Natural Scene Character Recognition Without Dependency on Specific Features

Muhammad Ali and Hassan Foroosh
Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, Florida, U.S.A.

Keywords: Natural Scene Text Recognition, Tensors, Rank-1 Decomposition, Holistic Character Recognition, Feature Independence.

Abstract: Current methods in scene character recognition heavily rely on discriminative power of local features, such as HoG, SIFT, Shape Contexts (SC), Geometric Blur (GB), etc. One of the problems with this approach is that the local features are rasterized in an ad hoc manner into a single vector perturbing thus spatial correlations that carry crucial information. To eliminate this feature dependency and associated problems, we propose a holistic solution as follows: For each character to be recognized, we stack a set of training images to form a 3-mode tensor. Each training tensor is then decomposed into a linear superposition of 'k' rank-1 matrices, whereby the rank-1 matrices form a basis, spanning solution subspace of the character class. For a test image to be classified, we obtain projections onto the pre-computed rank-1 bases of each class, and recognize it as the class for which inner-product of mixing vectors is maximized. We use challenging natural scene character datasets, namely Chars74K, ICDAR2003, and SVT-CHAR. We achieve results better than several baseline methods based on local features (e.g. HoG) and show leave-random-one-out-cross validation yield even better recognition performance, justifying thus our intuition of the importance of feature-independency and preservation of spatial correlations in recognition.

1 INTRODUCTION

Natural scene text recognition is a challenging problem in computer vision, machine learning and image processing. Ubiquitous availability of digital cameras on mobile devices e.g. phones and glasses, has computer vision applications like assisted navigation for visually impaired people, e.g. OrCam[1] device mounted on glasses etc. Moreover, huge online image data repositories can be mined for textual content to automatically generate useful information for marketing, archival, and other purposes. Other utilities of natural scene text recognition include and automatic reading of informational signs for automobile drivers or driverless cars.

Successful commercial applications of document text recognition ensued interest in solving the more general problem of natural scene text recognition. Owing to its complexity and challenges, the problem has been broken up into the following four sub problems:

1. Cropped Character Recognition
2. Cropped Word Recognition
3. Scene Text Detection
4. Full-image Scene Text Recognition

There is another related problem proposed by (Wang et al., 2011) to recognize words in an image given a short vocabulary pertaining to the image.

The introduction of ICDAR2003 reading competition (Lucas et al., 2003) and the associated dataset stirred up research interest in document recognition community and subsequently other researchers came forward and proposed their methods and/or datasets. For example (Weinman et al., 2009) used their sign reading dataset (WLM dataset), (de Campos et al., 2009) proposed Chars74K, and (Wang et al., 2011) came up with their Street View Text (SVT) dataset. More recently (Nagy et al., 2011) put forth NEOCR dataset.

Figure 1 shows sample images from popular natural scene text character datasets like Chars74K and ICDAR. The challenge is obvious from the sample noisy images which exhibit low resolution, variable typefaces, illumination effects, perspective distortions, all sorts of structured and/or random noise.

[1] http://www.orcam.com

Ali M. and Foroosh H..
Natural Scene Character Recognition Without Dependency on Specific Features.
DOI: 10.5220/0005305603680376
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 368-376
ISBN: 978-989-758-090-1

Figure 1: Sample characters from Chars74K (English) and ICDAR2003 datasets.

In this paper we address the sub problem 1 above and use ICDAR2003 robust character dataset, the Chars74K dataset, and the SVT-CHAR, a derivative from SVT dataset, which has been annotated by (Mishra et al., 2012) to report our results.

The paper is organized as follows: Section 2 describes the related research work done on this problem. Section 3 gives an overview of the tensor rank problem. Section 4 gives describes our method to solve the problem. Section 5 presents experimental setup, results, and discussion. Section 6 gives a recap of this paper along with future research directions.

2 RELATED WORK

Since the introduction of ICDAR 2003 Robust Reading Competition and the associated challenge datasets (Lucas et al. 2003), the area of scene text recognition has seen an increase in research efforts to solve the problem. Various solutions have been proposed for the sub problem of robust character recognition.

