The Impact of OCR Accuracy and Feature Transformation on Automatic Text Classification

Mayo Murata, Lazaro S.P. Busagala, Wataru Ohyama,
Tetsushi Wakabayashi, and Fumitaka Kimura

Mie University, Faculty of Engineering,
1577 Kurimamachiya-cho, Tsu-shi, Mie 5148507, Japan
{mayo, busagala, ohyama, waka, kimura}@hi.info.mie-u.ac.jp

Abstract. Digitization process of various printed documents involves generating texts by an OCR system for different applications including full-text retrieval and document organizations. However, OCR-generated texts have errors as per present OCR technology. Moreover, previous studies have revealed that as OCR accuracy decreases the classification performance also decreases. The reason for this is the use of absolute word frequency as feature vector. Representing OCR texts using absolute word frequency has limitations such as dependency on text length and word recognition rate consequently lower classification performance due to higher within-class variances. We describe feature transformation techniques which do not have such limitations and present improved experimental results from all used classifiers.

1 Introduction

In recent years, the main means of information exchange has been changing from the traditional printed information to the digital data. This is due to the fact that digital data such as text, image, and audio can be transferred and retrieved faster, more flexibly and more easily. Activities such as digital publishing and the digital library might become the main sources of information in the near future. As the matter of fact digitization projects have been taking place [7], [8]. Since there have been a need to make archives accessible through digital information systems, other traditional libraries might be considering converting printed archives into digital data. Digitized materials might need techniques from automatic text classification (ATC) to be applied to different domain applications such as automatic indexing for Boolean information retrieval systems; document organization; information filtering and hierarchical categorization of web pages.

When working with printed documents there might be two ways to generate digital texts which are keying texts into computer system and using optical character recognition (OCR) systems where by text materials are extracted from digital text images. LDI project team [7] argues that Harvard University Library keying process cost approximately 10-13 times more expensive per page than using uncorrected OCR. They refer to uncorrected OCR due to the fact that OCR-generated texts generally have errors [1], [2]. The authors in [3] showed the impact of OCR accuracy on automatic text classification such that as OCR recognition rates dropped down, the classification

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 506 – 517, 2006.

performance decreased. In this paper, we describe feature transformation techniques for OCR-generated texts and present improved experimental results from all used classifiers.

This paper is organized as follows. The next section presents a brief survey on relevant research works in the literature. Section 3 describes the feature transformation techniques used. In section 4 we present the experimental setup. The results and a short discussion are given in section 5. We conclude and describe future study in section 6.

2 Related Works

This paper describes techniques for transforming features from OCR-generated documents. The literature shows rare research works done previously on OCR in relation to automatic text classification (ATC). This section gives a brief survey from research works that might be relevant.

The work in [9] reports on OCR text representation for learning with a focus on different techniques for automatic construction of relevant features from Germany language documents. Their study considered various features including all words, elimination of stop-words, morphological and composite analysis and use of *n*-grams. Although some important results are given, the fact that they used different language datasets, their work is remarkably different in various ways. Not only didn't they perform feature transformation techniques but also they didn't use the benchmark collection to text categorization from which we generated image text documents to study the impact of transformed features on OCR-generated documents.

Frasconi et al. [10], [11] performed experiments on text categorization for multi-page documents extracted by an OCR system. They used contrarily untransformed word counts i.e., bag of words to represent the texts. They also used information gain technique for feature selection to reduce the number of features hence dimension reduction. However we employ principal component analysis (PCA) after using term selection techniques for dimension reduction.

The authors in [3] investigated the impact of OCR accuracy on automatic text classification using absolute word frequency as an OCR text representation technique. Since absolute frequency depends on text length we give techniques to solve this problem (see section 3).

Most of the works (*if* not all) mentioned above and that in [12], exhibit notable differences with this paper. The biggest difference is that they reported experimental results from OCR texts represented by untransformed features. Hence we focus on transformed features for representing OCR texts. Experimental results reveal improved classification performance.

