World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:1, 2015

Recognition of Grocery Products in Images Captured by Cellular Phones

Farshideh Einsele, Hassan Foroosh

***Abstract*—In this paper, we present a robust algorithm to recognize extracted text from grocery product images captured by mobile phone cameras. Recognition of such text is challenging since text in grocery product images varies in its size, orientation, style, illumination, and can suffer from perspective distortion. Pre-processing is performed to make the characters scale and rotation invariant. Since text degradations can not be appropriately defined using well-known geometric transformations such as translation, rotation, affine transformation and shearing, we use the whole character black pixels as our feature vector. Classification is performed with minimum distance classifier using the maximum likelihood criterion, which delivers very promising Character Recognition Rate (CRR) of 89%. We achieve considerably higher Word Recognition Rate (WRR) of 99% when using lower level linguistic knowledge about product words during the recognition process.**

***Keywords*-Camera-based OCR, Feature extraction, Document and image processing.**

I. Introduction

THE use of digital cameras to capture text from natural images is gaining an emerging interest. Digital cameras are easy-to-use versatile tools and are likely to replace classical scanners in OCR applications. Scanner-based OCR technology has made significant progress delivering very accurate results and is affordable at low-prices. However, text extracted from digital cameras in natural unconstrained images is more complex to recognize as it varies in size, style, orientation, illumination and can suffer from perspective distortion and can be surrounded by shadows or even partly occluded. Therefore, camera-based OCR proves to be very challenging and can rarely make use of existing methods in the scanner-based OCR technology. Some existing works in the literature on camera-based OCR concentrate on location and rectification of text areas [1] and rely on the application of classical OCRs for the character recognition. This method can only be used in limited scenarios where such OCRs perform well. Consequently, the recognition of text in camera-based natural images depends on the specific underlying scenario. Other research activities report on applying different algorithms for character recognition from that used in the classical OCRs.
Generally speaking, in a statistical character recognition approach, feature extraction and classification methods are the key components of such a system. Therefore feature extraction methods play a crucial role in properly discriminating the character classes. In addition, our experiments have shown that character recognition accuracy in camera-based images is strongly dependent upon two factors: 1) the chosen feature extraction method and 2) the underlying database. Trier et al. in [2] provide a thorough survey of feature extraction methods for character recognition. One important feature extraction method is using invariants. Invariants have approximately the same values for samples of the same characters that differ by geometric transformations like translation, rotation, scaling, stretching, mirroring, shearing and skewing of characters in a specific application. Lu and Tan in [3] propose three perspective invariants including character ascender and descender, character centroid runs and water reservoirs. They use a CART structure to classify lowercase characters using their perspective invariants and perform experiments on 40 document images from books, the web and proceedings using a digital camera of 7 mega pixel. Since gathering and annotating a large number of images from natural scenes can be costly and time-consuming, many researchers have acquired a database containing synthetic data having exactly defined geometric transformations like rotation, shearing, scaling and translation. Ucida et al. in [4], [5] and [6] propose another set of invariants for camera-based character recognition that are tolerant to geometric transformations and perspective deteriorations. Their proposed invariants are area ratios and cross ratios for each character pattern. They produce original characters of same height using some computer fonts and extract their invariants and test the robustness of their method on a synthetically affine transformed character database. Flusser and Suk in [7] and [8] report promising results using affine moment invariants as features to recognize characters in a synthetic database containing degraded computer fonts using combined affine transformations like shearing and rotation. Although invariants are powerful features in the above mentioned studies, they may fail when dealing with the degree of variability of character styles faced in a real world database. Other features can be extracted from grayscale images. Sun et al. in [9] report a grayscale feature extraction method based on dual eigenspace decomposition. They perform their classification experiments on a large training set containing synthetic degraded characters with different degradation levels. However, the discriminative power of the proposed feature extraction method strongly relies on a synthetic training set. Other feature extraction methods work only on binary images like those gained from binary contours. Zhang et al. in [10] extract shape signatures like central distance, complex coordinates, curvature and cumulative angle functions from shape contours. Since shape invariance is difficult to achieve in the spatial domain, they transform their shape signatures

F. Einsele is with the Section Business of Bern University of Applied Sciences,3005 Bern, Switzerland (e-mail: farshideh.einsele@bfh.ch).
H. Foroosh is with School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA (e-mail: foroosh@cs.ucf.edu).