Some researchers used off-the-shelf OCRs to recognize characters. (Chen and Yuille, 2004) used an adaptive version of Niblack's binarization algorithm (Niblack, 1985) on the detected textual regions and then employed commercial OCRs for final recognition. The reported results with ABBYY (www.abbyy.com) were good for their dataset (collected from cameras mounted on blind people.) Later performance of ABBYY reported by (Wang and Belongie, 2010) and (de Campos et al., 2009) showed its poor performance on more challenging ICDAR and Chars74K datasets.

Overall, the literature in natural scene character recognition is dominated by local feature-based methods: These methods mainly focus on extracting a feature vector, e.g. a Histogram of oriented Gradients (HoG) (Dalal and Triggs, 2005) or some variant of it, from a character image and then using some classifier, e.g. Nearest Neighbor, SVM etc., to recognize the character. (de Campos et al., 2009) used various feature descriptors including Shape Contexts (SC), Scale Invariant Feature Transform (SIFT), Geometric Blur (GB), etc. in combination with bag-of-visual-words model. The results, however, left a lot of room for improvement. (Weinman et al., 2009) used a probabilistic framework wherein they utilized Gabor filters in their similarity model to recognize characters in their dataset. (Wang and Belongie, 2010) showed better performance than (de Campos et al., 2009) by incorporating HoG features. (Neumann and Matas, 2011) used maximally stable extremal regions (MSER) to create MSER mask and then got features along its boundary which they used in SVM for classification. (Donoser et al., 2008) used MSERs in conjunction with simple template matching to get initial character recognition results which are subsequently improved by exploiting web search engines to get final recognition results. In addition, unsupervised feature learning system has been proposed by (Coates et al., 2011) that utilizes a variant of K-means clustering to first build a dictionary then map all character images to a new representation in the dictionary.

In this paper, we propose a novel approach for scene character recognition based on rank-1 decomposition of image tensor. Our results show that the proposed method effectively captures character shape variations occurring in natural scene images in a holistic manner thus avoiding the problems associated with ad hoc rasterisation of local image features.

We report our best performance on the aforementioned popular datasets using leave-random-one-out cross-validation, justifying thus our solution to achieve feature-independency for better recognition.

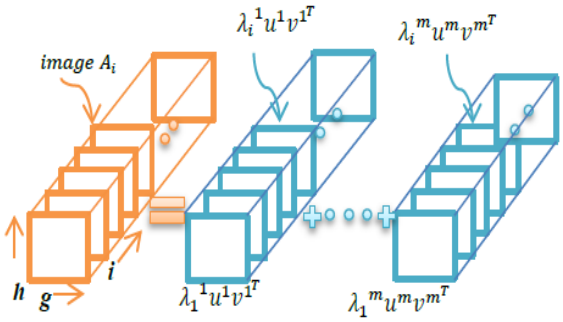

Figure 2: Rank-1 decomposition of a 3-mode tensor.

3 RANK-1 TENSOR DECOMPOSITION

The rank-1 decomposition has been shown to be effective for face recognition by (Shashua and Levin, 2001) and (Sun et al., 2011) have used it for action recognition. In the following, we give its novel application to characters extracted from natural scene images.

Consider a set of character images $\{A_i\}$, where $i = 1, \dots, d_3$ and the dimensions of images be $d_1 \times d_2$. Let the images be stacked together as slices of a tensor T whose elements are $T_{g,h,i}$, where $g = 1, \dots, d_1$ and $h = 1, \dots, d_2$.

The following expresses tensor T as the sum of k rank-1 tensors:

$$T = \sum_{m=1}^{k} \lambda^m \otimes u^m \otimes v^m \quad (1)$$

where u, v are the basis vectors and λ represents the mixing coefficients. The smallest k for which equation (1) holds is called the rank of T. Figure 2 illustrates the process of rank-1 decomposition of T.

For two-image tensor, polynomial time algorithms are available for low rank factorization. However, when the number of slices (or images) of T are more than '2' the problem of finding such a superposition of low rank tensors is NP-hard (Hazan et al., 2005). Various algorithms have been proposed to get the rank-1 factors of a multi-image (number of images >2) tensor depending upon how the solution space is constrained. For example, High-Order SVD (HOSVD) (Xianqian and Sidiropoulos, 2001) enforces orthognality constraints among the basis vectors. (Shashua and Levin, 2001) give algorithms to the effect of getting desirable SVD like extension to the multi-image tensor decomposition.

