Mobile Video Capture of Multi-page Documents

Jayant Kumar
University of Maryland,
College Park, USA
jayant@umiacs.umd.edu

Raja Bala, Hengzhou Ding, Phillip Emmett
Xerox Research Center Webster, USA
{Raja.Bala, Hengzhou.Ding,
Phillip.Emmett}@xerox.com

Abstract

This paper presents a mobile application for capturing images of printed multi-page documents with a smartphone camera. With today's available document capture applications, the user has to carefully capture individual photographs of each page and assemble them into a document, leading to a cumbersome and time consuming user experience. We propose a novel approach of using video to capture multipage documents. Our algorithm automatically selects the best still images corresponding to individual pages of the document from the video. The technique combines video motion analysis, inertial sensor signals, and an image quality (IQ) prediction technique to select the best page images from the video. For the latter, we extend a previous no-reference IQ prediction algorithm to suit the needs of our video application. The algorithm has been implemented on an iPhone 4S. Individual pages are successfully extracted for a wide variety of multi-page documents. OCR analysis shows that the quality of document images produced by our app is comparable to that of standard still captures. At the same time, user studies confirm that in the majority of trials, video capture provides an experience that is faster and more convenient than multiple still captures.

1. Introduction

With ongoing advances in mobile camera technology, smartphones are being increasingly used to capture images of documents in a variety of consumer and business applications. Examples of such documents include bank applications, checks, insurance claims, receipts, etc. A variety of mobile apps for document scanning are available in the market, notable ones being *Camscanner*, *MDScan*, *JotNot*, and various apps from banks like *Chase*. When the document comprises multiple pages, all of these apps require the user to take a sequence of still shots, one per page. This process can be time-consuming and cumbersome for the user, especially when the number of pages becomes large, as would be the case for certain bank and loan application forms.

We propose a novel means for mobile capture of multi-page documents that uses video instead of still photography. We hypothesize that video capture provides a much superior user experience as it eliminates the need to manually frame and capture multiple still shots. We exploit the fact that most modern smartphones capture video in HD resolution, and have verified this provides adequate document image quality (IQ) for many business applications. What remains is an automatic way to select a set of high quality images of individual pages of the document from the video sequence. We propose an algorithm that combines video motion analysis, inertial sensor signals, and an extension of a recent learning-based IQ prediction technique to select the best page images from the video. To our knowledge, there does not exist such a video-based app for document page capture.

The mobile application works as follows. The user launches the app on his/her smartphone, points the device at the document and starts the video capture as s/he flips through the pages of the document. Measurements from the smartphone's inertial sensors are simultaneously recorded with the video. Once recording is completed, the app invokes the automatic page selection algorithm and presents images corresponding to individual pages to the user for further processing (e.g. geometric correction, noise removal, etc.) and approval.

2. Related Work

In a general sense, our problem is related to that of video storyboarding [1] and summarization [2] which addresses the task of extracting key frames corresponding to important events from video data to form a summary representation. In our application, the important event occurs when the mobile camera "sees" a clear view of a document page; while events corresponding to page turns must be discarded. In this regard, we note related work by Yamada et al. [3] wherein a static overhead camera captures video of a user interacting with a document in a controlled office setting. Page turn events are detected via a series of frame differencing operations, and hand detection is accomplished by detecting regions with predefined skin color in the video frames. The key distinction is that in our application, video is captured by a mobile device in arbitrary environments; and hence our algorithm must be more robust with respect to capture conditions such as camera shake, widely varying lighting, etc.

A key component of our algorithm is the ability to analyze and predict image quality. There is a significant body of literature in this area [6-13]. In our application we are primarily dealing with document images containing significant text content, hence we seek a metric that measures text quality [9,10,12]. Furthermore, since the user can capture video of arbitrary printed matter, we are interested in the class of "no-reference" IQ analysis techniques that do not require the presence of a reference image to compare against [7,8,10,11]. Additionally, since mobile capture can engender a wide range of distortions such as camera motion, focus blur, and lighting effects, the use of hand-crafted rules for characterizing these distortions often do not generalize well. Finally the technique must be computationally efficient to lend itself to a mobile implementation. Considering all these factors, we adopt an unsupervised feature learning technique developed by Ye et al. [8,10] that predicts optical character recognition (OCR) performance of document images. We then extend their approach to address the video frame selection problem.

3. Page Selection Algorithm

Our algorithm comprises the following cascade of operations that successively eliminate unwanted frames from the video sequence:

1. Detection and removal of frames involving page turning and hand presence.
2. Use of inertial sensor data to remove frames exhibiting camera motion.
3. Calculation of IQ scores using a machine learning technique.

