Leveraging the Crowd for Annotation of Retinal Images*

George Leifman[1], Tristan Swedish, Karin Roesch and Ramesh Raskar

Abstract— **Medical data presents a number of challenges. It tends to be unstructured, noisy and protected. To train algorithms to understand medical images, doctors can label the condition associated with a particular image, but obtaining enough labels can be difficult. We propose an annotation approach which starts with a small pool of expertly annotated images and uses their expertise to rate the performance of crowd-sourced annotations. In this paper we demonstrate how to apply our approach for annotation of large-scale datasets of retinal images. We introduce a novel data validation procedure which is designed to cope with noisy ground-truth data and with non-consistent input from both experts and crowd-workers.**

I. INTRODUCTION

Medical data is complex and requires specialized expertise for collection, analysis and interpretation. Recent rapid progress in development and deployment of personal healthcare devices has led to exponential increase in the amount of collected digital medical data.

As a result, automatic image processing and computer vision tools are required to process this data. In addition to various issues of storing and processing large medical data, such as security and accessability, one of the most concerning issues is the lack of good annotation of the data. Despite recent progress in developing unsupervised learning techniques, labeled data is still essential for both developing and testing automatic algorithms. In this paper, we address the aspect of labeling and annotating medical datasets. We demonstrate the principles of our approach in the field of retinal imaging, which has attracted a lot of attention in computer vision [1], [2], [3], [4].

Engaging only trained specialists to annotate large data sets is usually expensive and time-consuming. Our annotation approach is based on crowdsourcing. Crowdsourcing is obtaining services, ideas, or content by soliciting contributions from a large group of people, usually online, rather than from traditional employees or suppliers. The term was originally coined in 2005-2006 from the combination of words: "crowd" and "outsourcing". Since then, crowdsourcing complex image analysis and computer vision tasks has emerged as an alternative to energy-inefficient and difficult-to-implement computational approaches.

The most popular crowdsourcing scientific task is the categorization of galaxies [5]. Great potential has been shown for various biomedical tasks, such as determination of protein folding [6] and classification of malaria-infected red blood cells [7], [8]. In healthcare, crowdsourcing has been recently used for clinical diagnosis [9] and for improvement of service efficiency [10].

A recent study [11] has demonstrated that crowdsourcing can be an effective, viable and inexpensive method for the preliminary analysis of retinal images. The authors show a crowd's identification of severe abnormalities is particularly accurate with a sensitivity of 98% and distinguished between normal and mildly abnormal images with a sensitivity between 61% and 72%. The study's aim was to assess the performance and repeatability of crowdsourcing for the classification of retinal fundus photography. Therefore, only 100 images were used. Our goal is to build a platform for labeling and annotating large medical datasets.

While using crowdsourcing, the main question is "Can we trust the crowd?". To make medical diagnosis reliable, we use an initial experts' input, and the confidence of crowdsourcers is verified by the experts' input.

Many systemic diseases, such as diabetes and hypertension, show manifestations in the eye. Diabetic retinopathy (DR) is a complication of diabetes and the second most common cause of blindness in the United States. According to the National Diabetes Statistics Report, the number of patients with diabetes in the United States is 29.1 million (or 9.3% of the population).

Public datasets: To our knowledge at least ten public datasets for evaluation of retinal imaging algorithms exist: STARE [12], DRIONS [13], DIARETDB1 [14], DRIVE [15], MESSIDOR [16], ONHSD [17]. HEIMED [18], ROC [19], ARIA [20], DR12 [21]. The detailed description of the datasets can be found in the Appendix.

Most of the datasets contain similar information, however the annotations were performed employing different approaches, resulting in various confidence of the data. This makes it almost impossible to produce a fair comparison of the results across different datasets. Our aim is to standardize the process and to provide a tool for annotation of large-scale datasets.

In this paper we propose a framework for labeling retinal images. Moreover, we introduce a novel approach for data validation, which includes validation of the ground-truth and the annotation input.