We modify the greedy algorithm given in (Shashua and Levin, 2001) to get the rank-1 decomposition of scene character image tensors. The iterative algorithm solves the following minimization problem to get the desired unit vectors (rank-1elements) uv^T and the mixing scalar vector $[\lambda_1, \dots, \lambda_p]$ associated with each image.

$$\sum_{i=1}^{p} \|A_i - \lambda_i u v^T\|_F^2 \quad (2)$$

where $A_1, A_2, \dots, A_p$ is the given set of images. The steps are summarized below:

1. Create the matrix $S = \sum_{i=1}^{p} A_i A_i^T$ and find the eigenvector corresponding to the largest eigenvalue. This becomes the unit vector u and captures the spatial redundancy in the image set
2. Using u from above, get the eigenvector v corresponding to the largest eigenvalue of the matrix MM^T, where the columns of M are $A_i^T u$. Hence v captures the temporal aspect of character images, e.g. font variations etc.
3. Next, find the scalar λ_i associated with each image as the inner product: $v^T A_i^T u$
4. Compute the residual image as: $A_i = A_i - \lambda_i u v^T$, replace it with the original image in the set and repeat the above steps until stopping criterion is met

(Shashua and Levin, 2001) do iterative refinement of the vectors u and v in step 1 and 2 respectively, around the initially estimated location before computing the mixing scalars. We avoid this because we empirically found that it exacerbates noise in scene character images and results in performance reduction.

The stopping criteria in step 4 could be the residual falling below a specified threshold or pre-specifying the number of rank-1 elements. We use the latter one and the impact of it on performance is further discussed in (Section 5.3.2).

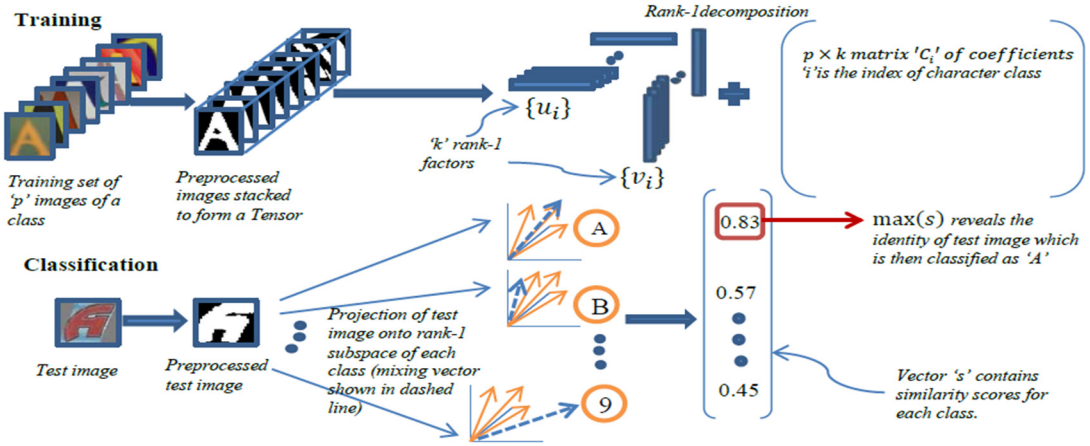

Figure 3: Illustration of our framework for scene character recognition.

4 SCENE CHARACTER RECOGNITION

Our framework for scene character recognition, as depicted in Figure 3, starts with preprocessing images, followed by tensor decomposition to extract rank-1 basis matrices that span the training images, projecting test images in each character class (subspace), and finally classifying the test image using inner product of mixing scalar vectors as a similarity measure.

4.1 Preprocessing

The cropped character images from natural scene datasets contain a lot of non-character structures and imperfect cropping artefacts which makes it difficult to effectively capture typeface and shape variations. Since the focus of our work is to demonstrate effectiveness of holistic recognition framework based on rank-1 tensor decomposition, we preprocess each image in training and testing sets to keep the images as noise free as possible. To this end we adopt binarization for image segmentation to reduce noise and extract, possibly only, character structures. This somehow lets us isolate the classification problem from the binarization problem.

