OCR Performance Prediction using a Bag of Allographs and Support Vector Regression

Tapan Kumar Bhowmik *, Thierry Paquet * and Nicolas Ragot †
* *LITIS EA-4108, Université de Rouen, France*
Email: tapan-kumar.bhowmik@univ-rouen.fr, Thierry.Paquet@univ-rouen.fr
† *LI EA-6300, Université François Rabelais Tours, France*
Email: nicolas.ragot@univ-tours.fr

***Abstract*—In this paper, we describe a novel and simple technique for prediction of OCR results without using any OCR. The technique uses a bag of allographs to characterize textual components. Then a support vector regression (SVR) technique is used to build a predictor based on the bag of allographs. The performance of the system is evaluated on a corpus of historical documents. The proposed technique produces correct prediction of OCR results on training and test documents within the range of standard deviation of 4.18% and 6.54% respectively. The proposed system has been designed as a tool to assist selection of corpora in libraries and specify the typical performance that can be expected on the selection.**

I. Introduction

In recent years, there is a huge interest in processing of historical documents into digital format as these old documents often have historical and cultural significance. The aim is to scan them and create digital libraries, thereby offering continuous electronic access to this important part of the cultural heritage. Efficient storage, indexing, retrieval, and management of these archives are extremely important in digital library for easy access. As a part of whole digitization process, in case of historical machine printed documents, OCR technology is used to convert document images into ASCII format for underlying indexing as well as easy storage and retrieval. Though many research in OCR technology have been done during last three decades and as a result many successful commercial OCR systems are developed but still there is some restriction to use them in the real applications. In fact, current OCR systems do not always perform well especially when the documents have noisy background such as text printed against shaded or texture background and/or embedded in images, complex and dense layout, text with non-text information, many irregular shapes due to typical fonts. Because of these reasons OCR technology faces difficulties to process historical documents. As a result, there can be some pages of documents for which the transcription is unusable. Thus, there is a need for some kind of monitoring system that can predict the OCR results, before processing masses of documents [1], which is often heavy and costly, and to check whether the estimated OCR accuracy claimed by a service provider is correct or not.

In this paper, we have developed such a monitoring system to predict the OCR results. Many research have been done in similar direction to assess the quality of the document regarding OCR [2], [3], [4], [5], but not to predict the OCR results. We assume that the performance of OCR varies mainly due to the variability of font, noise, image quality and typographic problems that have a direct impact on text recognition. To predict the OCR results, we characterize the documents on the basis of these phenomena. In order to do that, we collect all kinds of distinct object patterns, called allographic components [6], [7], [8] from the corpus. The distinct object patterns are identified by a similarity measure. The distribution of these patterns is estimated in each document and presented in the form of a feature vector for describing the document. Later, a SVR technique is used to build the predictor with the feature vectors.

The rest of the paper is organized as follows. In Section II, we describe our methodology of proposed OCR performance predictor system. The experimental results and analysis are presented in Section III. Finally, conclusions and further scope of research are discussed in Section IV.

II. Methodology

In this section, we describe the four major steps that are associated for building the predictor. They are: *A.* detection of allographic components in the gray scale document images, *B.* building of an allograph library from the corpus (a bag of allographs), *C.* representation of the document with bag of allographs (document vector generation) and *D.* building the predictor with training data. The overall predictor system is illustrated in Figure 1.

