**Peer Review File**

**Article information: https://dx.doi.org/10.21037/atm-21-6672**

**Reviewer A**

This paper presents a creative method of efficiently digitizing tabular medical records written in Chinese.
Processing photographed documents, not scanned ones, is an area of technical difficulty. The authors challenged this problematic task and achieved excellent results. Another advantage is that the processing speed is relatively fast while using deep learning.

However, considering that the area covered by the authors is medical care, it is necessary to dive into detail that the accuracy of 91.10%.

**Comment 1:** Would you please give an example of a well-processed document and a document that doesn't? And could you provide your opinion on its importance from a medical point of view and how it will be handled in the future?
**Reply 1:** Thank you for pointing out this problem in our manuscript. According to your suggestion, we drew Figures 7 and 8 to show good and bad examples, respectively. The reason our system does not perform well in Figure 8 is the presence of multiple lines of text in the cell. In response to this situation, we will train a table structure recognition model to recognize the table structure and merge multiple lines of text in cells in the future.
**Changes in the text:**
a. We added some content, "**DeepSSR shows excellent performance in the structured recognition of most UPBMR images (Figure 7). However, DeepSSR does not perform well for recognizing images with multiple lines of text in a cell (Figure 8)**" (see Page 18, line 359).
b. We added some discussion, "**3) DeepSSR performs poorly for images with multiple lines of text in one cell. Although these images are less numerous in real life, their medical value is equally important. In the future, we will train a table structure recognition model to recognize cells in a table, and then merge multiple lines of text in the cells**" (see Page 23, line 456).
c. We drew Figures 7 and 8 (see Page 28, line 515).

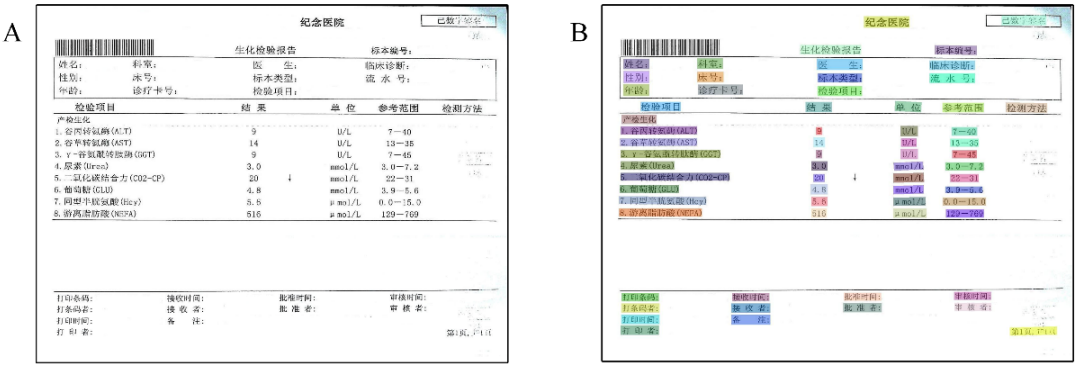



Figure 7. An example of a well-processed document. (A) is the input image. (B) is the character recognition result. (C) is structured data.