The frame with the highest IQ score for each page is selected. An inadequate page IQ score triggers appropriate feedback to the user for possible re-capture. The above three steps are described in further detail below.

3.1. Detection of Page Turn and Hand Interaction

The first task is to detect page turn events as these frames must be eliminated, and furthermore, they mark the boundaries of video segments corresponding to individual pages.

The input video is subsampled in the spatial and temporal dimensions in order to constrain computational cost. The absolute difference between adjacent frames is computed, followed by Gaussian smoothing using a 3×3 kernel and $\sigma_x = \sigma_y = 0.5$. This yields large blobs corresponding to large-scale motion in a page turn event.

Figure 1. Binarized blob images corresponding to page-turn (left) and non-page-turn (right) events

Morphological erosion using a circular structuring element of diameter 7 pixels is then applied to remove fine-scale motion arising from shake and/or jitter. The resulting image is binarized with an empirically determined threshold, and blobs are accumulated over a number of successive frames via a logical OR operation. In our application, integration over 6 frames provides a reliable indicator of page turns. Fig. 1 shows the binarized blob images corresponding respectively to a page-turn and non-page-turn event. Fig. 2 shows a plot of the stacked blob size as a function of frame number for a 4 page simplex document. The three prominent peaks correspond to true page-turn events.

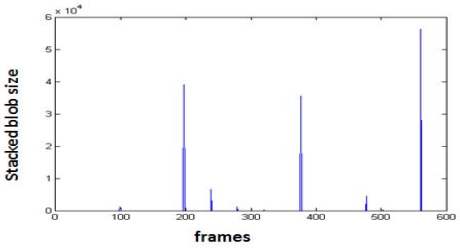


Figure 2. Stacked blob size vs. video frames for a 4 page simplex document.

Figure 3. Hand interaction detection: (a) reference frame (b) test frame containing hand interaction.

The next event that we wish to detect and remove is partial occlusion from the user's hand interacting with the document. This normally occurs just prior to a page turn, and is not necessarily caught by the aforementioned motion detection algorithm. We search for frames occurring just prior to page-turn that exhibit significant luminance change with respect to a reference frame posited to have no hand interaction. Specifically, the frame is divided into 3 × 3 sub-blocks and the standard deviation σ of luminance values within each sub-block is computed for both reference and test frames (see Fig. 3). If the difference in σ is larger than a predefined threshold then the frame is marked as a hand-interaction event and eliminated. The reference frame is randomly selected from

the first few frames immediately after a page turn event (where we do not expect to see hand interaction). In the case of the first page, the reference is selected randomly from the first 1-2 seconds of capture. The main advantage of this simple approach is its time-efficiency which is critical for mobile devices.

3.2. Camera Motion Detection

The use of the smartphone's inertial sensors to detect and recognize user movement is an area of active exploration [4, 5]. In our application, we use inertial sensor data as auxiliary input to eliminate frames with significant camera motion during video capture. The iPhone 4S emits acceleration data calibrated for gravity (G) (i.e. with the effect of gravity removed) in x, y, and z directions. We record and compute the magnitude of the 3D acceleration vector at the instant that each video frame is grabbed. Frames whose acceleration magnitude exceeds an empirically determined threshold of 0.02G are eliminated. Fig. 4 plots acceleration magnitude for a sample video capture. Also shown are portions of two frames exhibiting low and high acceleration magnitudes respectively. The inertial sensor clearly offers a simple and effective way to eliminate frames with objectionable motion blur.

3.3. Image Quality Prediction

Ideally, the frames that have survived up to this point should be free of page turn and camera motion effects, but may still vary significantly in quality due to effects of camera focus, shadows, etc. We thus propose the use of an additional measure of document IQ for ranking and selecting the best frame for each page. An absence of high quality frames can also be used to provide an alert that the user may need to recapture certain pages.

We first describe the approach taken by Ye et al. [10] followed by an important extension we have developed for the purpose of video frame selection. The original algorithm comprises two phases: unsupervised feature learning and document classification.

3.3.1 Unsupervised feature learning. In this phase, training data is generated by acquiring image frames from representative mobile video captures of printed text documents. The images are processed through an OCR engine to obtain OCR accuracy scores based on known ground truth. Patches of size M×M are extracted at random locations from the training images. In order to avoid selecting patches in non-informative regions, only patches whose pixel variance exceeds a threshold are selected.

