Paper Review

Weighted Combination of Per-frame Recognition Results for Text Recognition in a Video Stream

Project 11: Vital Monitor and ID Detection through
Machine Vision for Improving EMS Communication Efficiency

Robert Huang
Biomedical Engineering Senior
rhuang22@jhu.edu

April 2021
EN 601.456 Computer Integrated Surgery II

Contents

1 Project Summary

Smart glass technology has only just recently been introduced into the field of emergency medicine and currently only serves to provide remote video feed to other healthcare professionals, either to obtain remote advice or as a form of training recruits. Consequently, there is opportunity to use the live feed, along with AI or other vision algorithms, to streamline certain processes faced in emergency medicine. In particular, the two main objectives of Project 11 are to use the camera feed from smart glasses to, firstly, grant remote doctors visual access to the information available on medical devices (such as ultrasound) in order to facilitate physician guidance, and to, secondly, automatically extract key information from standard documents like driver's licenses so that medics may spend less time on paperwork and more time on treating patients.

2 Paper Selection

The paper I have chosen is:

O. Petrova, K. Bulatov, V. Arlazarov, V. Arlazarov, Weighted combination of per-frame recognition results for text recognition in a video stream, Computer Optics. 45 (2021) 77–89. doi:10.18287/2412-6179-co-795.

The algorithms designed to solve the project's objectives rely extensively on an accurate Optical Character Recognition (OCR) technique to extract and classify the textual elements from both identification and vitals monitors. Unfortunately, even state of the art OCR techniques have significant accuracy drops in terms of classification of letters and words if the environment around a textual object is nonoptimal, such as in the presence of poor lighting, uneven illumination or perspective distortion, all of which are salient considerations in the unpredictable field of emergency medicine. The paper I have chosen seeks to mediate this problem.

In the paper, the authors describe a novel 'frame weighing' technique that takes the per-frame Optical Character Recognition (OCR) textual results of non-ideal video feed, specifically that taken from a mobile camera, then adds weights to the results in order to more accurately extract the textual elements from a document. Using video, it is possible to incorporate a type of 'voting' system, whereby each captured frame, taken in slightly different conditions by the user moving the desired document or the camera, can contribute to more accurately classify text. The authors then assess this technique with different parameters by using the MIDV-500 and MIDV-2019 dataset of identity documents captured with a mobile device camera in nonoptimal conditions. Ultimately, their results indicate marked improvement in recognition.

3 Paper Overview

1. **Background**

   The authors note that previously sophisticated technology, such as smartphones are becoming more commonplace. The demand for the use of OCR, which can save time, money, and effort from documentation for companies and individual persons, has also increased. Consequently, most documents are now captured not with specialized equipment, but rather mobile phones in uncontrolled environments. Yet, as OCRs are increasingly used for identity documentation, errors caused by uneven lighting or altered perspectives are becoming more costly. Still, one advantage that mobile phones have yet to take into account is their ability to take videos, which can grab multiple images of the same document. This makes it possible to incorporate different illuminations, angles, and focus characteristics into the text recognition algorithm, thus allowing one to reduce the sporadic errors of an OCR-system.

   The authors then describe the general scope of the project. They first describe the OCR flow, why each step is taken, and common techniques used to perform the step: preprocessing, text-field localization, segmenting string images into characters, character recognition, and finally post-processing. They then indicate that there exist two different categories of multiple frame: image combination or recognized text based. Image combination involves creating a high-resolution image by blurring and distorting together multiple lower resolution frames. However, as these techniques require significant camera orientation accuracy, image combination is difficult for mobile devices. Therefore, the authors will investigate the recognized text based multiple frame incorporation technique. An example of the technique, called ROVER (Recognizer Output Voting Error Reduction), has two steps: align the output text so that each detected text is paired with its respective text from other images, and then use a voting procedure to select the best characters. The authors then ponder if a weighting system could be added to discard unreliable frames and keep valuable frames.

2. **Error Analysis**

   The authors describe the different sources of error for document recognition, which are split between physical difficulty and recognition-stages difficulty. Physical difficulty, noted when even a human cannot read the text, is characterized by glare or defocused images. Glare, which cuts off characters, can be resolved through the voting system. Defocused images, which, when unweighted can add nonsensical 'votes', need to be discarded. Recognition-stages difficulty is characterized by any stage of the OCR flow producing erroneous results. For instance, if the document localization process slightly incorrectly identifies the boundary of the document, the text will be skewed.