4.1.1 Image Segmentation

Binarization has been used to segment textual information from natural scene images, e.g. see [(Chen and Yuille, 2004), (Mishra et al. 2011), (Kita and Wakahara, 2010), (Field and Learned-Miller, 2013)]. Binarization of natural scene character images is a challenging problem in itself. Hence, for the purpose of this paper, we employ a simple and novel combination of the methods of (Yokobayashi and Wakahara, 2005) and (Otsu, 1979) in an effort to segment each image to get textual foreground (in white). Using these methods, we get both the binary image and its inverted version. For these binary images, we then perform a connected component analysis based on the observation that cropped characters mostly fall in the middle of the image. We consider any small pixel group as noise if its size is less than a small fraction (<5%) of the size of the largest central connected component.

[2]http://www.mathworks.com.au/matlabcentral/fileexchange/34767. Last visited: 10 December 2014.

4.1.2 Size Normalization

We then normalize each image following a convex-hull based method given by D'Errico[2]. The intuition is the fact that convex hull contains an edge of a rectangle that bounds an image. We find the one that has the least perimeter. We then try to make the image upright by rotating the bounding rectangle so that its major axis is vertical. This somehow corrects the slant in the input image. Following this, we resize the image to 32x32 pixels.

Figure 4: Preprocessing of scene character images.

4.1.3 Selection of Correct Segmentation

The four binary images from the above steps contain a potential candidate that we select as a correct binary image. In the case of training, we use a reference image of a class to decide on the correct binary image. The reference image we use is a character (for each class) in Arial font centered in the image using ImageMagick[3]. To get the 'correctly' binarized image, we subtract each binary image from the reference image and pick the one with the least Frobenius norm. For testing we simply check segmentation for all reference images and select the one with the least Frobenius norm. Some results of preprocessing are shown in Figure 4.

4.2 Training

For training, we stack the preprocessed images belonging to each character class to form a mode-3 tensor. For each tensor we then apply rank-1 decomposition with a specified number of rank-1 matrices and we keep this number same across all the classes. Hence the output of training is the specified number of rank-1 matrices that form the subspace basis for each class along with a set of scalar coefficients yielding the mixing vector for each image in that class. Figure 3 top part illustrates the training process. The sensitivity of number of rank-1 elements to the accuracy on test data is discussed in (Section 5.3.2).

[3] http://www.imagemagick.org

4.3 Classification

To classify a test image, we first preprocess it and then project it onto the rank-1 subspace of each character class. The projection here means to get inner product between the rank-1 factors and the test image to get the mixing coefficients by the expression: $\lambda_i = v^T A_i^T u$; where λ_i is the i^{th} mixing coefficient and A_i is the corresponding residual image.

When $i = 1$, A_i = given test image. For $i \geq 2$, $A_i = A_{i-1} - \lambda_{i-1} u v^T$. In this way, we get 62 vectors for each test image (one per each class). We then measure similarity of each test vector using the inner product with the training vectors of each class and record the maximum. The final classification is given by taking the maximum over all classes. The process is illustrated in the bottom part of Figure 3.

5 EXPERIMENTS

We evaluated our approach on three popular scene character datasets Chars74K[4], ICDAR[5], and SVT-CHAR[6]. We used various experimental settings to report our results on these datasets. We also compare our method with several baseline methods in scene character recognition.

5.1 Datasets

The English subset of Chars74K dataset consists of 12503 characters. Characters have been cropped from 1922 images of advertisement signs and products from stores etc. The dataset is not split in training and testing sets, rather the authors give their proposed training and testing splits for comparison with their results. There is, however, a split between '*GoodImg*' and '*BadImg*' and as obvious from the names, the respective splits contain 'good' and less noisy (7705 images) as well as 'bad' more noisy images (4798 images) for a total of 12503 images.