II. DATA ACQUISITION

Medical image data presents an interesting challenge since oftentimes the knowledge is only held by physician experts. We built a web-based framework for obtaining reliable image annotation. Our goal is to present users with Human Intelligence Tasks (HITs) that sample their knowledge in

*Research supported by Vodafone Americas Foundation

[1] The authors are at the MIT Media Lab, Massachusetts Institute of Technology, Cambridge MA 02139. (Corresponding author email: gleifman at mit.edu).

interpreting images. Following the success of crowdsourcing efforts such as developed by Mavandadi et al. [8], we recruit a small number of experts to build a HIT test set that helps establish crowdsourced annotation session quality. We then use session quality to weight crowdsourced annotations through a voting system that determines which annotations are reliable. New annotations that have been deemed reliable are then added to the HIT test set.

We collect user annotations through a custom built web application (http://ilabelit.media.mit.edu), enabling users to manipulate images in a way suitable for medical images. Our approach allows fine-tuning of the interface presented to crowdsourced users, not available from other HIT collection platforms such as Amazon's Mechanical Turk.

The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board.

We use the web app development stack called MEAN (MongoDB, Express.js, Angular.js, Node.js), which allows for rapid extension of our interface and backend.

We designed two different types of user interface for different labeling tasks. Global labels are applied using the Diagnostic Labeling Tool (DLT), and drawn or point annotations are applied using the Feature Labeling Tool (FLT). In our case, each tool is used to generate an independent dataset from a common pool of images. The DLT provides a simplified interface that is ideal for quickly tagging images, and the FLT is best suited for carefully marking anatomical features. In either case, our methods can be applied with input from either tool.

The DLT consists of a simple image manipulation area and two labeling fields (see Figure 1). The user is asked to fill information about the image quality into the first field and either healthy or labeled disease in the second field.

The DLT was designed to provide global labels, but point and line drawn annotations can be added by the FLT (see Figure 2). A feature can be selected, and then point marks can be added by clicking on the image. If the feature takes up a larger area, multiple clicks will produce an array of point marks connected by lines.

Our database consists of a web store of images connected to our server using the Google Cloud Storage API. De-identified image file names and metadata are stored as entries in the MongoDB database. The web app client generates an annotation entry in collaboration with the user, which is dynamically saved to the server using POST requests.

III. DATA VALIDATION

Our goal is to build reliable ground-truth data. Usually, there is no ultimate, unique ground-truth in medical annotations. Therefore, our ground-truth consists of annotations that have high degree of consensus between experts and crowd-workers. We developed an approach to validate the annotation input from various annotators. Unlike Mavandadi et al. [8] we start from the ground-truth of various publicly available databases (see Appendix). As annotations in the public datasets are prone to mistakes and need to be validated

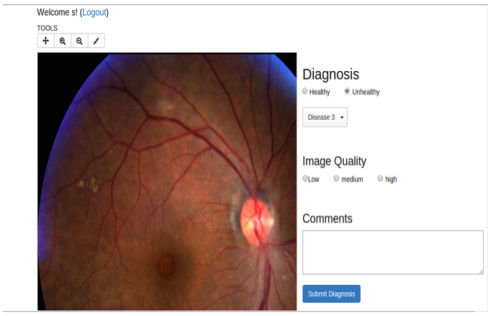

Fig. 1. **Diagnostic Labeling Tool (DLT):** We show the basic layout for assigning disease state to retinal images. In the top left, the user has access to basic image manipulation tools, such as drag and zoom. Along the right of the image the user selects either "Healthy" or "Unhealthy". The user is presented with an additional drop down menu to select a diagnosis when the image is tagged as "Unhealthy". Furthermore, the user can input the image quality and any comments they have.

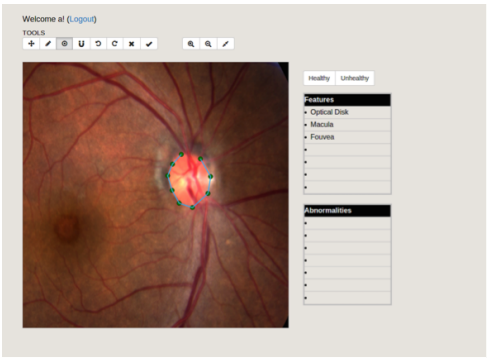

Fig. 2. **Feature Labeling Tool (FLT):** The user interface for annotating anatomical features. The user has access to basic image manipulation tools in order to inspect the image. We've also added "dot" and "line" tools which allow the user to create point annotations or draw an outline of a particular feature or abnormality. Feature and Abnormality labels are added to a growing list on the right. Each set of annotations is saved as a data structure in a database on the server.

our approach of data validation includes validation of the ground-truth and the annotation input (see Figure 3).