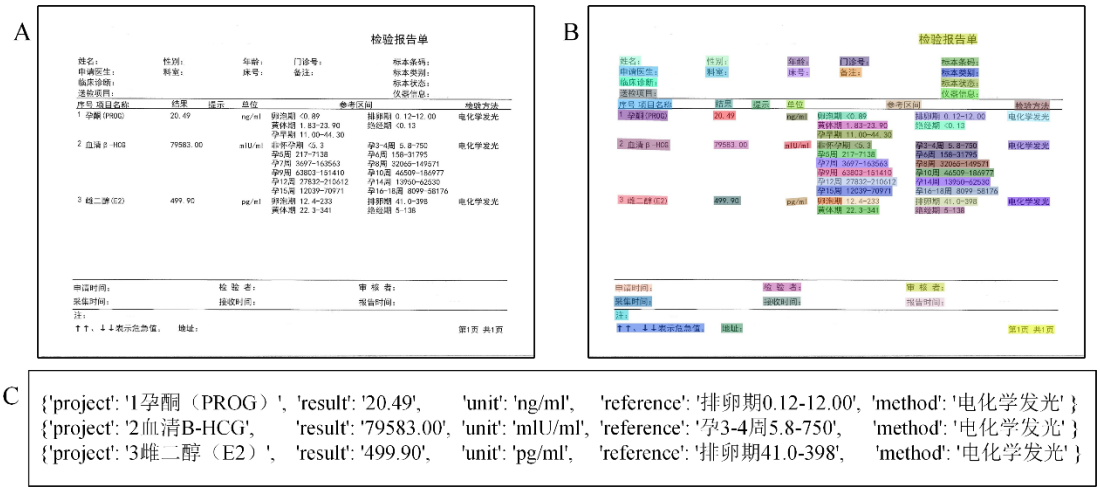



Figure 8. An example of a poorly processed document. (A) is the input image. (B) is the character recognition result. (C) is structured data.

**Reviewer B**

This paper discussed an important topic of paper-based medical records recognition. It is necessary since the heterogeneity of different records from different organizations can bring barriers to information communication and cognitive burdens to the health care providers.

The authors did a great job describing the parts of their model on the four tasks - image preprocessing, table detection, text recognition, and text box assignment. And they provided reasonable evaluations on some parts of their models.

I just have a few minor points.

**Comment 2:** The latest works in the literature review ranged from 2018 to 2020, which was not until the very recent. Were you able to check if there has been any new achievement available in the recent six months? Besides, I understand that there might not be a lot of works on this topic in the medical field. Still, it will be great to mention it if any of the works you reviewed were medically related (like using a medical dataset, etc.)

**Reply 2:** Thank you for pointing out this problem in our manuscript. We searched a lot of literature published after 2020, and finally found an article related to our research direction (White-Dzuro CG, Schultz JD, Ye C. et al. Extracting medical information from paper COVID-19 assessment forms. Appl Clin Inform 2021). White-Dzuro et al. used template matching to extract the information in the paper COVID-19 assessment forms. While their system performs well, it only works for a single template image. Compared with their method, our system is more suitable for complex scenarios.

**Changes in the text:** We added the content, "**White-Dzuro et al. developed an optical mark recognition/optical character recognition system to extract medical information from paper COVID-19 assessment forms. The system scans a fixed area in the image by aligning the template to identify textual information, which only works for a single template. Different from the approach of White-Dzuro et al., our system uses a deep learning approach to identify important information areas in the report. Therefore, our system is more robust and suitable for recognizing medical report images of various structures**" (see Page 22, line 437).

**Comment 3:** Differentiable Binarization (DB) seems an important part of your model. I suggest you add more details about it to help the readers understand how it works and why it renders good results.

**Reply 3:** Thank you for the above suggestion. Based on your request, we have added a comparison of the traditional method with the DB algorithm and given the reasons why

the DB algorithm has advantages.

**Changes in the text:** We added some details, "**After obtaining the probability map, the traditional method calculates the binary map through a fixed threshold. The DB algorithm predicts the threshold of each position in the image through the network to generate a dynamic threshold map. The results calculated according to the dynamic threshold are more suitable for complex and changeable detection scenarios**" (see Page 13, line 262).

**Comment 4:** In the evaluation section, results on AP50 were given to demonstrate the performance of YOLOv3-MobileNet. Adding a few more AP with different thresholds could be more convincing.

**Reply 4:** Thank you for the above suggestion. In the evaluation part, we use AP50, AP75, and FPS to demonstrate the performance of YOLOv3-MobileNet. In terms of AP75, the performance of the YOLOv3-MobileNet is not the highest, but only 0.2% lower. And the speed advantage of the model is huge.