The ICDAR2003 robust character dataset contains 11615 images of cropped scene characters and the dataset comes split into training and testing subsets. Characters have mostly been cropped from images of books titles, storefronts and signs and exhibit great variability in terms of resolution, illumination, color, etc. The test set has 5340 images in total but those belonging to 62 classes (A-Z, a-z, and 0-9) are just 5379.

The SVT-CHAR dataset consists of 3796 character images cut out and annotated by (Mishra et al., 2012) from cropped word images of the Street View Text (SVT) dataset. The original SVT dataset was harvested from Google Street View images of businesses and storefronts by (Wang et al., 2011). There is no training portion of SVT-CHAR and results have been reported only using it as a test set.

5.2 Results

For all experiments, we report results using 500 rank-1 factors for tensor decomposition. The impact of number of rank-1 factors on accuracy is further discussed in Section (5.3).

In our first experiment, we used the whole ICDAR2003 training set to get the rank-1 factors for each class of characters. The accuracy on the test set was 69% (see Table 1). In Figure 5, the lines parallel to the main diagonal of the confusion matrix reflect ambiguities due to character case, e.g. small case 'c' confused with 'C', etc. (Wang et al., 2012) reported accuracy of 83.9% on a modified version of the ICDAR2003 test set, but they re-cropped all images for their experiments and their set contains 5198 images, which is less than those in ICDAR2003 test set. In Table 1, we also report our results on the training and test splits proposed by (de Campos et al., 2009) for Chars74K, viz., Chars74K-15, where the suffix '15' specifies the number of training and test samples to be used for the experiment.

Table 1: Recognition performance on ICDAR2003 and Chars74K datasets.

Method	ICDAR	Chars74K-15
GB+NN (de Campos et al., 2009)	41%	47.1%
HoG+NN (Wang and Belongie 2010)	51.5%	58%
SYNTH+FERNS (Wang et al., 2011)	52%	47%
NATIVE+FERNS (Wang et al., 2011)	64%	54%
MSER (Neumann and Matas, 2011)	67%	-
Proposed RANK-1	**69%**	**57.1%**

[4] http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k

[5] http://algoval.essex.ac.uk/icdar/datasets.html

[6] http://vision.ucsd.edu/~kai/svt

Our second experiment was on SVT-CHAR. Since this is just a test set, therefore, we formed its training by combining the training sets of ICDAR2003 and the 'GoodImg'portion of the Chars74K dataset. We trained our factors on all 62 classes, to fairly compare our results with the reported ones, despite the fact that the SVT-CHAR does not contain any digit classes. We got 64% accuracy and the results are shown in second column of Table 2.

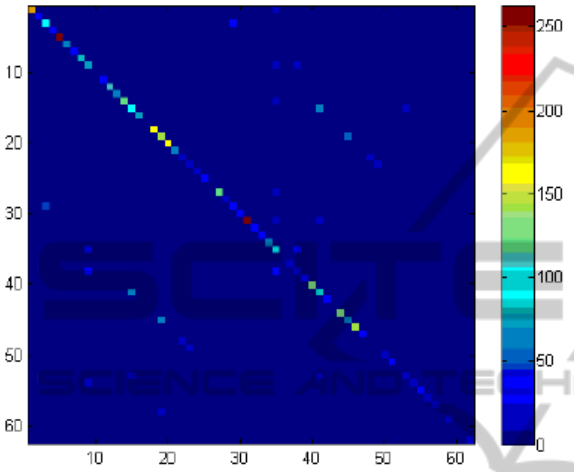

Figure 5: Confusion matrix for ICDAR2003 test set Numbers 1-62 show character classes A-Z, a-z,0-9. Lines parallel to the main diagonal show character confusions.

Table 2: Recognition performance on Chars74K-15 test split.

Method	Chars74K	SVT-CHAR
ABBYY FineReader[7]	31%	15.4%
GB+NN (de Campos et al., 2009)	54.3%	-
HoG+SVM (Mishra et. al., 2012)	-	61.9%
MSER (Neumann and Matas, 2011)	71.6%	-
Proposed RANK-1	**68.5%**	**64%**

In our third experiment, we used Chars74K-15 test split while training on all Chars74K but those images that are in the test set. Column 1 of Table 2 shows some improvement in accuracy as compared with other baseline methods and our earlier results given in Table 1.