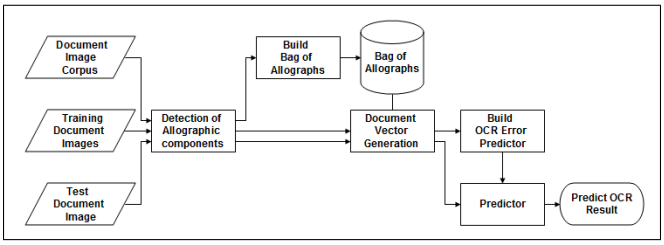

Figure 1. Illustration of entire process for OCR accuracy prediction system

A. Detection of allographic components in gray scale images

For an input document image, we detect all the primary object regions by means of allographic components that occur in the document. This process is done via an object localization algorithm. The purpose of this algorithm is to find out all the primary object regions present in the document. The object may be an individual character or part of a character or part of multiple characters or any other significant component. In order to do that, we first identify the edge pixels by Canny edge detector and then group the edge pixels based on K-connected neighbors and finally merge the grouped pixel regions on the basis of certain criteria. The value of K is usually 8, but it can be extended more say, 24, 48 etc. depending on the requirement. The higher value of $K(> 8)$ can make larger group even though all the pixels in that group are not 8-connected. There are two threshold values associated with Canny edge detector: upper and lower thresholds. The lower threshold is chosen as 0.4 times of upper threshold. We set the upper threshold (τ_1) to a very low value (e.g., $\tau_1 = 0.3$). The reason is to incorporate the weak edges also into the algorithm. Though a lower threshold introduces many noisy edge pixels but the true object edges are always detected even when the foreground to background contrast is very low. For merging the regions, we develop an algorithm which is similar to the well known connected component analysis algorithm where instead of grouping or labeling the pixels on the basis of connectivity of neighbor pixels, the regions are grouped on the basis of a distance measure of neighboring regions. The distance between two regions R_i and R_j is defined as

$$dist(R_i, R_j) = \frac{2 * area(R_i \bigcap R_j)}{area(R_i) + area(R_j)} \tag{1}$$

If the region R_i is matched perfectly with region R_j, then the distance is 1 otherwise it is < 1 and for non-overlapping case it is always 0. Two neighbors regions are said to be connected or same label in the context of merging if the following condition holds.

$$conn(R_i, R_j) = \begin{cases} 1 & \text{if } \xi < dist(R_i, R_j) \leq 1, \xi \in [0, 1] \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In particular for this study, we set ξ to 0.1 when either region R_i is the inner region of R_j or R_j is the inner region of R_i otherwise ξ is set to 0.

B. Building bag of allographs

So far, we have described the localization algorithm for locating the objects present in a document image. Each of these objects in gray scale is called an allographic component. Now we build an allographic component library, called bag of allographs. In order to do that, we group all the objects from a corpus on the basis of similarity measures by template matching. A bag of allographs is created from a single element of each group. In context of template matching, there are many similarity measures described in the literature [9]. But, in this study we use the normalized cross correlation (NCC) coefficient measure for obtaining the similarity between the object pattern images. The normalized cross correlation coefficient is defined as

$$\rho = \frac{\sum_{x,y}(I - \overline{I})(T - \overline{T})}{\sqrt{\sum_{x,y}(I - \overline{I})^2 \sum_{x,y}(T - \overline{T})^2}}, \tag{3}$$

where T is the template image, I is the larger image, $\overline{I} = \frac{1}{N}\sum I$, $\overline{T} = \frac{1}{N}\sum T$, N is the number of pixels in T and (x, y) are summing over the range of area of T. In common practice, a template image is matched against a larger image by shifting its position pixel by pixel over its possible entire search space. Instead of shifting the template image, here we normalize the size of the second image according to the template image size by wavelet transformation [10]. It is also to be noted that the normalization process is applied for matching only when both images are compatible in size. Otherwise, the input image is considered as a non-match pattern object with respect to that particular template. The two images $I_1(W_1, H_1)$ and $I_2(W_2, H_2)$, where W and H indicate width and height of an image, are said to be compatible in size if the following condition holds:

$$comp(I_1, I_2) = \begin{cases} 1 & \text{if } \frac{Min(W_1,W_2)}{Max(W_1,W_2)}, \frac{Min(H_1,H_2)}{Max(H_1,H_2)} \geq \text{AR} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where, $AR \in [0, 1]$ is the aspect ratio threshold. For building a bag of allographs, we start from a single object pattern, and add the patterns successively in the library as a new pattern group if a sufficiently good match ($AR \geq 0.25$ and $\rho \geq \rho^{'} = 0.85$) is not found comparing with the existing patterns in the library. Otherwise, the pattern is considered as a member of an existing pattern group.

