN) and multiple visual features (see García et al. [8, 9]). On a past database a good performance of 68% was obtained for the same task.

Manual Subfigure Classification Validation Similar to [6] a figure classification validation step was carried out to assure the data quality. In this case the subfigures were presented together with the automatically labeled class in a crowdsourcing task. The question asked to the contributors was the following:

- Does the figure correspond to the stated category?:

- Yes, perfect classification;
- No, wrong category;
- Not sure.

Manual Subfigure Classification One last crowdsourcing task was created to classify the figures not marked as correctly classified in the previous step. This task was slower than the previous steps. Contributors were asked to classify each of the images according to the full hierarchy shown in Figure 2. A hierarchy was proposed in the interface to simplify the task (see Figure 5), so more than one click was necessary for the classification, with three levels for diagnostic images and two levels for general illustrations. In a similar task in 2015 we realized for crowdsourcing the contributors used the categories requiring few clicks much more often, which led to changes in the setup. As crowdsourcing pays per annotated image there is a risk to have people use the fastest way to categorize if there are differences. Thus the structure was slightly changed to have the same number of clicks for each of the classes in 2016, which avoided this bias.

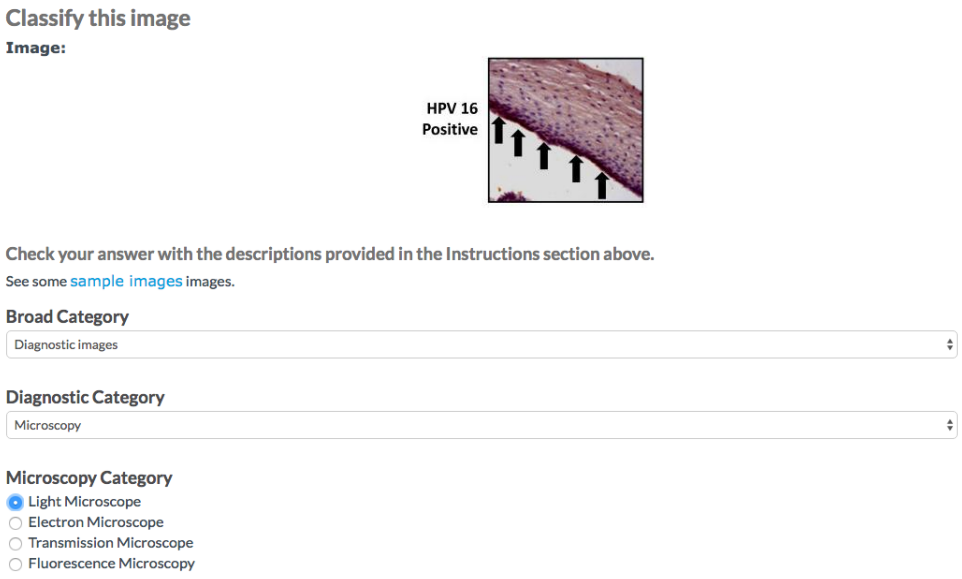


Fig. 5. Screenshot of a crowdsourcing task that aims at classifying biomedical figures from the literature into image types.

Manual Class Balancing After the previous step several of the classes were not represented or contained only very few images. Therefore, compound figures

containing the image types that were underrepresented were manually selected from the database to better represent these classes.

Compound Figure Multi–label Assignment To finalize the annotation process, each compound figure was assigned with the labels of all subfigures that it contains. Like this we can validate not only images that separate and then classify subfigures but also multi–class labeling based on entire figures.

2.3 Crowdsourcing Quality Control

A quality control (QC) is needed when using crowdsourcing to ensure the success of the annotation task [13], particularly with medical images where some domain knowledge is very benefitial [1]. QC approches were applied during design–time and runtime [2].

First, tasks were designed to be as simple as possible to make sure the persons understand the tasks quickly and correctly. This is the reason to divide the process into several subtasks. Automatic steps were added to limit the manual tasks where possible and reducing the number of figures to be manually classified since this is the most challenging step of the process. A detailed and unambiguous description of the tasks was provided to the participants and in case of doubt the participants could access this description at any moment. In particular, the description included several figure examples of each case or modality. In addition, Crowdflower provides feedback from several experts on the task design. Contributors were limited to the internal team of biomedical imaging experts or contributors with specified reputation level to optmize the quality.