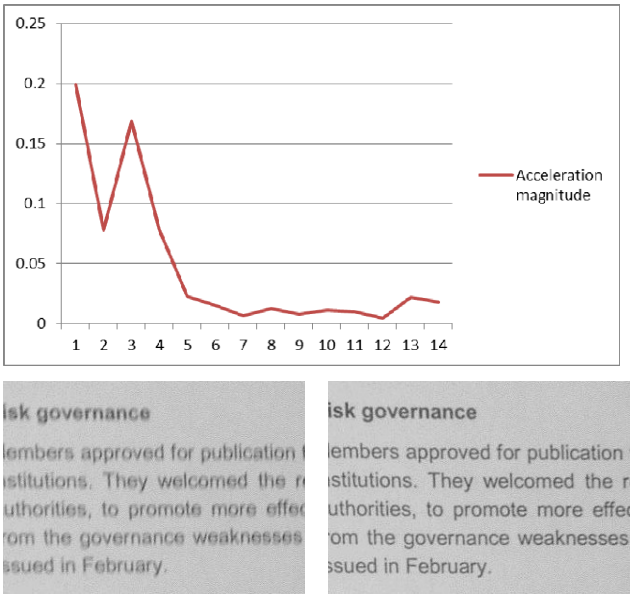


Figure 4. Top: Acceleration magnitude recorded during capture of a document. Bottom: video frames corresponding to 2nd frame (high motion) and 11th frame (low motion).

Patches are normalized to zero mean and unit standard deviation, reshaped into $M^2 \times 1$ vectors, and whitened using Zero Component Analysis to minimize correlation amongst vector elements. The resulting vectors $\mathbf{v}_i$ are clustered using K-means clustering, to produce a codebook $\{\mathbf{d}_1, \ldots, \mathbf{d}_K\}$ that captures representative document features such as edges and corners under varying types and degrees of distortion such as blur, contrast, noise, etc. (see Fig. 5.).

Next, a similarity measure $\mathbf{s}_i$ is computed between image patch $\mathbf{v}_i$ and the codebook entries $\mathbf{d}_i$ via inner products:

$$\mathbf{s}_i = [\mathbf{v}_i.\mathbf{d}_1, \ldots, \mathbf{v}_i.\mathbf{d}_K]^T \quad (1)$$

The positive and negative elements of $\mathbf{s}_i$ are separated into a modified vector $\mathbf{c}_i$ of dimension 2K as follows:

$$\mathbf{c}_i = [\max(\mathbf{s}_i[1],0), \ldots, \max(\mathbf{s}_i[K],0), \max(-\mathbf{s}_i[1],0), \ldots, \max(-\mathbf{s}_i[K],0)]^T \quad (2)$$

This kind of rectification has been shown to improve the performance of prediction [8]. Finally, the feature descriptor for the entire image is computed as an aggregate of local patch descriptors $\mathbf{c}_i$. Ye et al. recommend a max-pooling strategy, while in our experiments we have found that the average of the $\mathbf{c}_i$ is a more robust feature. The number of local patch descriptors to be used in the aggregation affects both the quality of the IQ prediction and computation time, and should be balanced appropriately. Fig. 5 summarizes the three step process involved in feature computation.

3.3.2 Document classification. The image features from the previous step are used to train a linear SVM [14] on

two broad classes: high-quality and low-quality documents. In [10], the two training classes are defined by selecting non-overlapping ranges of OCR scores. They also propose an alternate approach using support vector regression (SVR) to predict a continuum of OCR scores.

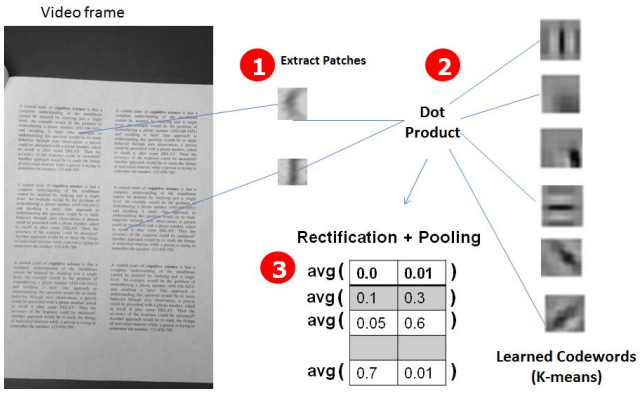


Figure 5. Steps involved in feature computation for learning IQ score

It should be noted that the intensive feature learning computation and SVM training are performed offline. The real-time operations are patch extraction, similarity measure computation, aggregate pooling, and linear SVM classification which are all efficient.

