A Self-Enhancing Approach for OCR

Introduction: In today's world finance is a fastest growing sector. A lot of financial activities take on the daily basis, the data related to these activities is immense. Which the help of advance technology these financial services has become faster and efficient, but still there are a lot of services and creditworthiness business in which physical documentations are used. Since it took time to verify those document manually thus adversely affecting service delivery. So, an alternate solution to this problem is automation. A lot of work is done on this problem but every solution got limited to standardization. This model will solve both these problems, it not only perform an effective OCR but also create a self-enhancing backbone to itself to overcome non-technical term's problem. This model works in two blocks, 1st block that will take the image of the document stored in database and extract the information which is technical and separate out the non-technical block for 2nd block to recognize. This modules uses multiple text extraction & recognition algorithms for this purpose in order to reduce errors and also uses prioritization process to manage effectiveness. For second layer, which is non-technical recognition, will use a dictionary approach for identification.

Objective: In this model, first we will design a UI which will keep track of events which will take place during entire process, logs of these event will be kept in records (data will be collected in standard format) and UI will keep showing the results and process execution. Second, we will develop several enhancement techniques for images (in case images are blurry, less clear) to improve the quality. This will help us in effective recognition. Now, several text detection algorithms will be implemented for recognition purpose. We will use multi-classification to take out the final text. For second block to work, we will develop a global dictionary that will collect the non-technical terms from this document and will be stored for a technical standardization. Each time a new non-technical terms appear in document, it will separate this terms and will recognize the rest, will keep that data for standardization purpose. So, if in future anything related to that terms appears, it will automatically been taken care of.

Steps Involved:

1. The 1st includes designing an UI. This UI will shows information related to the current document which will be tested by our model. UI will include several functionalities which will ease its use for users. This UI can be modified according to its usages and requirements.
2. Now, we will implement several image enhancement techniques for better classification and recognition. This process includes but not limited to- filtering, skew-correction, binarization etc.
3. This step is for separating standardized textual-information from non-standardized one. This process will take these two information apart, 1st information will be used to complete existing standardized data while other will be saved in dictionary (a dictionary of non-technical standardized data) to be modified and added as a new entry to an standardize database.
4. We will use multiple OCR algorithms for this purpose, matching results and deciding final text information will be done, based on multi-class classification. Some Examples- OCR, MSER, Edge & blocks methods etc.
5. Now, we will collect non-standardized data and create new standard attribute which will cover this information. After all information collected, verification steps will be carried out to authenticate the information.

Benefits:

1. UI will ease the use this model, logs will helps to keep track of process. This model can also be implemented in system that organization is using. UI will keep the information of validating textual info of standardized and non-standardized separately, which will helps us to find out the reason of validation failure, and false negative can be improved easily. If it validate successfully then new standardized attribute (coming out of dictionary data) can be added to old standardized database.
2. Multi-class classification and use multiple recognition techniques will reduce errors in confusion matrix thus high accuracy. This model can enhance any standardized database based on the type of line of business this data is used.
3. Use of dictionary will help us to add more technical terms in standardized database.

And list goes on……

Please have some reference to my own two project in recognition (text and image) PROJECT1 and PROJECT2 , (Note: project1 might show some insufficient data, sorry for such case).