A Web-Based Platform for Distributed Annotation of Computerized Tomography Scans

Nicholas Heller(✉), Panagiotis Stanitsas, Vassilios Morellas, and Nikolaos Papanikolopoulos

University of Minnesota Center for Distributed Robotics, 117 Pleasant St SE, Minneapolis, MN 55455, USA
{helle246,stani078,morellas,papan001}@umn.edu
http://distrob.cs.umn.edu

Abstract. Computer Aided Diagnosis (CAD) systems are adopting advancements at the forefront of computer vision and machine learning towards assisting medical experts with providing faster diagnoses. The success of CAD systems heavily relies on the availability of high-quality annotated data. Towards supporting the annotation process among teams of medical experts, we present a web-based platform developed for distributed annotation of medical images. We capitalize on the HTML5 canvas to allow for medical experts to quickly perform segmentation of regions of interest. Experimental evaluation of the proposed platform show a significant reduction in the time required to perform the annotation of abdominal computerized tomography images. Furthermore, we evaluate the relationship between the size of the harvested regions and the quality of the annotations. Finally, we present additional functionality of the developed platform for the closer examination of 3D point clouds for kidney cancer.

1 Introduction

Medical imaging modalities contain a wealth of useful information for the diagnosis of a wide range of ailments, rendering them an essential component of the diagnostic process. A plethora of tools for the accurate identification of risk markers for different pathologies in medical images has been developed (e.g. [2–6]). Such inference schemes require large amounts of annotated data, which are used for the training of supervised or semi-supervised models. Unfortunately, the very high cost of the annotation process associated with medical images results in a lack of publicly available benchmarks (e.g. [8,9]). The high cost can be attributed to the requirement of highly trained experts for providing the annotations. This scarcity of annotated data is prohibitive to the development of Computer Aided Diagnosis (CAD) for a variety of pathologies.

The overall objective of this study is concerned with the development of a CAD scheme for the localization and the health assessment of kidneys from abdominal Computerized Tomography (CT) scans. In that direction, two sub-problems can be identified; first the accurate localization and segmentation of

© Springer International Publishing AG 2017
M.J. Cardoso et al. (Eds.): CVII-STENT/LABELS 2017, LNCS 10552, pp. 136–145, 2017.
DOI: 10.1007/978-3-319-67534-3_15

the organ (kidney) and the aorta and, second, the automated identification of abnormal masses (malignant tissue, benign tissue and cysts).

With the support of our medical collaborators, a collection of several hundred abdominal CT scans of kidney cancer patients has been acquired. A majority of the patients' pathologies are clear cell Renal Cell Carcinomas (RCCs) but papillary RCCs, angiomyolipomas, renal oncocytomas, and papillary urothelials are represented as well. Our intention is to create a rich collection of accurate delineations of abnormalities developed by the kidneys. This introduces an annotation burden which is distributed among urologists at different locations.

A large variety of tools is available for the generic annotation of images. Such tools were designed with much different tasks in mind and have a large number of extraneous features which, for an application like the one in hand, would unnecessarily increase the complexity of the annotation sessions. Two examples of such tools are the GNU Image Manipulation Program[1] and Adobe Photoshop[2].

Furthermore, the anticipated high data volume creates the need for a centralized storage and backup platform. In that way, users are not required to manually update annotation repositories after each session, and only necessitates redundancies at the server level, rather than the personal computer level.

2 Related Work

A number of specialized tools tailored to the task of high-volume image annotation have been created. One such platform is the Automated Slide Analysis Platform (ASAP)[3]. ASAP was built for the specific task of annotating whole slide histopathology images. It includes a large collection of tools for this task including the ability to draw polygons from sets of points and to create splines.