3 Feature Transformation Techniques

3.1 Normalization to Relative Word Frequency

The limitation of using absolute word frequency is dependency on text length consequently lower separability of feature space. Fig. 1 shows how the feature vectors length differ for a given sample distribution of absolute word frequency. Relative word

frequency y_i in expression (1) does not depend on text length such that the within-class variances are smaller than for absolute word frequency hence feature space can gain more separability.

$$y_i = \frac{x_i}{\sum_{i=1}^{n} x_i} \tag{1}$$

Whereby, x_i is the absolute word frequency of word i and n is the number of different words. Fig. 1 shows how the feature vectors have smaller variations in lengths after feature transformations.

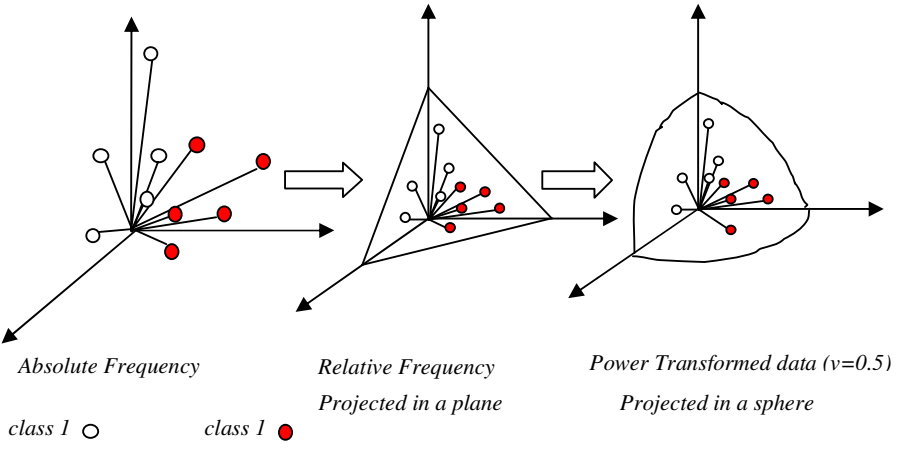


Fig. 1. Shows how the absolute frequency can be converted to relative frequency then to power transformed features in 3-dimensional space. When the feature is transformed, vector length variation in the category becomes smaller than the absolute frequency. The separability of the feature space increases hence classification rates can be expected to be improved too.

3.2 Power Transformation

Another variable transformation is the power transformation which is expressed as:

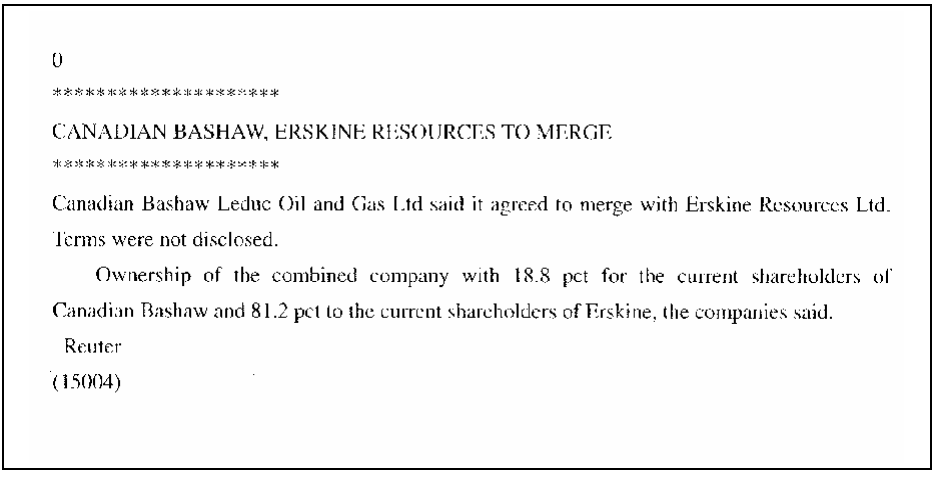
$$z_i = x_i^v \quad (0 < v < 1) \tag{2}$$

This was employed to improve the classification accuracy. Power transformation improves the symmetricy of the distribution of the frequency $x_i \geq 0$ which is noticeably asymmetric near the origin. The final effect is to improve performance of parametric classifiers derived on Gaussian assumption.