C. Document vector generation

We represent a document image as a distribution of allographs. The intuition is that the frequency of occurrence of allographs can play an important role in evaluating the OCR performance. If we consider N allographs for representing the documents, then the size of the feature vector is $N + 1$. The size is increased by one to calculate the frequency of patterns that does not match with any allograph patterns. To calculate the frequency of occurrence of allographs for a document, we follow the same process as the way the bag of allographs is built. Instead of adding new pattern in the library when there is no match found, we use an additional counter to calculate the frequency of occurrence of non-matched pattern objects. For faster document vector generation, instead of matching each and every object patterns in a document with the bag of allographs, we first find all the pattern groups that exist in the document. Now for

a single candidate element in each group, we do the same matching process with the bag of allographs. The frequency of allograph patterns are calculated by incrementing the counts with the number of group members. The normalized value of the frequencies is considered as a feature vector for representing the document. Therefore, our final document vector is of size $N+1$ of the form

$$f(D) = (f_{p_1}, f_{p_2}, \ldots f_{p_N}, f_{p_{N+1}}) \quad (5)$$

D. Building OCR Performance Predictor

In the previous section, we have described how to generate a document vector for an input document using bag of allographs. Let us assume that $\mathbf{x}_i$ is the document vector corresponding to the document D_i, where $\mathbf{x}_i$ has the form $(f_{p_1}, f_{p_2}, \ldots f_{p_N}, f_{p_{N+1}})$. Let us assume that y_i is the percentage of accuracy produced by some OCR on the document D_i. Suppose we have a set of training documents $D = \{D_1, D_2, \ldots, D_n\}$. Therefore, we have a training set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_n, y_n)\}$. Now, the task here is to build a predictor or in mathematical term, we can say to find a function $g(\mathbf{x})$ that produce the target value y_i with a certain deviation for all the training samples, i.e., $y_i = g(\mathbf{x}) + \zeta_i$, where ζ_i is the noise variable. In order to find such a function, we use support vector regression (SVR) technique [11]. The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function. Let Φ be a non-linear mapping from input space to some high-dimensional feature space. For the linear regressor (in feature space) defined by $f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b$, we wish to minimize

$$L(\xi, \xi^*, \mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \quad (6)$$

subject to $(\forall i \in \{1, 2, \ldots, n\})$

$$y_i - f(\mathbf{x}_i) \leq \varepsilon + \xi_i^*, f(\mathbf{x}_i) - y_i \leq \varepsilon + \xi_i, \xi_i, \xi_i^* \geq 0 \quad (7)$$

where $\langle ., . \rangle$ denotes the dot products, ε denotes the width of the error-insensitive zone of the cost function and ξ_i and ξ_i^* are slack variables measuring the deviation of $y_i - f(\mathbf{x}_i)$ from the boundaries of the error-insensitive zone. The constant $C > 0$ determines the trade-off between the flatness of $f(\mathbf{x})$ and the amount up to which deviations larger than ε are tolerated. Now this primal problem is converted into the dual quadratic optimization problem where two Lagrange multipliers α_i and α_i^* are introduced and calculated. Finally, we obtain a support vector regression of the form

$$f(\mathbf{x}) = \sum_{i}^{n}(\alpha_i - \alpha_i^*)(\mathbf{x}_i.\mathbf{x}) + b \quad (8)$$

and for non-linear mapping the regression equation becomes

$$f(\mathbf{x}) = \sum_{i}^{n}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i.\mathbf{x}) + b, \quad (9)$$

where $K(\mathbf{x}, \mathbf{x}^{'}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}^{'}) \rangle$. The details derivation in each step and further computing the value of b are described in the literature [12].