According to our partnering medical experts, certain features of ASAP are more relevant to the annotation of histopathological data. In our case, the most convenient way to segment our regions of interest was to simply draw outlines and fill them. Therefore, many of ASAP's features are vestigial to our task and would introduce unnecessary complexity. Additionally, ASAP is a desktop tool which requires the users to store a copy of the data locally. This is not ideal for our task for the reasons discussed in the previous section. Further, in order to save an annotation and move on to the next image or feature, at least 5 clicks are required by the user, on top of decisions he or she must make about where to store the annotation and which file to open next. This introduces a significant amount of unnecessary work which frustrates users and reduces efficiency.

Another platform that was created for this task is MIT CSAIL's LabelMe [1] website. This platform is well-made and better suited for our task than ASAP since it is web-based with central data management and requires only a single

[1] https://www.gimp.org/.
[2] https://www.adobe.com/products/photoshop.html.
[3] https://github.com/GeertLitjens/ASAP.

click to move to the next image. However, it is missing features which are critically important to our design objective. For instance, the tool only supports drawing using point-defined polygons. According to the experts we talked to, this is not ideal. Additionally, LabelMe draws with full opacity, and a simple experiment showed us that full opacity leads to higher variability among annotators and overall lower accuracies. Furthermore, the LabelMe interface does not have a "previous" button which medical experts told us was essential to their ability to accurately annotate, presumably so that they could conveniently flip back and forth between sequential frames in order to make better informed decisions about which regions are which.

In contrast, our platform was designed with the following three core requirements, namely, (i) distributed capabilities, (ii) robust and secure centralized data storage and, (iii) a lightweight interface focusing on the task in hand. Our use of the HTML5 canvas element makes this realizable. Additionally, in order to ensure a user-friendly presentation, our platform capitalizes on the Bootstrap[4] framework.

3 The Interface

The interface of the developed scheme was based on the the Bootstrap framework. In particular, we used Start Bootstrap's SB Admin template[5], since it allows for the *landing* page to provide the user with information on the state of the system. In our case, this is to display the annotation progress on a particular project. This landing page is depicted in Fig. 1. When the user clicks on the

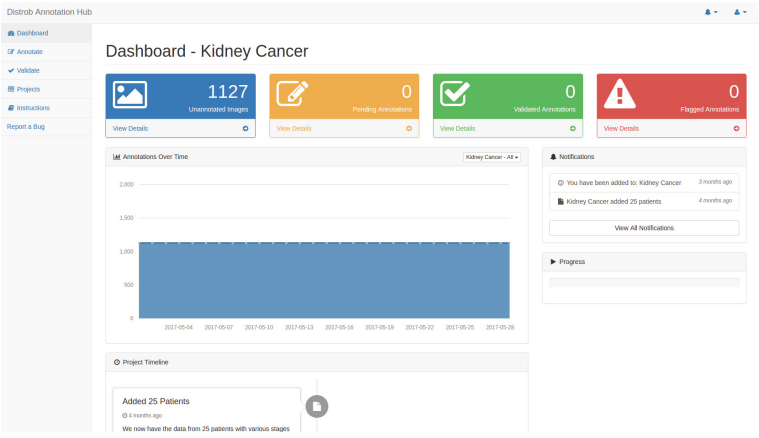

Fig. 1. The four colored cards correspond to the number of images belonging to each of the four bins: unannotated, pending, validated, and rejected. The proportions in each bin are visualized by the graph below the cards. This screen capture was taken when no images were yet annotated. (Color figure online)

[4] http://getbootstrap.com/.
[5] https://startbootstrap.com/template-overviews/sb-admin/.

unannotated card or the *annotate* button on the top left, it brings them to the image-set selection page. Here, the user sees a vertical list of image-sets, each corresponding to a set of slices from a single CT scan. If an image-set has been annotated by another user in the past hour, it shows the name of the user who made the most recent annotation and the time at which it was submitted. The user also has the option of selecting *auto* in which case the system will direct the user to either the last set he/she annotated, or a random unannotated set. A screen capture of this page is depicted in Fig. 2.