To do so we created two data structures: Ground-Truth Structure (GTS) and Temporal Ground-Truth Structure (TGTS). The GTS structure includes the images associated with their corresponding high-confident labels. Each label can be (0) Healthy, (1) Questionable or (2) Unhealthy. For each image we store the number of annotators that agreed on the associated label. The GTS is initialized with the data from the various publicly available databases. The number of annotators is set according to the number of experts as mentioned in the description of the dataset.

The TGTS contains images with either no label at all or with labels that are not high-confident. For each image in this structure we keep a list of labels from the annotators. Images are moved from TGTS to GTS when a certain level of confidence of the corresponding labels is reached.

We assume that even the most trustworthy annotator can be incorrect due to various factors. Therefore we combine the labels from each annotator in sessions and calculate the confidence for each session. In each session the annotator is

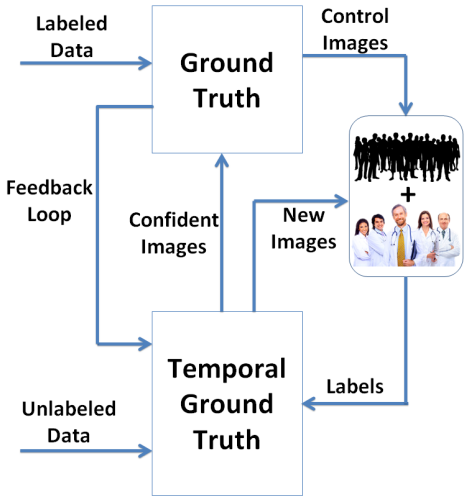

Fig. 3. **Data validation**: The GTS is initialized with labeled data and the TGTS is initialized with unlabeled images. A degree of agreement is calculated for each annotator's session. Labels from sessions with high degree of agreement are added to the TGTS. TGTS images with consistent labels are moved to the GTS. A fraction of GTS images that have the lowest number of labels is moved back to TGTS.

presented with a mixture of images from the GTS and from the TGTS. The confidence of the session is calculated as a degree of agreement between the labels from the annotator and the labels from the GTS (only the portion of images that are taken from GTS is included). The degree of agreement is calculated using Weighted Kappa coefficients [22].

Kappa coefficients [23], also known as Cohen's kappa coefficient, is a statistical measure of inter-rater agreement or inter-annotator agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since it takes into account the agreement occurring by chance. Kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}, \tag{1}$$

where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly choosing each category. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters other than what would be expected by chance (as defined by $Pr(e)$), κ = 0.

Weighted kappa [22] is useful when labels are ordered. Three matrices are involved: The matrix of observed scores, the matrix of expected scores based on chance agreement, and the weight matrix. Weight matrix cells located on the diagonal (upper-left to bottom-right) represent agreement and thus contain zeros. Off-diagonal cells contain weights indicating the seriousness of that disagreement. The equation for weighted κ is:

$$\kappa = 1 - \frac{(1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij})}{(1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij})}, \tag{2}$$

where k is number of codes and w_{ij}, x_{ij} and m_{ij} are elements in the weight, observed, and expected matrices, respectively. When diagonal cells contain weights of 0 and all off-diagonal cells weights of 1, this formula produces the same value of kappa as the calculation given above. In cases of many gradations the TrueSkill ranking [24] approach can be used to get more robust results.

The above scores for Kappa coefficient calculation are used for the Diagnostic Labeling Tool. While working with Feature Labeling Tool, Kappa coefficients must be updated with another type of scores. For example, following [25], to indicate how similar an observer's segmentation is to an expected segmentation - accuracy and precision scores can be used as follows. Let A represent the expected region and B - the observed segmentation. The accuracy is calculated as $\frac{|A \cap B|}{|A|}$ and the precision is $\frac{|A \cap B|}{|A \cup B|}$.