III. Experiments

In our experiment, we use a database of historical documents which contains works on art and humanity from French National Library (BnF). For indexing and searching of textual information of digital documents, as a part of digitization process, Optical Character Recognition (OCR) technology is used to convert textual contents into ASCII format. This conversion process is carried out by OCR providers. Such OCR outputs are used in this experiment to build the predictor as well as evaluate the prediction results. In the present study, we have used 22 books that include 1015 pages. In the predictor system, we use a set of threshold parameters $\lambda = \{\tau_1 = 0.3, K = 8, \xi = 0.1, \rho^{'} = 0.85, AR = 0.25\}$ in the different stages of our algorithms as described before. But these are not much sensitive in the sense that it is set prior to the system configuration and does not vary from document to document.

For developing the bag of allographs, we select manually 250 pages from all the books. The pages are chosen in such a way that it includes almost all the variety of documents such as font style, font size, background that exist in the corpus. Here, the size of the feature vector depends on the number of distinct allographs considered. So, there is always an open issue in deciding what would be the optimum number of distinct allographs. It is obvious that the computational cost grows with the increasing number of distinct allographs in the bag. For selecting the optimum number of allographs, a straight-forward approach is to choose all the distinct patterns that occur in the corpus. But in practice, it is found that many patterns occur only once in the whole corpus, which cannot play any significant role to predict the OCR results. These are considered as the noise patterns. After removing all these noise patterns from the bag, a total of 2112 significant distinct allographs are found in the corpus of 250 document pages. However, we observe that the frequency of many allographs among 2112 are still very low compared to others and these allographs are likely to be less significant. In this stage, we introduce another user-defined parameter p, called corpus coverage probability to determine the number of significant allographs to be included. Here, our assumption is that allographs with higher frequencies have more significance. Different values of the corpus coverage probability (p) versus the number of determined distinct allographs (N) are plotted in Figure 2. From the Figure 2, it is found that if we select 98% ($p = .98$) object patterns that occur in the whole corpus, then we only need to include $N = 724$ number of most significant distinct allographs in the bag.

To generate a feature vector for a document, our localization algorithm first finds all the object regions in the

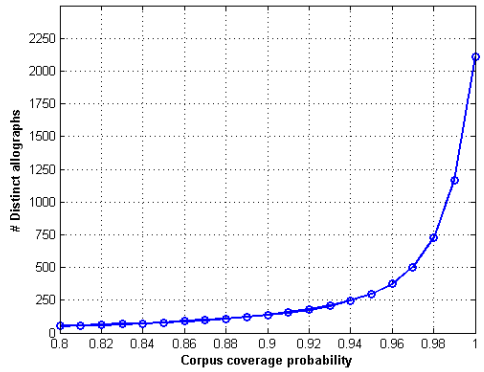

Figure 2. Corpus coverage probability vs number of distinct allographs

document. Such object regions produced by our algorithm are shown in the Figure 3. Next, we generate a feature vector

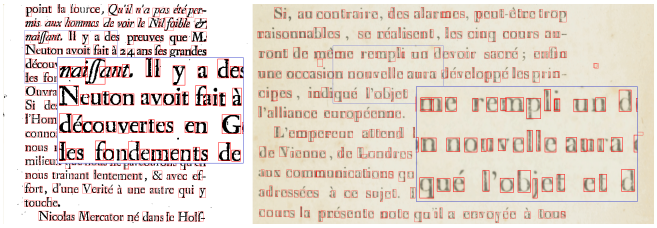

Figure 3. Object localization are shown for two pages from different Books. Left side page has fair background whereas right side has complex background.

for the document which is accomplished by matching all the localized objects with allograph patterns. While generating the document vector, the less-significant patterns are put together into a separate bin as a non-match pattern. For this experiment we choose $p = .98$. So, our document vector size is $N+1 = 725$. Few such significant allograph patterns among them are shown in the Figure 4.