Once the user selects an image-set to annotate, it brings them to the page depicted in Fig. 3. Here, he/she is presented with an image in the center of the screen, with thumbnails of the features already annotated below it, and a toolbar above it. Among the tools are *previous* and *next* buttons, a bar of small thumbnails of each slice to choose from, and *submit* buttons for each feature. The user may use the *bucket icon* to switch his/her tool to a bucket fill, or simply by right clicking which also performs this action.

The platform makes use of the CSS3 *filter* element to adjust its brightness and contrast. Medical experts have particular preferences for brightness and contrast for CT images that depend on which part of the body it depicts, and which organs they are studying. We selected abdomen brightness and contrast values (170% and 300%, respectively) by iteratively adjusting and getting feedback from expert urologists.

For this annotation task, we would like segmentation data for five regions of interest: left kidney, right kidney, left mass, right mass, and aorta. If a particular region doesn't exist in an image, the user simply omits that submission, or submits a blank canvas. Once an annotation is submitted, it falls to its respective thumbnail under the large image. Until then, those thumbnails remain gray placeholders.

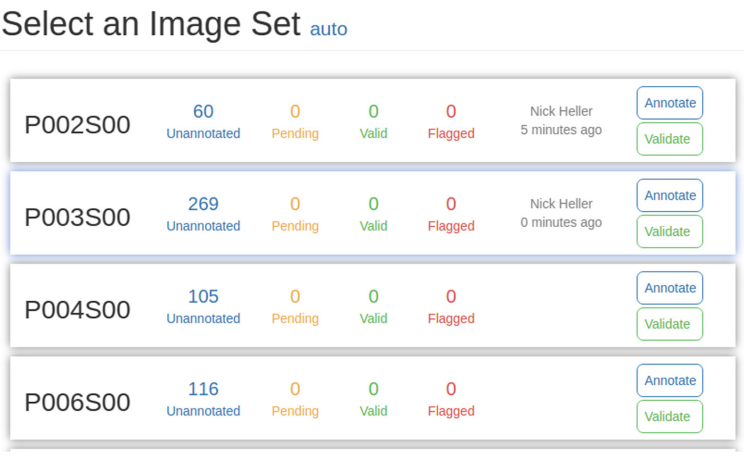

Fig. 2. The leftmost text of an image-set is that image-set ID, the first denoting patient 2, set 0. Next in from the left is a breakdown of the bins each image in the set resides in. Next is the aforementioned notice of recent activity. Finally we have the annotate and validate buttons.

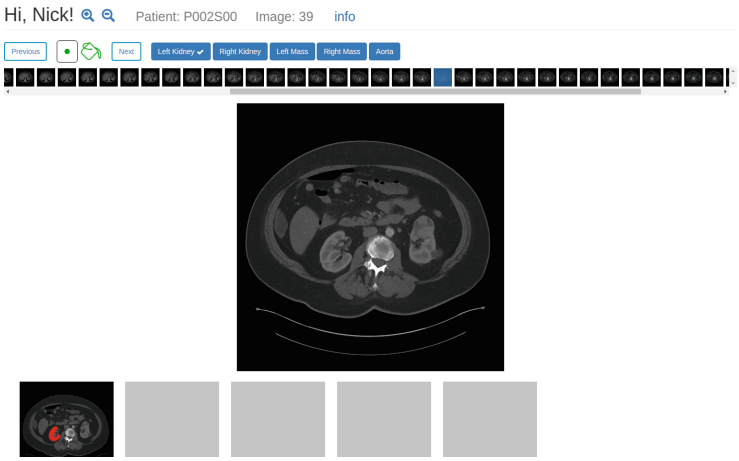

Fig. 3. The interface with which users create their drawn annotations.

The users also have the option to validate annotations. This is a simple binary feedback process on an interface that follows the design of the annotation interface, but instead of five submit buttons, there are only two: accept or reject, after which it stores the response and presents the next annotation. This feature was deployed for ensuring the high quality of the harvested annotations.