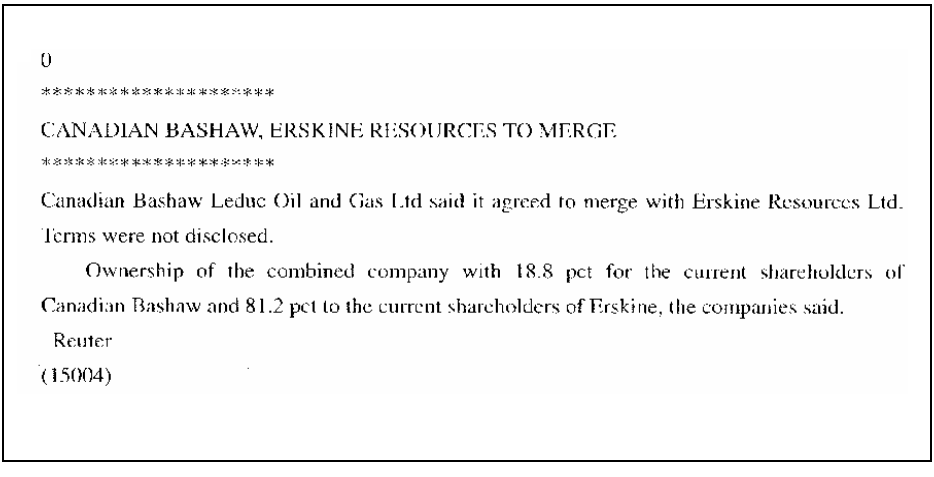
4 Experimental Setup

4.1 Used Data

In order to study the impact of transformed features for OCR-generated documents in the automatic text classification, a training sample is required. Therefore, we used the Reuters-21578 text benchmark collection for English text classification. The Reuters-21578 is composed of 21578 articles manually classified to 135 categories.



(a) 300 dpi



(b) 140 dpi

Fig. 2. Examples of the text images

In the experiments, a total of 750 articles i.e., 150 articles per category randomly selected from five categories (acq, crude, earn, grain, trade), were used. Since the sample size is not large enough, the sample was divided into three subsets each of which included 50 articles per category. When a subset was tested, the rest of the two subsets were used as learning sample in order to keep the learning sample size as large as possible while keeping the independency between the samples for learning and testing. Classification tests were repeated for three subsets and the average classification rates were computed.

4.2 Experiments

Generally there are three steps to be followed in the experiments. These include text image generation, OCR text generation and the automatic text classification. The followings are descriptions of these steps.

4.2.1 Text Image Generation

Textual documents from Reuter's collection were printed out. Paper texts were digitized using a scanner into images of different resolutions including 300 dpi, 200dpi, 150dpi, 145dpi, 140dpi, 135dpi and 130dpi. Figure 2(a) and (b) show the example of the text images of 300dpi and 140dpi respectively.