Our data validation algorithm can be summarized as follows. We start with two structures, GTS and TGTS. The GTS is initialized with the data from publicly available datasets and the TGTS is initialized with unlabeled images. For each annotator's session we calculate the session degree of agreement using weighted Kappa coefficients. If a session degree of agreement is high enough (> 0.8) we add the session labels to the TGTS. After each session we move TGTS images that have enough labels (> 10) and the labels are consistent ($> 90\%$ agreement) to the GTS. All the constants are established empirically.

Finally, to be able to constantly verify the labeling in the GTS, we move a fraction (5%) of GTS images that have the lowest number of labels to the TGTS. This feedback loop, not only allows verification of annotators' input, but also allows adjustment to the original labeling.

IV. CONCLUSIONS

We propose a general framework for labeling medical data. Moreover, we introduce a novel approach for data validation that includes validation of the ground-truth and the annotation input. Our validation procedure is designed to cope with noisy ground-truth data and with non-consistent input from both experts and crowdworkers. We illustrate our approach using the example of retinal imaging, but the approach is applicable to almost any kind of medical data.

Future work: Our initial goal is to build a large dataset of annotated retinal images and release it for public use. We will use it as a test case to then build a more general framework that can be used for various medical data.

We would like to explore the potential of our approach in a tele-medicine setting. Having enough crowdworkers and being really confident about their annotations, we can expect to get an almost real-time diagnosis for any new data.

Finally, we would like to explore a way to incorporate a machine learning annotation algorithm into our scheme.

APPENDIX

The following datasets of retinal imaging are available: STARE [12], DRIONS [13], DIARETDB1 [14],

DRIVE [15], MESSIDOR [16], ONHSD [17]. HEI-MED [18], ROC [19], ARIA [20], DR12 [21].

STARE includes around 400 images and diagnosis for each image. In addition, for a subset of images, it has textual expert annotations of the features, blood vessel segmentation, artery/vein labelings and optic nerve detection. In DRIONS 110 fundus images with their optical disk manually segmented by two different specialists are included. DIARETDB1 consists of 89 fundus images; 84 contain microaneurysms and 5 do not contain any signs of the Diabetic Retinopathy. DRIVE contains 40 retinal images; 33 do not show any sign of Diabetic Retinopathy and 7 show signs of mild early Diabetic Retinopathy. MESSIDOR contains 1200 images, with two diagnosis provided: retinopathy grade and risk of macular edema. ONHSD contains 99 images from 50 patients; 96 images have discernable Optic Nerve Hypoplasia (ONH). HEI-MED includes data from 169 patients with various levels of Diabetic Macular Edema (DME). ROC contains 100 photographs, marked as containing microaneurysms. ARIA is organised into three categories: age-related macular degeneration (AMD) subjects (n=23), healthy control-group subjects (n=61), and diabetic subjects (n=59). DR12 dataset includes two subsets DR1 and DR2. The DR1 subset has 5776 images, where 1392 represent poor quality and 3084 images of the periphery. The DR2 subset comprises 920 images, where 260 are not centered on the macula and 660 images centered on the macula (466 good and 194 low quality). In February 2015 a DR detection competition was announced by Kaggle [26]. Their dataset has 35000 images rated with 5 levels of Diabetic Retinopathy.

ACKNOWLEDGMENTS

We would like to thank Vishal Gupta, Fabin Rasheed and Dhruv Joshi at the Srujana - Center for Innovation, LVPEI for help with the development of the annotation application. We thank the members of the Camera Culture Group at the MIT Media Lab for helpful discussions and clinical collaborators for their feedback.

REFERENCES

[1] R. N. Maamari, J. D. Keenan, D. A. Fletcher, and T. P. Margolis, "A mobile phone-based retinal camera for portable wide field imaging," *British Journal of Ophthalmology*, 2013.

[2] V. F. Pamplona, A. Mohan, M. M. Oliveira, and R. Raskar, "NETRA: interactive display for estimating refractive errors and focal range," in *TOG*, vol. 29, p. 77, ACM, 2010.

[3] T. Swedish, K. Roesch, H. Ik Lee, and R. Raskar, "Self directed eye alignment using reciprocal eye box imaging," in *TOG*, ACM, 2015.