3.3.3 Extensions for video application. The original algorithm poses two important limitations for our application. First, the output of the SVM predictor is binary valued, and therefore does not provide an informed way to select among a set of video frames wherein multiple frames will invariably produce identical classifier outputs. The use of SVR could address this issue, however this method is highly sensitive to the probability distribution of OCR scores in the training data. We observed a sparse distribution with strong peaks in certain score intervals, and very low counts in others, which is insufficient for SVR training. The second limitation is that the IQ prediction is global, i.e. it is obtained from pooling descriptors across the whole page. In our experiments we frequently encountered images where effects such as focus blur and shadows were highly localized.

To address these issues, we extend the IQ prediction framework as follows. The image is divided into spatially non-overlapping sub-blocks, and the IQ predictor is applied to each block. Rather than returning a binary label, SVM returns the geometric margin, i.e. the signed distance from the input feature vector to the separating hyperplane given by:

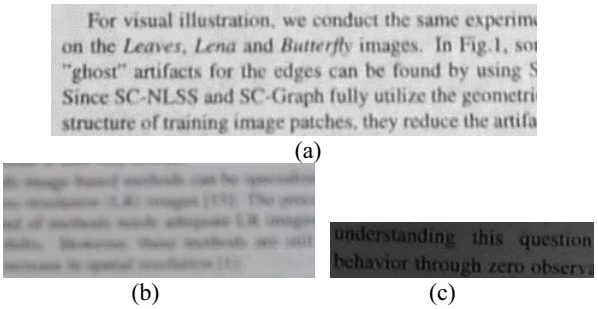
$$q_i = \mathbf{w}^T \mathbf{x}_i + b \qquad (3)$$

where $\mathbf{x}_i$ is the feature corresponding to the *i*-th subblock; q_i is the margin, and $\mathbf{w}$ and b are the SVM hyperplane parameters. The sign of q_i indicates the output class to which the image is assigned, and the magnitude conveys the confidence of the assignment. The optimal sub-block size trades off the extent of patch feature aggregation vs. the spatial resolution of the IQ metric, and was obtained via cross-validation experiments.

Our proposed extension thus provides both a meaningful continuously valued prediction as well as spatial dependence, thereby overcoming the two limitations of the original technique. In our implementation, we obtain a frame quality score Q by averaging the q_i across sub-blocks, and then select the frame with maximum Q for each page.

4. Experiments

The proposed algorithm was implemented as an iPhone 4S application. Video was captured in HD resolution and spatially subsampled by a factor of 4 in both x- and y-directions, and temporally subsampled to 6 frames per second within the app prior to processing. The image processing and IQ prediction algorithms were implemented using OpenCV Version 2.4.9. OCR on images was performed with ABBYY FineReader [15] and character level accuracy was obtained using the ISRI-OCR evaluation tool [16]. Parameters in our experiments were set using cross-validation on the training set.



(a) (b) (c)

Figure 6. Image segments extracted from training data set showing (a) high quality (b) focal blur (c) shadow

4.1. Image Quality Prediction

A data set was collected comprising smartphone video footage of a variety of multi-page text-intensive documents with different font sizes and styles, producing a total of 4800 video frames. The video was captured under a variety of representative lighting conditions and distortions such as shake, translation, motion, glare, shadow etc. Fig. 6 shows examples of training image patches corresponding to different capture conditions.
We randomly selected a set of 1448 images for learning the IQ score. We used a patch size of M=11 and codebook size of K = 50 for feature learning. A total of 1000 local patch descriptors were aggregated to obtain the

image feature. A two-class linear SVM classifier was trained with one class containing high quality samples with OCR accuracy greater than 90%, and the other containing low quality samples with OCR accuracy less than 10%. The images were then divided into 2×3 spatial sub-blocks, SVM margins were computed as in Equation (3), and averaged across sub-blocks to obtain a predicted score Q. Fig. 7 is a plot of Q vs. true OCR accuracy for a number of training samples. Superimposed on this plot is a logistic fit given by the function:

$$f(t) = \frac{100}{1+\exp\{-at-b\}} \quad (4)$$

where parameters a and b are obtained via least squares regression. The root-mean-square error of the logistic fit for an independent set of samples was 14.7 (where the quantity being predicted is percentage OCR accuracy on a scale from 0-100). Clearly the model cannot predict OCR score to a very fine precision. Note also that prediction error generally is higher in the mid-range of OCR accuracies, thus reducing the confidence of prediction in this regime.