into Fourier domain and ignore the phase information to gain rotation invariant features. Scale and translation invariance is gained with normalization and considering the centroid as the center of the coordinate system. The experimental results show that the centroid and complex coordinate signatures have a high precision and recall rate whereas the curvature and cumulative angular functions deliver unreliable results. Dionisio et al. in [11] also report a contour-based shape classification technique based on polygon approximation that is invariant under rotation and scaling. The vertices of polygon approximation are formed by high curvature points of the profile and are selected by the Fourier transform of the object contour. A series of features are computed from the polygonal approximation and a minimum distance classifier is used for object recognition. Although such contour-based invariants deliver promising results using Fourier descriptors for character recognition and are also reported in the survey of Trier et al. [2], the reported test and training databases are synthetically deformed patterns and the features are invariant with respect to translation, scaling, rotation and do not consider other transformations coming from real world captured images (e.g. shearing, shadowing, bad illumination and perspective distortion). Besides Trier et al.state in [2] that a statistical classification system should consider the so called *curse of dimensionality* meaning that it should be training-based containing a minimum number of patterns that is 8-10 times bigger than the number of the chosen features. As already stated, when dealing with a database containing real world character images, database generation is an expensive and time-consuming drawback. To sum up, the above mentioned works have been performed either by considering a synthetically degraded database or are training-based approaches. In this paper, we present a method for camera-based character recognition that uses a small *real world* database extracted from images of grocery products captured by a cellular phone with a resolution of 5 mega pixels. The presented method does not need a training set that should rely on the previously described term of *curse of dimensionality*. Therefore our proposed method can be applied directly to the extracted text with no use of cost-intensive image enhancement algorithms and delivers promising results. The remainder of this paper is organized as follows: in section II, we introduce the specificities of text in product images. Section III reports about our proposed character recognition algorithm including used feature extraction and classification methods. Section IV presents our evaluation results and section V is about our conclusions and a short sketch of our future works.

II. Challenges of Product text recognition

We extract text from images taken with cell phones from grocery products. Text extraction from camera-based images is a relatively well researched area with plenty of existing works in the literature [12]. However, text extraction methods from camera-based images is tightly related to a specific application and there does not exist a valid generic method for the extraction of text within different camera-based scenarios. We therefore use a text extraction algorithm that has been developed for the specificities of the text from grocery products and is explained in detail in [13]. The resolution of the used cell phone camera is 5 mega pixels and the images are taken from different angles with the camera having a similar distance to products as the one a common grocery shopper would have when he crosses grocery aisles. The extracted text has mostly a height between 20-50 pixels and characters can be mostly labeled and segmented using connected components algorithms. Table I shows some extracted words in our database.

TABLE I
EXTRACTED GROCERY PRODUCT WORDS

A. Character Specifications

As can be seen in Table I, some text, especially with small height or cursive style can contain characters that touch adjacent ones and therefore character segmentation prior to recognition would not be applicable in such cases.. However, in this initial study, we decided to concentrate on the cases where characters are higher than 20 pixels and are not cursive. As previously stated, such characters can be isolated using connected components labeling. We labeled the extracted characters and built a character database for our further investigations. The isolated characters show different kind of variabilities as shown in Table II. The first row of Table II contains the original characters. The second row shows their contours and the third row their skeletons. One can see that the the variability in character contours and skeletons can not be described by geometric transformations such as scaling, translation, rotation, affine transformation and other perspective transformations. Therefore, as stated in the previous section, a feature extraction method using such well-known invariants like aspect ratios or moment invariants are not discriminative enough for our character recognition task. We report in the next section about our performance investigations of such invariants.

III. Character recognition algorithm

A. Blockdiagram

Fig. 1 shows the blockdiagram of the proposed character recognition approach

TABLE II
VARIABILITIES OF CHARACTER "A" IN DIFFERENT GROCERY PRODUCT TEXT

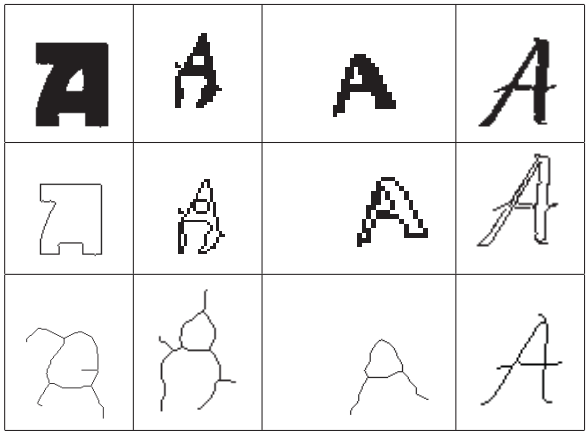
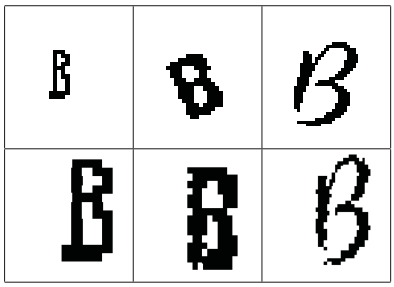


Fig. 1. Blockdiagram Character Recognition

B. Pre-processing

The original character images as shown in Table II show a large degree of variability regarding their scales, orientations and perspective transformations. We normalize the images to the same scale and rotate them with an angle that is calculated based on the difference of their major axis with x or y axis, to achieve similarity for scale, translation, rotation, affine and perspective transformation. Table III shows the result of the normalization process. We normalized the images by the scale of 50 x 50 pixels.