Figure 4. Most and least significant 45 allograph patterns in each are shown from the bag of 724 allographs.

For building the predictor, we consider 504 document pages (around 50% of the whole corpus) from different variety of documents that includes wide range of OCR accuracy from 0 to 99%. The rest of the document pages (511 pages) are used to evaluate our prediction results. It is also to be noted that for a blank document, the target accuracy is set to 0. To build SVR predictor, we select a polynomial kernel of degree 3 for non-linear mapping. The optimal learning parameters (C, ε) for polynomial kernel of SVR are selected using 5-fold cross validation experiments with grid search method on the training set. The experimental result for searching parameters is shown in Figure 5, where the minimum average root mean squared error (RMSE) is found to be 4.22. Finally, we train the SVR predictor model

	$\varepsilon = 0.01$	0.5	1	1.5	2
$C = 0.001$	8.09096	8.09537	8.08012	8.05134	8.01924
0.01	5.22991	5.21674	5.1919	5.27069	5.29366
0.5	4.23876	4.22618	**4.22291**	4.24198	4.29372
1	4.37193	4.33886	4.30934	4.31934	4.40169
2	4.57667	4.60476	4.57254	4.54984	4.56416

Figure 5. The average RMSE values obtained with grid search method are shown. Based on 5-fold cross validation error, the best parameters are $(C = 0.5, \varepsilon = 1.0)$.

by setting the optimal parameters $(C = 0.5, \varepsilon = 1.0)$. The prediction results on the 504 training document pages are shown in the Figure 6, where the average root mean squared (RMSE) error is 4.18.

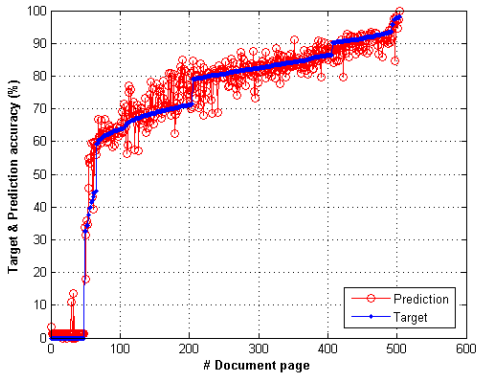

Figure 6. Prediction of OCR result on 504 training document images with SVR model. Here, the average root mean squared error (RMSE) of prediction is 4.18. Blue color indicates the target OCR accuracy value and red color indicates the corresponding prediction value.

To evaluate/test our model, we use the remaining 511 $(1015 - 504 = 511)$ documents. The prediction results on these test documents are shown in the Figure 7, where average root mean squared error (RMSE) is 6.54. We further analyze the prediction results by calculating the RMSE value in different ranges of true OCR accuracy values (see in Table I) and find that the precision is always better in higher accuracy documents and it is minimum (RMSE = 3.12) for the documents with accuracy ranging from 90 to 100%. This

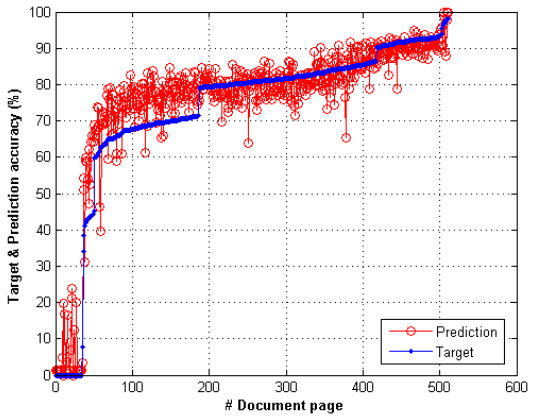

Figure 7. Prediction of OCR result on 511 test document images with SVR. Here, the average root mean squared error (RMSE) of prediction is 6.54. Blue color indicates the target OCR accuracy value and red color indicates the corresponding prediction value.