For runtime QC, the following tools provided by Crowdflower were used:

- Output agreement: two contributors had to independently provide the same result to consider an answer as correct.
- Control with known ground truth: tasks of the same type with known answers are proposed at the beginning and randomly during the job execution to check the quality of the answers of each contributor. A 70% accuracy was the minimum required to be maintained throughout the job as Crowdflower suggests; a few images can be subjective and could be added to more than one class and for this reason the threshold was not stricter.
- Monitor answer patterns: specific answer such as 'not sure' or 'other' were monitored; 17% was the acceptable range of answers like "Not sure" or "Other" and otherwise a contributor was removed.

Allahbakhsh et al. [2] propose that domain experts check the contribution quality. Therefore, to finalize the quality control, an expert review was carried out. An expert in biomedical imaging manually checked the contributions quickly to ensure the high quality of the annotations.

3 Results

This section describes the results obtained in the data classification and annotation steps described in Section 2.

15,403 compound figures were initially selected and automatically separated from the ImageCLEFmed 2013 database [7]. After the compound figure separation step, $\sim$ 57% of the figures were correctly separated based on a manual validation. This task was carried out using the free internal Crowdflower interface that can be used for known set of people. Eight experts in biomedical imaging verified the separation of the figures in $\sim$ 98 hours. A subset was selected to be separated into subfigures and the subfigures were automatically classified. In the subfigure classification validation process $\sim$ 56% were defined as correctly classified into the correct figure type. More than 100 contributors validated the classification in $\sim$ 49 hours with a cost of 396.68$. The incorrectly classified subfigures were manually classified into exact figure types via crowdsourcing. To evaluate the correct design of the task, the first 1,149 subfigures were classified using the internal interface by 5 experts in $\sim$ 5 hours. Then, the remaining subfigures ($\sim$ 9800) were classified by more than 100 contributors in 427 hours with a cost of 472.66$. After this process, a manual expert review was needed to solve subfigure classification mistakes.

As the final selection of subfigures did not contain all figure types and was very unbalanced it was decided to manually add additional compound figures that contain relatively rare subfigure types. 122 compound figures containing the following categories were added and then manually separated and classified: angiography; computerized tomography; magnetic resonance; ultrasound; electroencephalography; mathematics program listing; and combined modalities in one image. Even with this balancing step, the class distribution remains uneven, as it is in the biomedical literature, even though it was slightly more balanced.

In total, 2,651 compound figures were annotated with multiple labels of their subfigures, containing 8,397 subfigures. These figures were distributed for the ImageCLEFmed 2016 multi–label and subfigure classification tasks[7] [11] together with the figure captions. In 2015, 1,568 were distributed for the ImageCLEFmed multi–label task [10]. These figures were distributed as a training set (containing 1,071 figures) and a test set (containing 497 figures). Their subfigures were released for the ImageCLEFmed 2015 subfigure classification task. The training set contained 4,532 subfigures and the test set 2,244 subfigures. In 2016, ImageCLEFmed used all the figures distributed in 2015 as training set and the additional annotated figures were distributed as test set. As a result, 1,568 figures were provided as training set and 1,083 as test set in the ImageCLEFmed 2016 multi–label tasks. The ImageCLEFmed 2016 subfigure classification task contained 6,776 subfigures in the training set and 4,166 subfigures in the test set.

In 2016, ImageCLEFmed proposed 5 tasks: compound figure detection; compound figure separation; multi–label classification; subfigure classification and caption prediction. This work describes the generation of the data for the multi–label classification task and therefore the subfigure classification tasks. The ImageCLEFmed multi–label classification task aims at labeling each compound figure with each of the modalities (see Figure 2) of the subfigures contained with-

[7] http://imageclef.org/2016/medical/

out knowing where the separation lines are. Furthermore, the ImageCLEFmed subfigure classification aims at classifying figures into the 30 image types of the proposed hierarchy.

Research groups could participate in these tasks and compare their research tools with those of other researchers on the same data and the same evaluation scenario. Four groups submitted 15 runs to the ImageCLEFmed multi–label task and ten groups submitted 45 runs to the ImageCLEFmed subfigure classification task. More information can be found in the working notes of CLEF 2016 [11].