TABLE III
NORMALIZED IMAGES OF CHARACTER 'B'



C. Feature Extraction

Feature extraction has shown to be the most challenging part of our character recognition approach. As previously explained, there exist a big variability for the same character. Thus, we sought for invariant features with a powerful discriminative nature. As already stated in the introduction, one can find various proposed feature extraction methods suitable for character recognition like the Fourier transformation of image contours or their skeletons, since rotation invariance can be achieved in Fourier space when ignoring the phase information and using the magnitude as shape descriptors. However, our experimental results delivered a character recognition rate lower than 50% when using such Fourier descriptors as features. In addition, we calculated the high curvature points and used various proposed methods like polygon approximation of the character shape in [11] or using Fourier descriptors of such points as proposed in [10]. In both cases the accuracy was again under 50% which is not an acceptable OCR-result. Additionally, we performed experiments using the moment invariants like Hu's moments, affine moment invariants and rotations-invariant moments such as Zernike moments. The images had been first normalized to be translation and scale-invariant. However, the obtained recognition rates were again unsatisfactory. We thus decided to use the entire set of image points as our features. Therefore we obtained a feature vector containing 50 x 50 = 2500 elements.

D. Classification Method

Assuming the data in each class is normal distributed, the class means and variances for each feature are initially estimated. Let $x = (x_1, x_2, ..., x_n)$ denote a feature vector with n feature values (in our case $n = 2500$) and σ_{ij}^2 and μ_{ij} denote the variance and mean value of feature x_j and class w_i, respectively. Then the corresponding marginal pdf is given by

$$f_{ij} = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp - \frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \quad (1)$$

When using the Bayesian decision theory using multivariate Gaussian distribution, it is shown in [14] that minimum-error-rate classification can be achieved by use of the discriminant functions

$$g_i(x) = -(x - \mu_i)^T \Sigma^{-1} (x - \mu_i) - log|\Sigma_i| + logP(\omega_i) \quad (2)$$

with ω_i being the ith class, μ_i its mean vector and Σ_i and $P(\omega_i)$ its covariance matrix and prior probability, respectively. The middle term in 2 is a constant and has the same value for all the classes and can be omitted. If all classes have the same priors $P(\omega_i)$, the last term can also be omitted. In this case, the discriminant functions are simplified as

$$g_i(x) = -(x - \mu_i)^T \Sigma^{-1} (x - \mu_i) \quad (3)$$

The discriminant functions signify that after calculating the variance-normalized Euclidean distances to all class centers, the sample vector x is assigned to the class with the shortest distance.

- **Modified discriminant function for statistically insufficient data**

  Since we assumed that the class pdf is a multivariate Gaussian distribution, we need to calculate mean vector and the covariance matrix to calculate the discriminant function in (3). When assuming an identity covariance matrix, our experiments delivered more reliable results. In

contrast, when including the covariance matrix computed from the data in (3), the results were worse. The explanation can be twofold: first we have a small database of 500 extracted words where some characters like 'Q', 'q', 'X', 'x' 'Z' and 'z' are under-represented, since such characters rarely occur in a product text. Second covariance is a second order statistics parameter that shows to be more noise sensitive than mean vector (first order) particularly when used in such a small data base such as ours. (3) is then simplified as

$$g_i(x) = -(x - \mu_i)^T (x - \mu_i) \quad (4)$$

IV. Evaluation results

A. Isolated Character Recognition

We performed experiments on a database containing 1,500 isolated characters using our simplified discriminant classifier as described in previous section. Fig. 2 shows an image of the confusion matrix of all 52 uppercase and lowercase characters. The diagonal nature of the confusion matrix clearly shows the high accuracy of the obtained results. The overall CRR is 89%. This rate is achieved by ignoring the confusion errors between some upper and lowercase characters that are of same shape (e.g. 'C'/'c' or 'O'/'o'). The justification here is that uppercase and lowercase of such characters differ only in their scale but they have been normalized to become scale-invariant in a first step and thus belong literally to the same class.