4 The Backend

In this section we briefly discuss the backend of this platform. The platform stack is Linux Apache MySQL PHP (LAMP) with some flat files of structured data (JSONs) used for configuration and databasing. The software would likely run slightly faster if the flat files were migrated to MySQL, but as of now, speed is not a major concern.

During annotation, the *brush strokes* and *fill* commands from the user are individually stored locally to allow for undo and redo operations. Once the annotation is submitted, it is stored on the server as a whole image.

5 Evaluation

There are two components to the task of evaluating this platform, namely (i) evaluate the interface of the platform from a general standpoint of interaction design and ease of use, and (ii) evaluate the interface's capacity for allowing users to produce highly accurate image annotations.

5.1 Evaluating Interaction Design and Ease of Use

In addition to the design guidelines given by medical experts during this platform's initial development, we conducted a heuristic evaluation using the Nielsen

Norman Group's 10 heuristics [7]. This technique was selected because it has been shown to be a very effective and low-cost method for identifying weaknesses in a user interface. In that way, the types of flaws that a user study might miss are also identified. As is standard practice, the platform was evaluated by 3 experts trained in heuristic evaluation. Each expert compiled a list of heuristic violations independently. Then, the collected information was consolidated and each violation received and ordinal (0-4) severity score. Those were then averaged to yield the final evaluation score.

Our heuristic evaluation identified 13 violations. Only one violation received a 3, the rest were rated 2 or lower, and highly ranked one was identified as a known issue which at the time of writing was being worked on and near resolution. For brevity, we refrain from listing each violation here, but some clusters we noticed were (i) our platform suffers from the so-called "pottery barn principle" where certain actions have no or limited undo functionality, so users sometimes feel as though they are walking around in a pottery barn, which significantly impairs the user experience, and (ii) our error messages lack informative and constructive feedback about how to proceed in the event of each error. Improvements which address these issues have been slated for development and will likely be deployed a few weeks after writing.

5.2 Evaluating Data Quality

It is important to ensure that the annotations completed with this platform accurately represent the intentions of the expert performing them. We identified region size as a factor which impacts annotation precision. Towards developing size guidelines for freehand annotations, we performed a study in which a single user annotated the same kidney 16 times at each of 8 different levels of zoom. In addition to the annotations, we recorded the time of continuous annotation that the user took during each of the sessions.

To measure precision, for every possible pair in the 16 annotations, we computed the proportion of pixels that are highlighted in one annotation but not in the other, to the number of pixels highlighted in the union of the annotations. We multiplied this by 100 and computed the mean over all pairs which we interpret as the average percentage of deviation at a given level of zoom. The results of this study are shown in Fig. 4.

Our results suggest that there is an inverse correlation between the size of the feature on the screen and the users' error in consistently annotating that feature. The near-highest level of consistency can be seen to occur at feature sizes larger than 10 cm. Further, there appears to be a positive correlation between the size of the feature and the average annotation time.

The focus of this work was to construct a platform for distributing the annotation load across different locations. We wanted to achieve this in such a way that minimized the time elapsed for the pertinent tasks to the annotation. These include saving the annotations properly, and finding and opening the next image. These tasks are cumbersome in the existing more general-purpose GNU Image Manipulation Program (GIMP). In a similar experiment with the same user, we

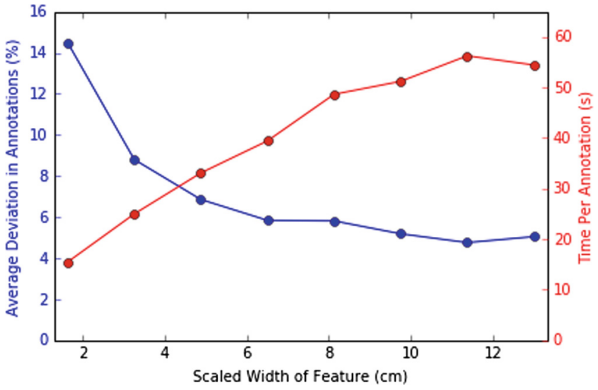

Fig. 4. The downward sloping line (blue) corresponds to the left y-axis and the upward sloping line (red) corresponds to the right y-axis. This chart suggests that for the task of highlighting kidneys, increasing the size of the kidney on the screen up to about 10 cm will improve the annotation consistency, but beyond that, little to no further gains can be made. (Color figure online)