CANADIAN BASHAW, ERSKINE RESOURCES TO MERGE
Canadian Bashaw Leduc Oil and Gas Ltd said it agreed to merge with Erskine Resources Ltd.
Terms were not disclosed.
Ownership of the combined company with 18.8 pet for the current shareholders of Canadian Bashaw and 81.2 pet to the current shareholders of Erskine, the companies said.
Reuter

(a) The ASCII text from text image of 300dpi

CANADIAN yASHAW. HR^KJNh kI';SOlJRCn,S TO V1HRGR
Canadian Bashaw Leduc Oil and Gas I.Id s;ilii IT agreed ro merge wirii Hnkine RL^OUI-CCS Ltd.
Forms were not disclosed.
Ownership of ihe combined company with 1^.8 pet for the cuncnt shareholders of C,m;idi,in B^sbaw and 81.2 pel to ihc eiirrL'nt sbaicholLlcrs cifF.rskrne, die companies said.
Rt-'uter

(b) The ASCII text from text image of 140dpi

Fig. 3. Examples of the ASCII texts converted by OCR software

4.2.2 Text Generation by an OCR System

The text images generated above were converted into ASCII texts by OCR software "OKREADER2000". Examples are given in Figure 3(a) and (b).

The obtained texts were compared with the original texts in the Reuters collection to compute the average character recognition rates and the average word recognition rates for each dpi value. The average character recognition rates can be defined by:

$$c = \frac{(s-t)}{s} \times 100 \tag{3}$$

Whereby, s and t are the numbers of total characters and the number of miss-recognized characters, respectively. The average word recognition rate can be defined by:

$$v = \frac{(w-u)}{w} \times 100 \tag{4}$$

Whereby, w and u are the numbers of total words and the number of miss-recognized words, respectively.

4.2.3 Automatic Text Classification

Automatic Text Classification was done as described in the following subsections:

I) Feature Vector Generation

First a lexicon consisting of all different words in a learning text set was generated and the alphabetic order of the word list was created. Then the feature vector was composed of the frequencies of the lexicon words found in textual documents. The features extracted can be represented in form of the feature vector X which can be denoted as:

$$X = (x_1, x_2, \cdots, x_n)^T \tag{5}$$

Whereby, n is dimensionality (size of lexicon), x_i is the frequency value of ith word and T refers to the transpose of a vector.

II) Dimension Reduction

Dimension reduction (DR) in Automatic Text Classification is essential due to the following among other basic two reasons. First in text classification high dimensionality of the term space may be problematic in terms of computational time and storage resources. Second, DR tends to reduce over-fitting. In the experiments we selected word frequencies, $n > 2$, and we used Principal Component Analysis (PCA) technique to extract significant components and to further reduce the dimensionality [3], [5].

III) Learning

Various classifiers were trained accordingly using a learning sample as follows. The Euclidean distance classifier involved computing the mean vector of each class. The linear discriminant function required computation for the weight vector determined by the mean vector of each class and the pooled within covariance matrix of all classes. Training the projection and the modified projection distances needed the computation of the eigenvectors and the eigenvalues of each of the individual category's covariance matrix [4]. Support Vector machines (SVMs) are methods that find the optimal hyperplane during training. In the experiments, C-support vector classification methods (C-SVC) with linear and radial basis (RBF) functions were used. Particularly, we used the SVM library (LIBSVM Version 2.33) developed by Chang and Lin [6].

VI) Classification

The feature vectors of reduced dimensionality were classified to the category with the distance (or the discriminant function) of which was minimum. Referring to the subject field manually given to each article in Reuters-21578, the classification rates, R were calculated by

$$R = \frac{x}{(x+y)} \times 100 \tag{6}$$

Whereby, x and y are the numbers of articles correctly classified and incorrectly classified, respectively [5].

5 Empirical Results

In this section we present experimental results from different features that include absolute word frequency, relative word frequency and their power transformations.

Table 1 shows the classifiers' classification rates versus character recognition rates and the word recognition rates from absolute word frequency at different resolutions. On this table it might be observed that, as the resolution of text images decreased, the character recognition and word recognition rates by an OCR system also decreased. In other words at relative higher resolutions, it was possible to obtain less recognition errors by OCR systems. Similarly, classification rates of OCR texts decreased with increase in OCR errors.