The above results clearly show that our approach does better if given more training samples which prompted us to do another experiment with leave-random-one-out cross-validation (CV) setting. We show results in Table 3 for both Chars74K and ICDAR2003. For ICDAR2003 we combined training and testing sets to get one big set for CV.

The results show median accuracy over 100 trials. The accuracy of 72.5% is the best result we are aware of on the whole Chars74K dataset (including both GoodImg and BadImg sets). On the other hand the results on ICDAR2003 under this setting show improvement and further propound our observation.

Figure 6: Letters correctly recognized by our method.

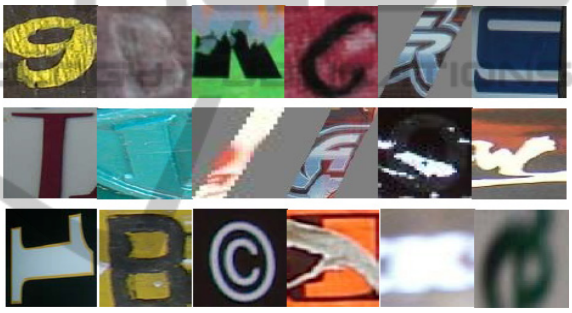

Figure 7: Some of the characters that could not be correctly recognized due to shape ambiguities, low contrast, occlusion, imperfect cropping, large rotations etc.

Table 3: Recognition performance using leave-random-one-out cross-validation (CV).

Method	ICDAR	Chars74K-15
Proposed RANK-1 + CV	76%	72.5%

Figure 6 shows some test samples from different datasets that our approach correctly recognized. Figure 7 shows the cases where our method failed. Some images here are not even easy human observers to recognize correctly due to low contrast, shape ambiguities, noise etc.

5.3 Discussion

5.3.1 Case Sensitivity: Why It Is Important?

Cropped characters from natural scene images exhibit extreme shape similarities for some classes. We

[7] http://www.abbyy.com

identified at least ten classes from English alphabet and two digit classes 0 and 1 that can become ambiguous in the absence of contextual clues. Consider Figure 8. It is extremely hard even for humans to distinguish between the pair of characters due to case similarity. Digit 'zero' and 'one' are also sometimes confused with letter 'O', 'I', and 'l (ell)' due to shape similarities. We, however, don't take into account digit and letter confusion and report only letter case insensitive accuracies in Table 4.

Figure 8: Case ambiguities in natural scene characters. Top row shows upper case while the bottom row shows lower case letters from ICDAR2003 dataset.

Moreover, from a recognition standpoint, the case distinction is immaterial even if eventually we are to recognize words from characters, unless we use case information to mark word boundaries. Table 4 shows that we get accuracy boost over our corresponding results in Tables 1 through 3.

Table 4: Recognition performance after removing case sensitivity.

Method	ICDAR	Chars74K-15
Proposed RANK-1	80%	66%
Proposed RANK-1 + CV	84%	78.7%
Method	Chars74K	SVT-CHAR
Proposed RANK-1	75.1%	73%

5.3.2 Rank-1 Elements and Number of Samples

We give number of rank-1 elements as input to the decomposition algorithm. Figure 9 shows the effect of the choice of rank-1 elements on accuracy for ICDAR test set and Chars74K set when tested on the proposed test split by (de Campos et al., 2009). As noted by (Shashua and Levin, 2001), addition of rank-1 elements helps capture temporal redundancies in the input image set that in our case occur due to font and shape changes. This eventually gets to increased accuracy. However, as shown in Figure 9, a point comes after which we get a kind of stagnation in accuracy. We, therefore, empirically fixed the number of rank-1 elements to '500'.