**Changes in the text:**

a. We modified the content, "**In terms of AP50, the YOLOv3-Mobilenet is higher than the Faster RCNN and the original YOLOv3. In terms of AP75 (average precision at IoU=0.75), the YOLOv3-Mobilenet is only 0.2% lower than the original YOLOv3. However, the YOLOv3-Mobilenet model is several times faster than the other two**" (see Page 17, line 341).

b. We modified the form (see Page 29, line 524).

Table 1. Comparison of table detection algorithms

| Detection algorithm | AP50/% | AP75/% | Test time of a single image per/ms |
|---|---|---|---|
| Faster RCNN | 96.5 | 94.1 | 29.50 |
| YOLOv3 | 97.5 | 94.9 | 14.03 |
| YOLOv3-Mobilenet | 97.8 | 94.7 | 5.82 |

**Comment 5:** The evaluation metrics seem robust generally. However, it may be worth it to discuss what could potentially impact the performance of your model. Surely some parts of your model trained on other datasets might not work the best for the current dataset, but I'm wondering if you have done any case analysis on some bad examples (with low accuracy) and see where they failed. E.g., given a specific paper-based record, you may investigate which elements from the image were not correctly recognized.

**Reply 5:** Thank you for the above suggestion. We drew Figure 8 to show bad examples and discussed their failure reasons and solutions in the text.

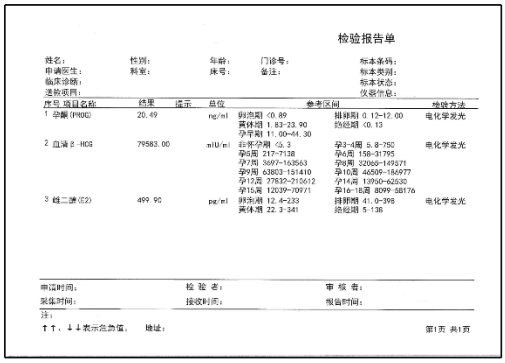
**Changes in the text:**

a. We added some content, "**DeepSSR shows excellent performance in the structured recognition of most UPBMR images (Figure 7). However, DeepSSR does not perform well for recognizing images with multiple lines of text in a cell (Figure 8)**" (see Page 18, line 359).

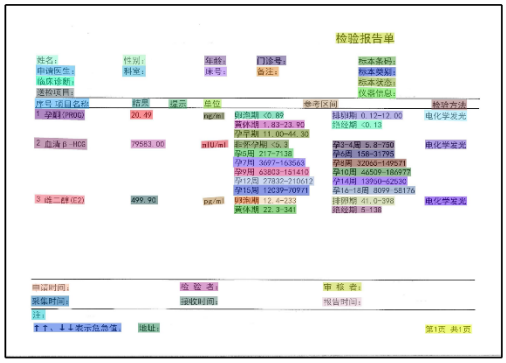
b. We added some discussion, "**3) DeepSSR performs poorly for images with multiple lines of text in one cell. Although these images are less numerous in real life, their medical value is equally important. In the future, we will train a table structure recognition model to recognize cells in a table, and then merge multiple lines of text in the cells**" (see Page 23, line 456).

c. We drew Figures 8 (see Page 28, line 519).

A




B




C

```
{'project': '1孕酮（PROG）', 'result': '20.49',    'unit': 'ng/ml',  'reference': '排卵期0.12-12.00', 'method': '电化学发光' }
{'project': '2血清B-HCG',    'result': '79583.00', 'unit': 'mIU/ml', 'reference': '孕3-4周5.8-750',   'method': '电化学发光' }
{'project': '3雌二醇（E2）', 'result': '499.90',   'unit': 'pg/ml',  'reference': '排卵期41.0-398',   'method': '电化学发光' }
```

Figure 8. An example of a poorly processed document. (A) is the input image. (B) is the character recognition result. (C) is structured data.

**Comment 6:** There are also some small issues with language usage. Most of them land in the usage of articles and commas.

**Reply 6:** Thank you for your comments. We have thoroughly checked and corrected the grammatical errors and typos we found in our revised manuscript.

**Changes in the text:** Due to grammatical problems, many revisions have been made. For details, please refer to the yellow part of the text.