Table I
PIECE-WISE MEAN SQUARED PREDICTION ERROR ON TEST DOCUMENT IMAGES

Accuracy (%)	**0-60**	**60-80**	**80-90**	**90-100**	**Average (0-100)**
RMSE	11.59	7.86	4.09	3.12	6.54
# Test Samples	51	177	189	94	511

result is interesting because people might be more interested to predict the results with high confidence for the documents which have accuracy range from 80 to 100% rather than for those which have accuracy below 80%.

IV. CONCLUSION

In this paper, we have described an efficient as well as a simple method for predicting OCR results. We introduce corpus coverage probability to select the size of the bag of allographs. In this context, we assume that the patterns which have higher frequency in the corpus indicate higher level of significance. Instead of considering the significance of patterns based on the frequency, there is also another possibility to find significant patterns in a discriminative way where we can choose the patterns which are significant for a particular document but not others. In this context, tf-idf measurement can be used to identify the most significant patterns. From the experiment, it is shown that our predictor gives good results but still there are many possibilities to improve the performance of the system. While building the allographs library we have considered only 250 pages but including more and more document pages can lead to a robust library which can represent a wide variety of documents in more significant way. Though, increasing the size of the library can make the system computationally expensive, but this problem can be overcome by incorporating a hashing technique where we can reduce the search space in a very efficient way. For building the predictor, we use only a single OCR, but training with multiple OCR results can make the predictor system more robust. So far, we have assumed that the accuracy of OCR varies due to the typography, typicality of fonts, noise, image quality, complex and dense layout but there are many other reasons say, mixing non-text information with text, skew, use of language model etc. that affect the OCR results. The errors get accumulated from different levels of whole processing chain in OCR into final output. So, a complete predictor system that evaluates errors from all kinds of difficulties in different levels of OCR is a new research problem in near future.

REFERENCES

[1] A. B. Salah, N. Ragot, and T. Paquet, "Adaptive detection of missed text areas in ocr outputs: application to the automatic assessment of ocr quality in mass digitization projects," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013, pp. 865 816–865 816.

[2] J. H. Michael Cannon and P. Kelly, "Quality assessment and restoration of typewritten document images," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 2, pp. 80–89, 1999.

[3] H. S. Baird, "The state of the art of document image degradation modelling," in *Proc. of 4th IAPR Int. Workshop on Document Analysis Systems, Rio de Janeiro*, 2000, pp. 1–16.

[4] T. Kanungo, R. M. Haralick, H. S. Baird, W. Stuezle, and D. Madigan, "A statistical, nonparametric methodology for document degradation model validation," *IEEE PAMI*, vol. 22, no. 11, pp. 1209–1223, 2000.

[5] L. Likforman-Sulem, J. Darbon, and E. H. B. Smith, "Enhancement of historical printed document images by combining total variation regularization and non-local means filtering," *Image Vision Comput.*, vol. 29, no. 5, pp. 351–363, 2011.

[6] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE PAMI*, vol. 29, no. 4, pp. 701–717, 2007.

[7] A. Bensefia, T. Paquet, and L. Heutte, "A writer identification and verification system," *Pattern Recognition Letters*, vol. 26, no. 13, pp. 2080–2092, 2005.

[8] S. Marinai, B. Miotti, and G. Soda, "Bag of characters and som clustering for script recognition and writer identification," in *Proc. of 20th Int. Conf. on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 2182–2185.

[9] R. Brunelli, *Template matching techniques in computer vision: theory and practice*. Wiley. com, 2009.

[10] T. K. Bhowmik, P. Ghanty, A. Roy, and S. K. Parui, "Svm-based hierarchical architectures for handwritten bangla character recognition," *IJDAR*, vol. 12, no. 2, pp. 97–108, 2009.

[11] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.

[12] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.