Table 1. OCR Text classification rates (%) for absolute frequency vs. character recognition rates (%) and word recognition rates (%) by an OCR system at different resolutions (dpi)

Resolution (dpi)	130	135	140	145	150	200	300
Word recognition rates	41	53.8	63.7	72.1	84.3	92.9	97.2
Character recognition rates	57.7	71.6	82.8	89.8	96	98.4	99.3
Euclidean distance	44.9	51.9	58.4	62.1	67.2	70.7	74.3
Linear discriminant function	65.7	74.8	80.1	86.0	88.3	89.9	91.1
projection distance	75.2	83.3	87.1	88.4	90.1	90.7	91.2
Modified projection distance	78.1	86.3	89.3	91.6	92.5	92.8	93.1
SVM-Linear	76.1	84.9	87.5	89.9	92.0	92.9	93.3
SVM- RBF	64.5	76.8	82.0	86.3	89.5	91.3	92.1

The summarized best classification rates from all features using different classifiers are given in table 2. It is notable that transformed features improved the performance of all used classifiers. Performing power transformation on relative frequency for example made all classification rates to rise as high as over 91%.

Table 1 and 2 also reveal that modified projection distance (MPD) outperformed all the classifiers used in terms of accuracy and robustness. In other words, this

classifier gave out the highest classification rates even when there were more OCR errors. For example when OCR word recognition rate was 41%, MPD was accurate by 78.1% - when absolute frequency was used as feature vector. This was improved to 91.7% [1] by employing power transformation on relative frequency (PTR). And when OCR word recognition rate was 97.2%, the MPD's classification accuracy was improved from 93.1% to 96.1% - when PTR was used as feature vectors.

Table 2. The summary of results showing the best classification rates in % at 300dpi

	Absolute Frequency		Relative Frequency	
Classifiers	without power transformation	with power transformation	without power transformation	with power transformation
Euclidean distance	74.3	89.6	86.0	91.1
Linear discriminant function	91.1	92.8	93.9	94.5
Projection distance	91.2	94.9	93.3	95.7
Modified projection Distance	93.1	95.3	94.7	96.1
SVM-Linear	93.3	95.1	94.3	95.3
SVM- RBF	92.1	93.5	94.4	95.3

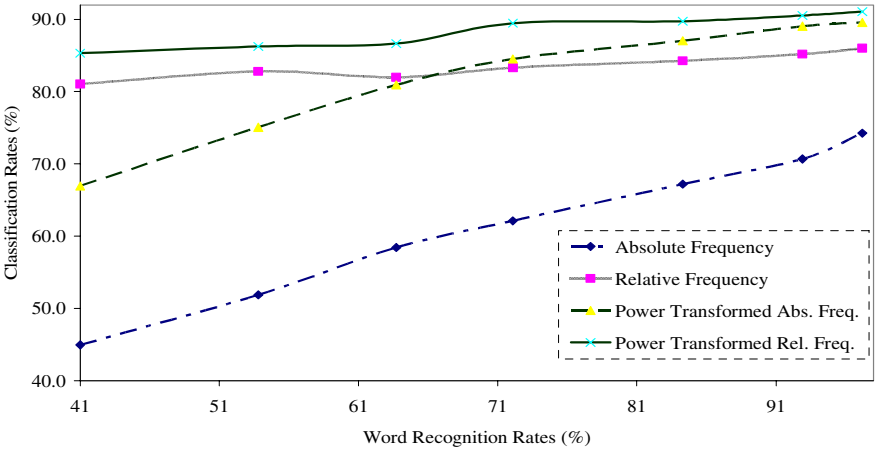


Fig. 4. Word recognition rates vs. classification rates for Euclidean distance classifier with improved results after feature transformation

The relationships between the OCR texts' classification rates of each classifier and word recognition rates are shown in figures 4 to 8. These figures also show the

[1] Note that 91.7% is a detail which is not reported in table 1 and 2. It was obtained at resolution 130dpi.

performance of each classifier with all feature types that is absolute word frequency, relative frequency and their power transformations.