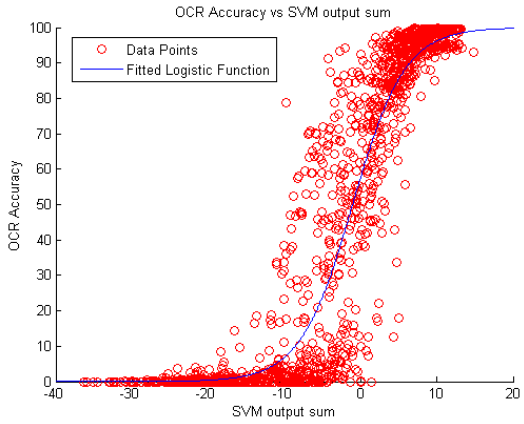


Figure 7. Plot of OCR accuracy vs. predicted IQ score (red circles) and logistic fit (blue curve).

For video frame selection, the IQ score is computed for all frames that survive the filters discussed in Sec. 2.1 and 2.2. For each document page, the frame with the highest IQ score is selected, and the corresponding score is referred to as the page score. Additionally we define two thresholds t_{high} and t_{low} to categorize pages into one of 3 broad classes based on their score: "high quality", "medium quality", and "low quality". If the page score falls in the "medium" category, the user is warned that the page may have to be re-captured. A page score in the "low" category automatically launches a re-capture step. The thresholds are tunable, and currently set to t_{high} = 90% and t_{low} = 10% in our mobile app.

4.2. User Experience Experiments

An experiment was conducted to assess the user experience when performing multi-page capture with our mobile video app vs. the standard process of capturing multiple still shots. Ten subjects participated in the experiment, all with experience in smartphone usage. Each subject was presented with five artifacts in random order: a duplex stapled 6-page conference paper, a simplex unstapled 4-page press release, a duplex 5-page brochure, an A3-sized bank application form (front/back), and a bank check (front/back). He/she was asked to capture the multi-page/multi-face artifact using the "still capture" mode as well as our video-based app. Initial training was given for subjects to become familiar with the interface. Capture sessions for each artifact were timed.

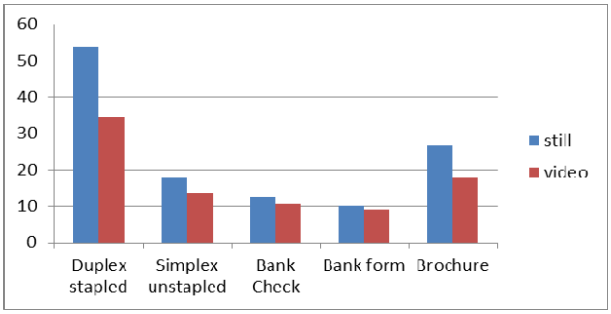


Figure 8. Average time (sec) for still vs. video capture.

Fig. 8 shows the average time across users for each artifact with the two capture modes. In most cases, video capture is noticeably faster, with timing differences being more pronounced for documents with a larger number of pages. This is not surprising since the number of "clicks" is substantially reduced with video capture.

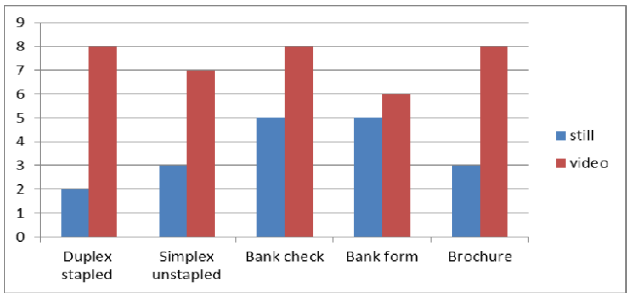


Figure 9. Preference for still and video capture mode. A score of 1 is given to the preferred mode for each subject. In case of ties, a score of 1 is assigned to both modes.

We also asked subjects to vote for their preference for the video vs. still capture mode. Ties were permitted. Fig. 9 plots the results of the votes, indicating again a strong overall preference for video capture.

4.3. Quantitative evaluation

We selected the 4-page press-release document for further analysis. Page images produced with the still- and video-capture apps by the ten subjects were run through ABBYY OCR. Statistics on per-page OCR accuracy are presented in Table 1. It is seen that the proposed video capture app performs comparably with the standard still capture mode.

Table 1. OCR accuracy for still vs. video capture.