3. **Problem Statement**

Here, the authors set up how they will think about the problem. Every character, x, is represented as a vector of probabilities, each probability indicating a certain letter.

$$x = \left(x_1, x_2, ..., x_K\right) \in [0.0, 1.0]^K, \quad \sum_{k=1}^{K} x_k = 1,$$

Figure 1: Character representation

Every text, or string of characters, is therefore represented as a 2D matrix, where each character estimate is stacked on top of one another.

$$X = \left(x_{jk}\right) \in \left[0.0, 1.0\right]^{M \times K}, \quad \forall j : \sum_{k=1}^{K} x_{jk} = 1,$$

Figure 2: String representation

The ROVER method is then described. We begin with an empty column vector or the first recognized result. When two recognized texts are obtained, due to variable length recognition, the texts are first aligned using the distance equation below. In essence, the algorithm aligns the text to maximize agreeability.

$$\rho\left(x^1, x^2\right) = \frac{1}{2} \sum_{k=0}^{K} \left| x_k^1 - x_k^2 \right|$$

Figure 3: Distance formula to align strings

Then, the algorithm combines the two texts using the weighted equation below to produce a resulting estimation of the desired text.

The authors explain that these weights should be proportional to the quality of the image and then explain that the problem is to find a way to obtain these weights.

4. **Weighting Model**

The authors rationalize why using all frames, rather than using only the ideal frames, is a good idea: there may not exist any ideal frames, and combining the recognition results

$$r = \left(r_k\right) \in \left[0.0, 1.0\right]^{K+1}, \quad \forall k : r_k = \frac{x_k^1 \cdot w(x^1) + x_k^2 \cdot w(x^2)}{w(x^1) + w(x^2)},$$

Figure 4: Combined String Formula

can give the correct result, such as in the case of a 'sliding highlight'. Still, the authors create the below model, which will firstly order the frames from the best to the worst quality (according to w), and then keep the best t results by zeroing the weight of the worst frames.

$$\pi(i) < \pi(j) \Leftrightarrow w(I_i(\overline{X}), X_i) \geq w(I_j(\overline{X}), X_j)$$

Figure 5: Ordering based on Weights/Quality of frame

$$w_i^{(t)} = \begin{cases} w(I_i(\overline{X}), X_i), & \text{if } \pi(i) \leq t, \\ 0, & \text{if } \pi(i) > t. \end{cases}$$

Figure 6: Weights with Threshold

5. **Weighting Criteria**

   Here, the authors describe how w, the weights, will be determined. The first criterion looks at the x,y,xy,and yx gradients, and, based on these values, determines how in-focus the image is.

$$\begin{aligned} G_{\mathrm{r,c}}^{V}(I_i(\overline{X})) &= \left|\mathrm{I}_{r+1,\mathrm{c}} - \mathrm{I}_{\mathrm{r,c}}\right|, \\ G_{\mathrm{r,c}}^{H}(I_i(\overline{X})) &= \left|\mathrm{I}_{\mathrm{r,c+1}} - \mathrm{I}_{\mathrm{r,c}}\right|, \\ G_{\mathrm{r,c}}^{D_1}(I_i(\overline{X})) &= (1/\sqrt{2})\left|\mathrm{I}_{r+1,\mathrm{c+1}} - \mathrm{I}_{\mathrm{r,c}}\right|, \\ G_{\mathrm{r,c}}^{D_2}(I_i(\overline{X})) &= (1/\sqrt{2})\left|\mathrm{I}_{\mathrm{r,c+1}} - \mathrm{I}_{r+1,\mathrm{c}}\right|, \end{aligned}$$

Figure 7: Gradient Calculations

, where q(G) is a 0.95-quantile of the gradient image G. The second criterion looks at how confident a text field is by looking at the least confident predicted character's confidence level.

6. **Per-Character Weighting**

   The authors note that individual characters in a text field may contain different levels of quality, and suggest that each character should have their own weight value, based on the

$$F(I_i(\bar{X})) = \min\Big\{q(G^V(I_i(\bar{X}))), q(G^H(I_i(\bar{X}))), \\ q(G^{D_1}(I_i(\bar{X}))), q(G^{D_2}(I_i(\bar{X})))\Big\},$$

Figure 8: Focus estimation weight criteria formula

focus of the character image, and their highest confidence level. With this new addition, the authors then describe the whole weighting process, including the initiation with an empty column, when weights will be assigned, how recognized text will be combined, how weights will be propagated after combination, and how non-aligned text will be handled. They also present the detailed flow chart below.

$$Q(X) = \min_{j=1}^{M}\left\{\max_{k=1}^{K} x_{jk}\right\}$$

Figure 9: Confidence level weight criteria formula

7. **Experimental Evaluation**

The authors first use the full-string weighting model on the datasets MIDV-500 and MIDV-2019, which respectively contain 500 smartphone videos of identity documents without significant distortion and 200 smartphone videos of identity documents with significant distortion. Per video, 30 frames that contained the whole document were considered and only the text fields of document numbers, numeric dates, latin name components, and machine-readable zone lines were read. Experiments were conducted with ten different parameter groups. Five were run with the focus weight criterion, and five were run with the confidence level weight criterion. Those five runs were separated into no weighting, choosing the best result, weighting the three best results, weighting the top 50% of results, and weighting every frame. The author then presents many graphs and tables that compare how changes in the use of the criterions and the number of frames weighted affect the final accuracy on both datasets. It becomes evident that the focus criterion is the better of the criterions, and that weighting the top 50% of the top results is the best amount of frames to use.