By observing figures 4 to 8 it can be learnt that the classification rates were significantly improved by using the relative frequency instead of the absolute word frequency. For instance the accuracy of Euclidean distance classifier was improved by 11.7% at 300dpi and by 36.2% at 130dpi. In addition, power transformation on absolute frequency (PTA) also improved the performance of all used classifiers. However, it is clear that the use of PTA gave out more classification errors when there were more OCR errors in generating texts.

Power transformation on relative frequency (PTR) further improved the classification accuracy of each classifier used. For example the accuracy of Euclidian distance classifier was finally improved cumulatively by 16.8% at 300dpi and by 40.4% at 130dpi. PTR improved classification accuracy such that all classifiers exhibited over 91% classification rates.

Not only did the classifiers performance rose by doing power transformation on relative frequency, but also the robustness of classifiers increased such that even when OCR systems gave a lot of unacceptable amount of errors the performances were considerably higher than using untransformed features in representing OCR texts for classification purposes. For example at highest level of OCR errors, when PTR was used, the worst classifier performed as high as 85.5% accuracy and the best classifier came up with 91.7% accuracy.

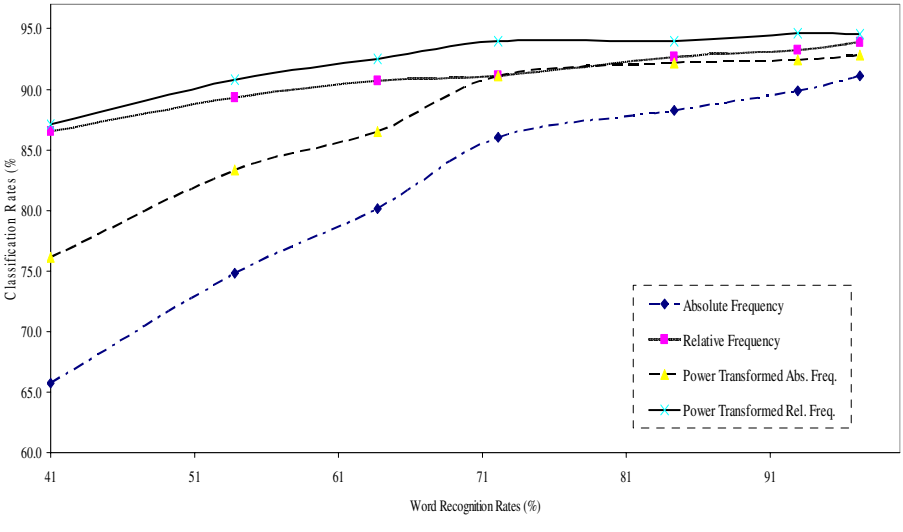


Fig. 5. Word recognition rates vs. text classification rate for linear discriminant function with improved performance after feature transformation

It is also interesting to note that transformed features particularly relative frequency do not heavily depend on word recognition rates by the OCR systems. In

such a way that, the differences in accuracy between the absolute frequency and the transformed features, increase as the word recognition rates by OCR systems decrease.