	75th Percentile	Median	25th Percentile
Still Capture	99.9	99.6	98.8
Video Capture	99.8	98.8	87.1

The Pearson linear correlation between predicted page scores and OCR accuracies was 0.85. Also noteworthy is that for two of the ten users, the predicted page IQ scores were in the "low" and "medium" categories, and these were indeed a result of poor capture. These subjects were prompted to repeat the capture to produce better video quality. The IQ score thus provides an effective instant feedback mechanism to users. Finally, 140 out of 150 page turns were correctly detected (i.e. 93 % success rate).

5. Conclusion

We have presented an application for acquiring an electronic version of a multi-page printed document using smartphone video capture. At the heart of the application is an algorithm for automatically extracting high quality page images from the video feed using efficient image processing, inertial sensor input, and an extension of a machine learning technique to predict page IQ. Experiments validate our hypothesis that video capture offers a superior experience while posing minimal sacrifice to image quality. While we have reported results primarily with paper documents, the technique is applicable for any printed material such as checks, business cards, credit cards, packages, etc. Also since the algorithm parameters have been chosen based on standard HD video output that is available on most modern smartphones, we expect the algorithm to be robust across different mobile platforms.

There are several avenues for future exploration. Currently the processing time for page selection is approximately equal to the video capture duration, but can be significantly sped up with a parallelized (e.g. GPU) implementation. Image enhancement techniques such as super-resolution reconstruction can be explored. Another aspect to be investigated is improving the accuracy and resolution of the IQ prediction model, and extending its applicability from text to pictorial content. Finally it would be beneficial to utilize the full spatial profile of the IQ metric for richer user feedback, as well as to facilitate fusion of multiple video frames with different spatial IQ profiles to produce superior quality.

Acknowledgements

We thank Florent Perronnin and the reviewers for their constructive comments and suggestions in improving the quality of this paper.

References

[1] G. Bozdagi, H. Yu, M. R. Campanelli, R. Bryll, and S. J. Harrington, US Patent 6,647,535, "Methods and systems for real-time storyboarding with a web page and graphical user interface for automatic video parsing and browsing", 2003.

[2] C. Cotsaces, N. Nikolaidis, and I. Pitas. "Video shot detection and condensed representation. a review." IEEE Signal Processing Magazine, 23.2 (2006): 28-37.

[3] K. Yamada and K. Ishikawa, "A method of analyzing the handling of paper documents in motion images", Proc. ICPR, Vol. 4, pp. 413-416, 2000.

[4] S. L. Lau and K. David, "Movement recognition using the accelerometer in smartphones." IEEE FNMS, pp.1-9, 2010.

[5] S. Dernbach, B. Das, N. Krishnan, Bl L. Thomas, D. J. Cook, J. Diane, "Simple and Complex Activity Recognition through Smart Phones," Proc. ICIE , pp. 214-221, 2012

[6] A. Rúa, E. Castro, and C. G. Mateo. "Quality-based score normalization and frame selection for video-based person authentication." Proc. BIM pp. 1-9, 2008.

[7] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, 2005.

[8] P. Ye, J. Kumar, L. Kang and D. Doermann. "Unsupervised Feature Learning Framework for No-reference Image Quality Assessment", Proc. CVPR, pp. 1098-1105, 2012.

[9] X. Peng, H. Cao, K. Subramanian, R. Prasad, and P. Natarajan. "Automated image quality assessment for camera-captured OCR." Proc. ICIP, pp. 2621-2624, 2011.

[10] P. Ye and D. Doermann. "Learning features for predicting OCR accuracy", Proc. ICPR, pp. 3204-3207, 2012.

[11] S. Suresh, R. Babu, H. J. Kim, "No-reference image quality assessment using modified extreme learning machine classifier", Journal of Applied Soft Computing, Vol. 9(2), pp. 541-552, 2009.

[12] H. S. Baird, "Document image quality: making fine discriminations", Proc. ICDAR, pp. 459-462, 1999.

[13] J. Kumar, F. Chen and D. Doermann, "Sharpness Estimation for Document and Scene Images", Proc. ICPR, pp. 3292-3295, 2012

[14] C. Cortes and V. Vapnik, "Support-vector networks", Machine Learning, Vol. 20, Issue 3, pp. 273-397, 1995.

[15] ABBYY Finereader 10 Professional Edition, Build 10.0.102.74, 2009.

[16] ISRI-OCR evaluation tool. 2010: http://code.google.com/p/isri-ocrevaluation-tools