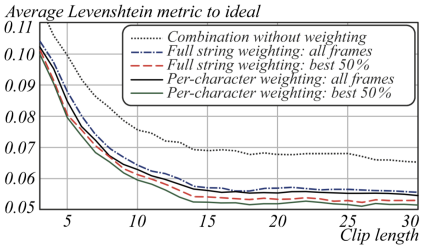

Fig. 14. Performance profiles for focus estimation weighting on MIDV-500

Figure 10: Performance Profile for Focus Estimation Weighting on MIDV-500 [Figure 14, 1]

Table 5. Mean Normalized Levenshtein metric distance to the correct result on MIDV-500 using focus estimation

Combination method	Mean Normalized Levenshtein metric					
	5 frames	10 frames	15 frames	20 frames	25 frames	30 frames
Without weighting	0.0995	0.0756	0.0689	0.0677	0.0680	0.0652
Full string weighting: all frames	0.0879	0.0643	0.0570	0.0569	0.0565	0.0555
Full string weighting: best 50%	0.0804	0.0612	0.0541	0.0533	0.0529	0.0529
Per-character weighting: all frames	0.0847	0.0628	0.0561	0.0553	0.0552	0.0545
Per-character weighting: best 50%	**0.0795**	**0.0595**	**0.0524**	**0.0518**	**0.0516**	**0.0515**

Figure 11: Mean Normalized Levenshtein metric distance to the correct result on MIDV-500 using focus estimation [Table 6, 1]

The authors then used the per-character weighting model on the datasets, while only using the focus weight criterion. Experiments were conducted with five categories: no weighting, full string weighting with all frames, full string weighting with the best 50% of frames, per-character weighting with all frames, and per-character weighting with the best 50% of frames. The author notes that the per-character weighting with 50% of frames achieves the best performance.

8. Discussion

   The authors discuss how their results show that weighted frames, focus estimation weight criterion, per-character weights, not using only the few top frames, and not using every frame results in better recognition of textual features, and why this is so. The author then discusses how the confidence level criterion performed poorly due to how lost characters (due to highlighting) in a string would actually increase the overall weight by removing the lost characters. Finally, the author clarifies that per-character weighting only achieves significantly better results than full string weighting when there is a significant distortion affecting a text field.

9. Conclusion

   The author concludes that the combination of the best 50% frames with per-character weighting and focus weighting will result in the best performance when integrating multiple frames of a text field. The author states that in the future, they plan to explore other weighting criteria, explore how deep learning may be used as a technique for frame integration, and evaluate how this model could be used for other applications.

4 Critiques

Overall this paper is very comprehensive and detailed in its description of a weighted model for OCR post-processing with video. There were many figures that helped illustrate presented points, and the graphs and tables were easily comprehensible. The flow however was somewhat stagnant. Though they give a brief outline after the background, the descriptions given serve more to describe a section in isolation rather than in the flow of the paper.

In fact, the paper introduces a lot of background and many existing techniques in easily understood vocabulary and models. Therefore, it serves well as an introductory paper for OCR. However, the main innovation in this paper is the addition of Per-character weighting to their previous method [2]. However, this point is not mentioned prior to section 5 and therefore, the paper does not describe their main point until very late into the paper.

Other critique is that the results of some papers are referenced in a conclusive form, but explicit numbers or percentages are never produced. For instance, when referencing their other paper "Methods of weighted combination for text field recognition in a video stream Proc", it is stated that "the weighted combination actually improves the recognition quality", but no numbers are given for comparison to this paper's proposed technique.

With regards to content and experimentation, the paper should have looked at more fields on the identification documents. The fields that are presented in the experimental section are generally the largest and most distinct fields. It also would have been good to look at how the algorithm affected characterization of other languages. The paper also only presents two criterion weights. Finally, the experiments do not consider how failed localization errors or other documentation recognition errors (rather than physical distortion) may impact the weighting algorithm.

5 Key Takeaways

During my implementation of the model described, it would be best to not only model the recognized text in the matrix fashion described, but also incorporate the conclusions made in the paper to achieve the best accuracy. In particular, I should use the best 50% frames, use per-character weighting, and use the focus weighting criterion.

6 References

- O. Petrova, K. Bulatov, V. Arlazarov, V. Arlazarov, Weighted combination of per-frame recognition results for text recognition in a video stream, Computer Optics. 45 (2021) 77–89. doi:10.18287/2412-6179-co-795.
- O. Petrova, K. Bulatov, V.L. Arlazarov, Methods of weighted combination for text field recognition in a video stream, Twelfth International Conference on Machine Vision (ICMV 2019). (2020). doi:10.1117/12.2559378.