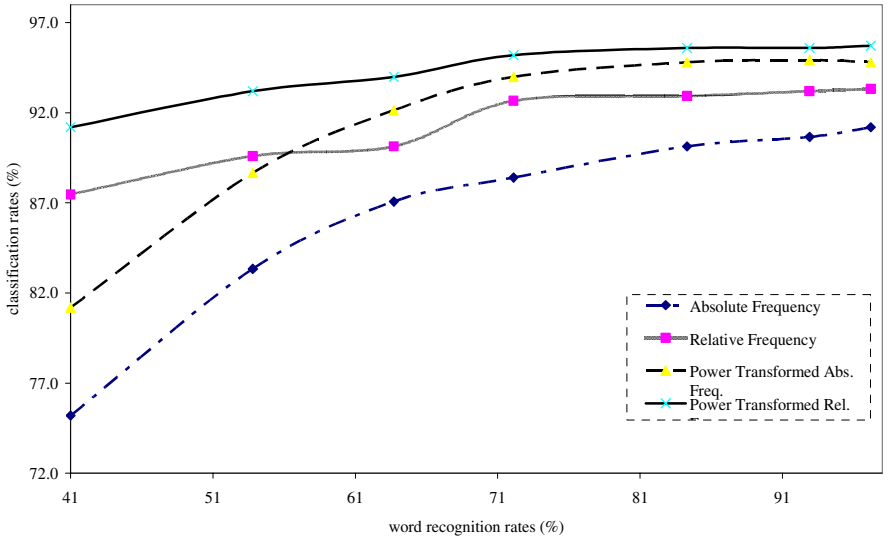


Fig. 6. Word recognition rates vs. text classification rate for projection distance with improved performance after feature transformation

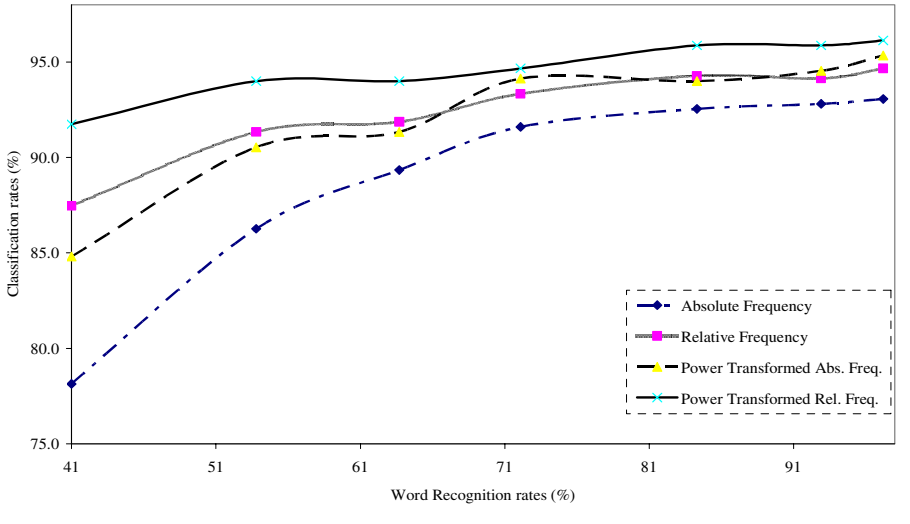


Fig. 7. Word recognition rates vs. classification rates for Modified Projection Distance with improved results after feature transformation

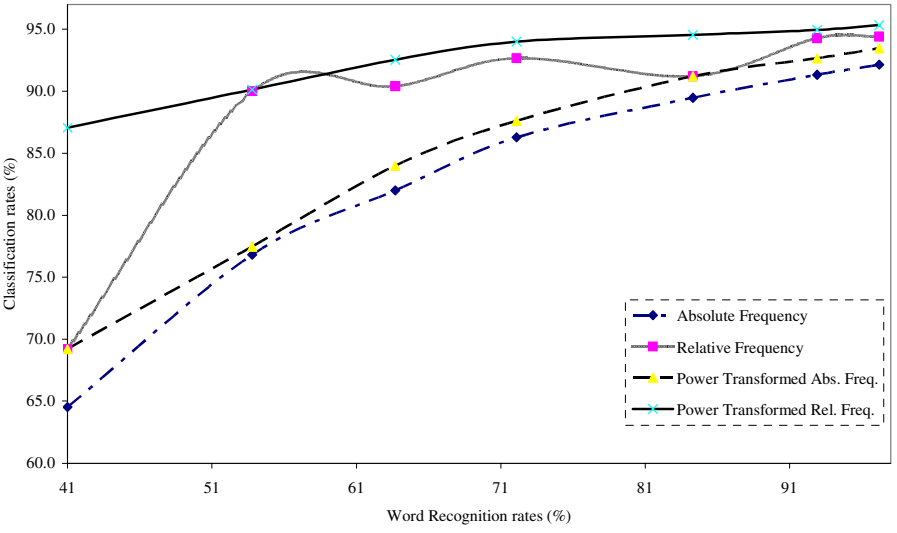


Fig. 8. Word recognition rates vs. classification rates for SVM-linear with improved results after feature transformation