[4] E. Zvornicanin, J. Zvornicanin, and B. Hadziefendic, "The use of smart phones in ophthalmology," *Acta Informatica Medica*, vol. 22, no. 3, p. 206, 2014.

[5] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, *et al.*, "Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189, 2008.

[6] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, *et al.*, "Predicting protein structures with a multiplayer online game," *Nature*, vol. 466, no. 7307, pp. 756–760, 2010.

[7] M. A. Luengo-Oroz, A. Arranz, and J. Frean, "Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears," *Journal of medical Internet research*, vol. 14, no. 6, 2012.

[8] S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen, and A. Ozcan, "Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study," vol. 7, no. 5, p. e37245, 2012.

[9] T. B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, J. E. Burns, and R. M. Summers, "Distributed human intelligence for colonic polyp classification in computer-aided detection for ct colonography," *Radiology*, vol. 262, no. 3, pp. 824–833, 2012.

[10] A. Turner, K. Kirchhoff, and D. Capurro, "Using crowdsourcing technology for testing multilingual public health promotion materials," *Journal of medical Internet research*, vol. 14, no. 3, 2012.

[11] D. Mitry, T. Peto, S. Hayat, J. E. Morgan, K.-T. Khaw, and P. J. Foster, "Crowdsourcing as a novel technique for retinal fundus photography classification," *PloS one*, vol. 8, no. 8, p. e71154, 2013.

[12] A. Hoover and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 8, pp. 951–958, 2003.

[13] E. J. Carmona, M. Rincón, J. García-Feijoó, and J. M. Martínez-de-la Casa, "Identification of the optic nerve head with genetic algorithms," *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 243–259, 2008.

[14] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "The DIARETDB1 diabetic retinopathy database and evaluation protocol.," in *BMVC*, 2007.

[15] J. Staal, M. D. Abràmoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 4, pp. 501–509, 2004.

[16] E. Decenciere, X. Zhang, G. Cazuguel, B. Laÿ, B. Cochener, C. Trone, P. Gain, J.-R. Ordónez-Varela, P. Massin, A. Erginay, *et al.*, "Feedback on a publicly distributed image database: The messidor database," pp. 231–234, 2014.

[17] J. Lowell, A. Hunter, D. Steel, A. Basu, R. Ryder, E. Fletcher, and L. Kennedy, "Optic nerve head segmentation," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 2, pp. 256–264, 2004.

[18] L. Giancardo, F. Meriaudeau, T. P. Karnowski, Y. Li, S. Garg, K. W. Tobin, and E. Chaum, "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," *Medical Image Analysis*, vol. 16, no. 1, pp. 216–226, 2012.

[19] M. Niemeijer, B. van Ginneken, M. Cree, A. Mizutani, G. Quellec, C. Sanchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, X. Wu, G. Cazuguel, J. You, A. Mayo, Q. Li, Y. Hatanaka, B. Cochener, C. Roux, F. Karray, M. Garcia, H. Fujita, and M. Abràmoff, "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 185–195, 2010.

[20] D. J. Farnell, F. Hatfield, P. Knox, M. Reakes, S. Spencer, D. Parry, and S. Harding, "Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators," *Journal of the Franklin institute*, vol. 345, no. 7, pp. 748–765, 2008.

[21] R. Pires, H. F. Jelinek, J. Wainer, and A. Rocha, "Retinal image quality analysis for automatic diabetic retinopathy detection," in *SIBGRAPI*, pp. 229–236, 2012.

[22] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.," *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[23] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[24] R. Herbrich, T. Minka, and T. Graepel, "Trueskill(tm): A bayesian skill rating system," in *Advances in Neural Information Processing Systems 20*, pp. 569–576, 2007.

[25] D. Gurari, S. K. Kim, E. Yang, B. Isenberg, T. A. Pham, A. Purwada, P. Solski, M. Walker, J. Y. Wong, and M. Betke, "Sage: An approach and implementation empowering quick and reliable quantitative analysis of segmentation quality," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pp. 475–481, IEEE, 2013.

[26] "Diabetic retinopathy detection competition," 2015. https://www.kaggle.com/c/diabetic-retinopathy-detection.