4 Conclusions

This article presents the steps used to annotate compound figures from the biomedical literature with figure type information and to separate compound figures with separation lines to cut them into all subfigures. As a result 2,651 compound figures were annotated with figure type information and all figures were made available for the ImageCLEFmed 2016 multi–label task. To ensure the quality of the annotation, the process was divided into multiple steps combining automatic tools (e.g. for figure separation and figure modality classification) and manual work to validate or label data. Crowdsourcing was used to accelerate the tasks with a limited cost. Therefore, it was very important to carry out QC. Thanks to the described process it was possible to annotate the figures automatically and thus limit the manual control to verify and correct the annotations. The created resources are now available for the medical image analysis and image retrieval community. This is a manually created gold standard to build tools to create more metadata for the over four million figures in PMC and the likely over 2 million compound figures containing an estimated 6–7 million additional subfigures. Providing detailed metadata for these figures can well help to make the knowledge contained in the figures accessible for research and clinical work.

Acknowledgments This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

References

1. Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N.: Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Transactions on Medical Imaging 35(5), 1313–1321 (2016)
2. Allahbakhsh, M., Benatallah, B., Ignjatovic, A., MotahariNezhad, H.R., Bertino, E., Dustdar, S.: Quality control in crowdsourcing systems: Issues and directions. IEEE Internet Computing (2), 76–81 (2013)
3. Chhatkuli, A., Markonis, D., Foncubierta-Rodríguez, A., Meriaudeau, F., Müller, H.: Separating compound figures in journal articles to allow for subfigure classification. In: SPIE Medical Imaging (2013)

4. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from national university of singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval. p. 48. ACM (2009)
5. Foncubierta-Rodríguez, A., Müller, H.: Ground Truth Generation in Medical Imaging: A Crowdsourcing Based Iterative Approach. In: Workshop on Crowdsourcing for Multimedia, ACM Multimedia (oct 2012)
6. García Seco de Herrera, A., Foncubierta-Rodríguez, A., Markonis, D., Schaer, R., Müller, H.: Crowdsourcing for Medical Image Classification. In: Annual Congress SGMI 2014 (2014)
7. García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum) (September 2013)
8. García Seco de Herrera, A., Markonis, D., Joyseeree, R., , Schaer, R., Foncubierta-Rodríguez, A., Müller, H.: Using semi–supervised learning for image modality classification. In: Multimodal Retrieval in the Medical Domain (MRMD) 2015. Lecture Notes in Computer Science, Springer (2015)
9. García Seco de Herrera, A., Markonis, D., Schaer, R., Eggel, I., Müller, H.: The medGIFT group in ImageCLEFmed 2013. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum) (September 2013)
10. García Seco de Herrera, A., Müller, H., Bromuri, S.: Overview of the ImageCLEF 2015 medical classification task. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum) (September 2015)
11. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum) (September 2016)
12. Kalpathy-Cramera, J., Hersh, W.: Automatic image modality based classification and annotation to improve medical image retrieval. Studies in Health Technology and Informatics 129, 1334–1338 (2007)
13. Lease, M.: On quality control and machine learning in crowdsourcing. Human Computation 11, 11 (2011)
14. MaierHein, L., Mersmann, S., Kondermann, D., Stock, C., Kenngott, H.G., Sanchez, A., Wagner, M., Preukschas, A., Wekerle, A.L., Helfert, S., et al.: Crowdsourcing for reference correspondence generation in endoscopic images. In: International Conference on Medical Image Computing and ComputerAssisted Intervention. pp. 349–356. Springer (2014)
15. Mitry, D., Peto, T., Hayat, S., Morgan, J.E., Khaw, K.T., Foster, P.J.: Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of images in the epic norfolk cohort on behalf of the UK biobank eye and vision consortium. PLOS ONE 8(8) (2013)
16. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on Multimedia information retrieval. pp. 557–566. MIR '10, ACM, New York, NY, USA (2010)
17. Tirilly, P., Lu, K., Mu, X., Zhao, T., Cao, Y.: On modality classification and its use in text–based image retrieval in medical databases. In: 9th International Workshop on Content–Based Multimedia Indexing (2011)
18. Wang, C., Yan, S., Zhang, L., Zhang, H.J.: Multilabel sparse coding for automatic image annotation. In: IEEE Conference onComputer Vision and Pattern Recognition. pp. 1643–1650. IEEE (2009)