6 Conclusion and Future Study

In this paper we have shown the impact of using transformed features for OCR-generated documents in automatic text classification. The findings show that using transformed features significantly improved the performance of all used classifiers. Even when OCR systems gave a lot of errors by representing texts with transformed features it was encouraging to obtain as higher classification rates as possible. The implications of these results are that, with error-prone OCR texts it is possible to automate the classification tasks and use the automation in different applications such as information retrieval, information filtering and document organization.

Future experiments will include increasing the sample size from more categories for real world applications in text classification. Also error correction of words by spelling check is also remaining as a future study to improve the text classification performance.

References

[1] Ohta,M., Takasu,A., Adachi,J.: "Retrieval Methods for English-Text with Missrecognized OCR Characters", *Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR),* pp.950-956, August 18-20, 1997,Ulm, Germany.

[2] Myka, A., Guntzer. U.: "Measuring the Effects of OCR Errors on Similarity Linking", *Proceedings* of the Fourth International Conference on Document Analysis and Recognition (ICDAR), pp.968-973, August 18-20, 1997, Ulm, Germany.

[3] Zu, G., Murata, M., Ohyama, W, Wakabayashi, T. and Kimura, F.: "The impact of OCR accuracy on Automatic Text Categorization", *Proceedings of Advanced Workshop Content Computing,* pp. 403-409, 2004.

[4] Fukumoto,T., Wakabayashi,T. Kimura,F. and Miyake,Y.: "Accuracy Improvement of Handwritten Character Recognition By GLVQ", *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition Proceedings*(IWFHR VII), 271-280 September 2000.

[5] Guowei Zu, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura,: "Accuracy improvement of automatic text classification based on feature transformation" *DocEng'03 (ACM Symposium on Document Engineering 2003*), pp.118-120, November 20–22, 2003, Grenoble, France

[6] C.C. Chang, and C.J. Lin : "LIBSVM -- A Library for Support Vector Machines (Version 2.33)", *http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html,* (2002.4)

[7] Library Digital Initiative Project Team, Harvad University Library: "Measuring Search Retrieval Accuracy of uncorrected OCR: Findings from the Harvad-Radcliffe Online Historical Reference Shelf Digitization Project" A research report available at *http://preserve.harvard.edu/resources/ocr_report.pdf,* (Aug. 2001)

[8] Bicknes, D.A: "Measuring the accuracy of the OCR in the Making of America". A research report available at http*://www.hti.umich.edu/m/moagrp/moaocr.html* (1998)

[9] Junker, M and Hoch, R.: "An experimental evaluation of OCR text representations for learning document classifiers". *International Journal on Document Analysis and Recognition.* Springer-Verlag, pp. 116-122, 1998

[10] Frasconi, P., Soda, G. and Vullo, A: "Text Categorization for Multi-page Documents: A Hybrid Naïve Bayes HMM Approach". *1st ACM-IEEE Joint Conference on Digital Libraries(JCDL'01)* Roanoke Virginia (June, 2001)

[11] Frasconi, P., Soda, G. and Vullo, A: "Hidden Markov Models for Text Categorization in Mult-page Documents". *Journal of Intelligent Information Systems, 18:2/3, 195–217, 200*2002).

[12] Taghva, K., Nartker, T., Borsack, J., Lumos, S., Condit, A. and Young, R.: "Evaluating Text Categorization in the Presence of OCR Errors," *Proceedings of the Symposium on Electronic Imaging Science and Technology*, pages 68-74, San Jose, CA, January 2001.