found the mean annotation time using GIMP to be ~106 s per region of interest at a scaled width of 8.125 cm and ~123 s per region of interest at a scaled width of 11.375 cm. This suggests that our platform provides a 54% time improvement over GIMP, while no significant difference in consistency was found.

In order to better understand the nature of the deviations, we conducted a follow-up study in which a user who was not familiar with the previous experiment selected a level of zoom at which he/she felt comfortable to perform accurate annotations, and provided 60 annotations of the same feature. A visualization of these annotations are shown in Fig. 5. This user was instructed to focus only on annotation consistency and told that time was not a factor.

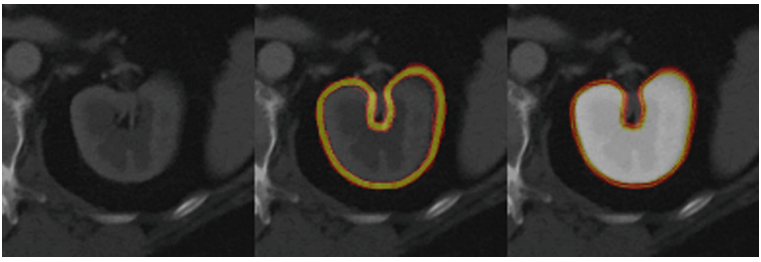

Fig. 5. The background of each image is identical. The left has no overlay, the right is overlaid with each pixel's variance, and the right is overlaid with each pixel's mean. The color-map used is OpenCV's COLORMAP_HOT. We omit a numerical scale since the translucent overlay invalidates any correspondence and the figure is only intended to show a trend. (Color figure online)

The level of zoom that the user selected corresponded to a feature width of less than 4 cm, and that user was very surprised to find that annotations varied, on average, by 11%. This suggests that users' intuition is not accurate at guessing the expected consistency amongst annotations, and that such evaluation studies are deemed necessary prior to investing large amounts of resources in labeling.

6 Future Work

In the near future we intend to release this project as open source software so that other groups can install and serve the platform for their own research purposes. Before we do this, however, there are a number of scheduled improvements to both code clarity and the platform itself.

6.1 UI Improvements

A limited number of potential improvements in the appearance and interaction patterns have been identified from both the heuristic evaluation and the user studies conducted. Most of those could be addressed with relatively little development time. The improvements we intend to make include (i) making the interface more conducive to annotating highly zoomed images, (ii) modifying error messages to be more informative and constructive, (iii) introducing additional functionality to enable users to undo/redo pieces of their brush strokes individually, and (iv) extending the platform such that new annotation projects can be added with a simple addition to a configuration file.

6.2 Added Functionality

The main focus of this work is to reduce the time required by experts to annotate regions of interest. With that in mind, we plan studying the possibility of developing schemes which suggest annotations for each region of interest. These would then be further tuned by the experts, rather than requiring the experts to start drawing from scratch, as implemented in the present version.

Furthermore, the development and evaluation of a system which offers a number of different annotation suggestions and asks the user to select the best among them, is under construction. This process could iterate until the expert is satisfied with the suggestion windows provided by the tool. Heuristically we believe that either of these schemes, or a combination of the two, would result in a significant time improvement over the current method, without compromising the annotation quality.

6.3 Further Evaluation of Annotation Quality

It is imperative that we not only ensure that annotations are performed quickly, but also that they accurately reflect the features they are attempting to segment. We plan to further study this issue through large scale auditing throughout the

annotation process. Certain randomly selected image-sets will be duplicated and blindly assigned to additional users to evaluate consistency and identify any biases that certain annotators might hold. This work will further inform our development efforts to mitigate this issue moving forward.