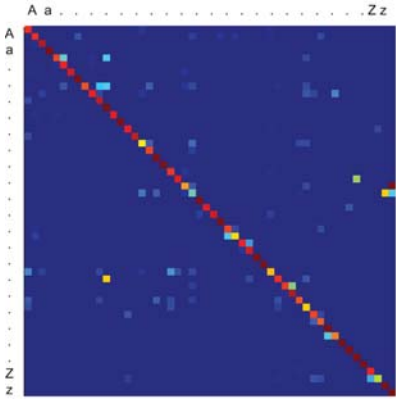


Fig. 2. Confusion Matrix of 52 uppercase and lowercase characters

B. Word Recognition

In order to improve the recognition rate of isolated characters, we built a dictionary of words including the ones in our database. The used dictionary includes between 500-1000 product words. Additionally, we built an error database of words containing characters that were recognized erroneously. By scanning through the database we calculate the probability of each *"wrong"* word being the *"right"* word in our dictionary provided that both words have the same character length. We obtain a significantly higher total WRR rate of approx. 99%. Thus using a 'dictionary' of products as a knowledge base shows to be highly beneficial to the recognition of product text images.

V. Conclusion and future works

We have introduced in this paper a method to recognize extracted text from product images obtained from a cell phone camera with a resolution of 5 mega Pixels. We have found that shape and moment invariants are not discriminative enough in our case. Invariants deliver promising results when using a synthetic database containing degraded text using *well-defined* geometric transformations like affine, rotation, shearing, scaling and translations. Our database instead contains *real world* text from product images that are not related by *well-defined* geometric transformations. Such extracted text can contain geometric transformations but can additionally suffer from other sources of noise like bad illumination, perspective distortions, or even shadows or occlusions. As a result, We have used after a pre-processing step the entire character points as our feature vector. Our experiments delivered a promising CRR of 89% for all characters including the ones in our database. Additionally, we gained a significantly higher WRR of above 99% by using a product word dictionary in the recognition process. However, the used dictionary has a relatively small sample size (500-1,000 words) and consequently the achieved WRR may drop when the size of the product dictionary will be increased to the size similar to a common world dictionary (some 10,000 words), since the probability of having multiple *right* words with the same character length increases. In the future work, we plan to combine the presented recognition method with our text extraction method introduced in [13] to simultaneously extract and recognize text in product images.

Acknowledgment

The authors would like to thank Swiss National Foundation for their friendly support of this project.

References

[1] M. Mirmehdi and P. Clarck, "Recognising text in real scenes," in *International Journal on Document Analysis and Recognition (IJDAR)*, 2001, pp. 243–257.
[2] O. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition – a survey," *Journal of Pattern Recognition, Elsevier ScienceICPR06)*, vol. 29, pp. 641–662, 1996.
[3] S. Lu and C. L. Tan, "Camera text recognition based on perspective invariants," in *Proc. of the 18th International Conference on Pattern Recognition (ICPR06)*, 2006.
[4] S. Omachi, M. Iwamura, S. Uchida, and K. Kise, "Affine invariant information embedment for accurate camera-based character recognition," in *Proc. of the 18th International Conference on Pattern Recognition (ICPR06)*, 2006, pp. 1098–1101.
[5] S. Uchida and M. Iwamura, "Data embedding for camera-based character recognition," in *Proc. of the 18th International Conference on Pattern Recognition (ICPR06)*, 2006, pp. 1098–1101.
[6] S. Uchida, M. Iwamura, S. Omachi, and K. Kise, "Ocr fonts revisited for camera-based character recognition," in *Proc. of the 18th International Conference on Pattern Recognition (ICPR06)*, 2006.
[7] J. Flusser and T. Suk, "Pattern recognition by affine moment invariants," *Pattern Recognition*, vol. 26, pp. 192–195, 1993.
[8] J. Flusser and T. Suk, "Graph method for generating affine moment invariants," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR04)*, 2004, pp. 167–174.
[9] J. Sun and S. N. Y. Hotta, Y. Katsuyama, "Camera based degraded text recognition using grayscale feature," in *Proc. of the 8th International Conference on Document Analysis and Recognition (ICDAR05)*, 2005.

[10] D. S. Zhang and G. Lu, "A comparative study on shape retrieval using fourier descriptors with different shape signatures," in *In Proc. of International Conference on Intelligent Multimedia and Distance Education (ICIMADE01)*, 2001, pp. 1–9.
[11] C. Dionisio and H. Kim, "A supervised shape classification technique invariant under rotation and scaling," in *Intl Telecommunications Symposium*, 2002.
[12] K. Jung, K. I. Kim, and A. Jain, "Text information extraction in images and video: A survey," in *Pattern Recognition*, vol. 37, 2004.
[13] F. Einsele and H. Foroosh, "Towards text extraction from low resolution cell phone images," in *submitted paper to IEEE International Conference on Image Processing (ICIP09)*, 2009.
[14] R. Duda and P. Hart, *Pattern classification and scene analysis*. Reading, MA: John Wisley & Sons, 1972.