As mentioned (in Section 5.2) above, the number of training samples per class also played an important role in boosting accuracy on the test set. We empirically observed that unless we add good (or less noisy) images to the training set, the decomposition process would be affected by the presence of even a small number of noisy images. This can be explained by the fact that as the number of less noisy images increases, the additive effect of noisy images is reduced and a good pattern of variation in character's font and shape becomes available and is effectively captured in the spatial and temporal components of the rank-1 decomposition. This is the reason why we used just the '*GoodImg*' part of Chars74K when we combined ICDAR2003 and Chars74K to train for SVT-CHAR. To further demonstrate this fact, we plot in Figure 10 the accuracy gain with increasing number of training images for the character class 'A'

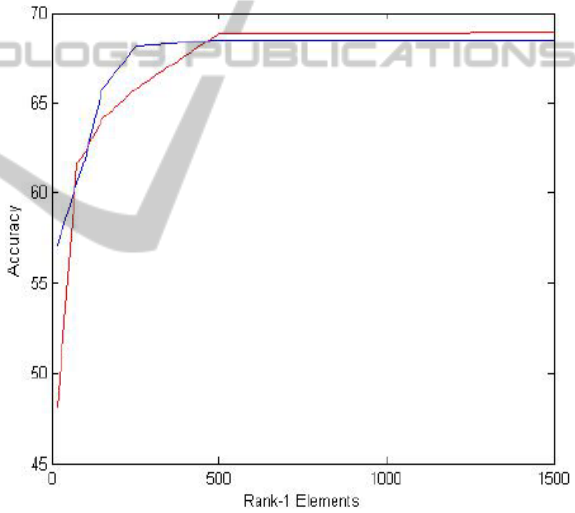

Figure 9: Number of Rank-1 Elements vs. Accuracy on ICDAR (shown in red) and Chars74K (shown in blue).

(this trend is also true for all other character classes). The accuracy here represents performance measured on the test samples of 'A' from (de Campos et al., 2009) test split after training over different number of available training samples of character 'A' from Chars74K. We varied the sample count from '15' (given in de Campos et al. training split) to '659' (all samples of 'A' excluding the '15' given in the test split). The plot validates our observation about gain in accuracy with the increase in number of samples for natural scene character recognition.

6 CONCLUSIONS

We proposed a holistic approach to solve natural scene character recognition that avoids dependency on specific features. Our method is based on multi-image tensor decomposition similar to (Shashua and Levin, 2001) with modification as to the way we get rank-1 matrices for natural scene images that contain a lot of variations and noise. Through our results we showed the potential of using image tensor decomposition to better capture shape and font variations in scene character images. We got better results than several baseline methods and achieved improved recognition performance on the datasets using leave-random-one-out cross-validation, justifying thus our intuition of the importance of feature-independency and preservation of spatial correlations in recognition.

In future we hope to get state-of-the-art performance using better image segmentation methods and also plan to incorporate recent advances in tensor decomposition domain in solving other sub problems of scene text recognition.

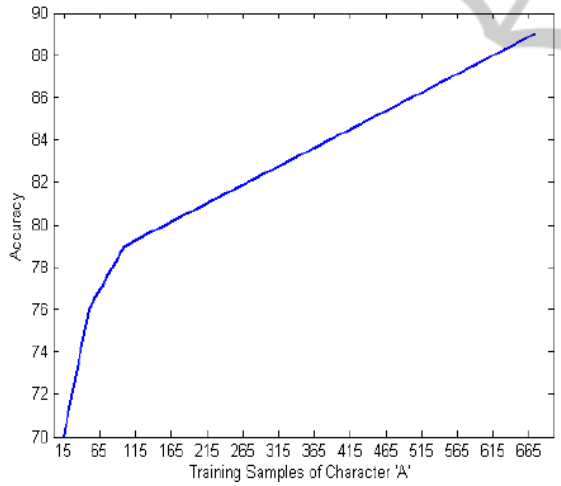

Figure 10: Accuracy vs. Number of training samples of 'A' from Chars74K dataset.

REFERENCES

Chen, X., and Yuille, A., 2004. Detecting and reading text in natural scenes. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004.* IEEE 2004. Vol. 2. pp. II–366.

Coates, A., Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang, T., Wu, D., and Ng, A., 2011. Text detection and character recognition in scene images with unsupervised feature learning. In *International Conference on Document Analysis and Recognition (ICDAR), 2011.* IEEE 2011, pp. 440–445.

Dalal, N., and Triggs, B., 2005. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision and Pattern Recognition (CVPR) 2005.* IEEE 2005, pp.886-893.

de Campos, T.E., Babu, B. R., and Varma, M., 2009. Character recognition in natural images. In *VISAPP (2)*, 2009, pp. 273–280.