6.4 Utilizing 3D Information

When paired with the annotation data–or conceivably, the segmentations produced by our network–the marching cubes [10] algorithm can be used to create a 3D reconstruction of the features. This reconstruction could be useful for informing treatment decisions or for giving surgeons a better visualization of an area they may be preparing to operate on. We wrote an offline script which, given these annotations, creates this reconstruction. An example is shown in Fig. 6.

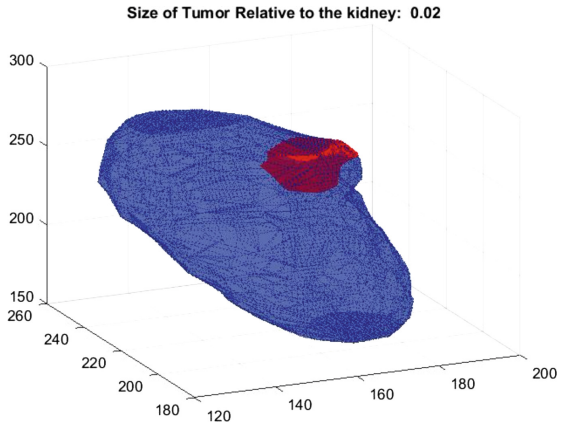

Fig. 6. A 3D reconstruction of a kidney (blue) and tumor (red) based on annotations of slices using our system. (Color figure online)

We plan to further explore ways to present these reconstructions to medical professionals so as to maximize their utility. One idea is to integrate this presentation into the current interface using the WebGL library. Another is to import our meshes into a virtual reality platform.

Acknowledgments. We thank Maxwell Fite and Stamatios Morellas for their expertise on heuristic evaluation, Drs. Christopher Weight, Niranjan Sathianathen, and Suprita Krishna for their feedback on the initial development process, and Samit Roy, Meera Sury, and Michael Tradewell for annotations completed thus far. "'This material is partially based upon work supported by the National Science Foundation through grants #CNS-0934327, #CNS-1039741, #SMA-1028076, #CNS-1338042, #CNS-1439728, #OISE-1551059, and #CNS-1514626.'"

References

1. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. Int. J. Comput. Vis. **77**, 157–173 (2008)
2. Shehata, M., Khalifa, F., Soliman, A., Abou El-Ghar, M., Dwyer, A., Gimelfarb, G., Keynton, R., El-Baz, A.: A promising non-invasive cad system for kidney function assessment. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 613–621. Springer, Cham (2016)
3. Dhungel, N., Carneiro, G., Bradley, A.P.: The automated learning of deep features for breast mass classification from mammograms. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 106–114. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_13
4. Wang, J., MacKenzie, J.D., Ramachandran, R., Chen, D.Z.: A deep learning approach for semantic segmentation in histology tissue images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 176–184. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_21
5. Xu, T., Zhang, H., Huang, X., Zhang, S., Metaxas, D.N.: Multimodal deep learning for cervical dysplasia diagnosis. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 115–123. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_14
6. Xu, Y., Li, Y., Liu, M., Wang, Y., Lai, M., Chang, E.I.-C.: Gland instance segmentation by deep multichannel side supervision. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 496–504. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_57
7. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: Proceedings of ACM CHI 1994 Conference, pp. 152–158 (1994)
8. CAMELYON16, ISBI challenge on cancer metastasis detection in lymph node. https://camelyon16.grand-challenge.org/
9. Multimodal brain tumor segmentation challenge 2017. http://braintumorsegmentation.org/
10. Lorensen, W., Cline, H.: Marching cubes: a high resolution 3D surface reconstruction algorithm. In: Proceedings of SIGGRAPH 1987, vol. 21, pp. 163–169 (1987)