Donoser, M., Bischof, H., and Wagner, S., 2008. Using web search engines to improve text recognition. In *19th International Conference on Pattern Recognition, ICPR 2008.* Vol. no. 14, pp. 8-11.

Hazan, T., Polak, S., and Shashua, A., 2005. Sparse Image Coding using a 3D Non-negative Tensor Factorization. In *International Conference on Computer Vision (ICCV), 2005.* IEEE 2005. Vol. 1, pp. 50–57.

Field, J., and Learned-Miller, E., 2013. Improving Open-Vocabulary Scene Text Recognition. In *International Conference on Document Analysis and Recognition (ICDAR) 2013.* IEEE 2013, pp. 604–608.

Kita, K., and Wakahara, T., 2010. Binarization of color characters in scene images using k-means clustering and support vector machines. In *International Conference on Pattern Recognition (ICPR), 2010.* IEEE 2010, pp. 3183–3186.

Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., and Young, R., 2003. ICDAR 2003 robust reading competitions. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition 2003.* IEEE 2003. Vol. 2, pp. 682–687.

Mishra, A., Alahari, K., and Jawahar, C., 2012. Top-down and bottom-up cues for scene text recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012.* IEEE 2012, pp. 2687–2694.

Mishra, A., Alahari, K., and Jawahar, C., 2011. An MRF model for binarization of natural scene text. In *International Conference on Document Analysis and Recognition (ICDAR), 2011.* IEEE 2011, pp. 11–16.

Nagy, R., Dicker, A., and Meyer-Wegener, K., 2011. NEOCR: A Configurable Dataset for Natural Image Text Recognition. In *CBDAR Workshop*, ICDAR 2011, pp. 53-58.

Neumann, L., and Matas, J., 2011. A method for text localization and recognition in real-world images. In *Computer Vision–ACCV 2010*, pp. 770–783.

Niblack, W., 1985. An introduction to digital image processing. Strandberg Publishing Company.

Otsu, N., 1979. A Threshold Selection Method from Gray-Level Histogram. In *Trans. System, Man and Cybernetics*. IEEE 1979. Vol.9, pp.62-69.

Shashua, A., and Levin, A., 2001. Linear Image Coding for Regression and Classification using the Tensor-rank Principle. In *International Conference on Computer Vision and Pattern Recognition (CVPR), 2001*. IEEE, 2001. Vol. 1, pp. I-42 - I-49.

Sun, C., Junejo, I. N., and Foroosh, H., 2011. Action Recognition using Rank-1 Approximation of Joint Self-Similarity Volume. In *International Conference on Computer Vision (ICCV) 2011*, pp. 1007-1012.

Wang, T., Wu, D., Coates, A., and Ng, A., 2012. End-to-End Text Recognition with Convolutional Neural Networks. In *International Conference on Pattern Recognition (ICPR), 2012*. IEEE 2012, pp. 330.

Wang, K., Babenko, B., and Belongie, S., 2011. End-to-end scene text recognition. In *International Conference Computer Vision (ICCV), 2011*. IEEE 2011, pp. 1457–1464.

Wang, K., and Belongie, S., 2010. Word spotting in the wild. In *Computer Vision–ECCV 2010*, pp. 591–604.

Weinman, J., Learned-Miller, E., and Hanson, A., 2009. Scene text recognition using similarity and a lexicon with sparse belief propagation. In *Pattern Analysis and Machine Intelligence TPAMI*. IEEE Transactions 2009. Vol. 31, no. 10, pp. 1733–1746.

Xianqian, L., and Sidiropoulos, N.D., 2001. Cramer-Rao lower bounds for low-rank decomposition of multidimensional arrays. In *Transactions on Signal Processing*. IEEE 2001. Vol. 49, Issue 9, pp. 2074-2086.

Yokobayashi M., and Wakahara, T., 2005. Segmentation and Recognition of Characters in Scene Images Using Selective Binarization in Color Space and GAT Correlation. In International Conference on Document Analysis and Recognition (ICDAR), 2005. IEEE 2005. pp. 167–171.