Statistical Modeling of the Relation Between Characters and Diacritics in Lampung Script

Akmal Junaidi, René Grzeszick and Gernot A. Fink
Pattern Recognition in Embedded Systems Group
Department of Computer Science
TU Dortmund
Dortmund, Germany
Email: {Akmal.Junaidi,Rene.Grzeszick,Gernot.Fink}@udo.edu

Szilárd Vajda
Lister Hill National Center for Biomedical Communications
US National Library of Medicine
National Institutes of Health
USA
Email: szilard.vajda@nih.gov

***Abstract*—Lampung Script is a non-cursive script where a rich set of diacritics is used to modify the syllable denoted by a character symbol. Consequently, the analysis of the relation between characters and diacritic marks associated with them plays an important role in the recognition process. As diacritics can appear in three different relative positions with respect to a character (top, bottom, and right) associating them correctly with a character is a challenging problem. In this paper we propose a novel approach for modeling the relations between characters and diacritics in handwritten Lampung documents. First, a document is segmented into characters and diacritic marks. Then every character defines a normalized coordinate system into which nearby diacritics can be mapped. The relation between a diacritic mark and its associated character can then be described by a statistical model. In a writer independent experimental evaluation we investigate models with different degrees of specialization with respect to their capability of predicting the correct character-to-diacritic associations. We achieve significant error rate reductions with respect to a naive association model using a nearest-neighbor criterion.**

I. Introduction

Handwritten character recognition systems have been progressively developed for more than 40 years. During this time, many researchers devoted their ideas on various scripts like Arabic [1], Chinese [2], Bangla [3], [4], Farsi [5], etc. Recently, many approaches have been developed and used for some key research areas like line extraction, word spotting, character recognition, writer identification, etc. Most of this research focused on the characters while ignoring particular marks around the characters, the so-called accents or diacritics. Only very limited research work was dedicated to characters with diacritics [6]. However, these marks play an important role to change the meaning or pronunciation of a word. Many languages from all around the world uses these symbols as an integral part of their writing system. For example, French, Greek, German, Czech, Hungarian, Spanish, Portuguese and Turkish in Europe, Arabic in the Middle East or Indic scripts, Vietnamese and Lampung in Asia.

While accents or diacritics occur only rarely in some of these scripts, they play a major role in others. For example, Lampung is a script written in the Lampung province in Indonesia, which is employing a rich collection of diacritics as an important part of the script (see Fig. 1 for illustration).

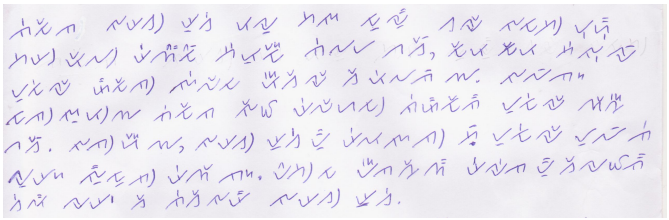


Fig. 1. An example of a handwritten Lampung document.

Unlike other scripts that have only top and/or bottom diacritics, this script has diacritics lying on the right side of the character as well. Moreover, diacritics can be found not only on one position around a character, but also at combinations of these locations. In this paper we will introduce a novel method for associating diacritics at varying positions with a Lampung character. Besides the limited research on diacritics in general, the extensive usage of diacritics in Lampung is a major motivation for our work on this script.

The rest of this paper is organized as follows. The next section discusses the related works with respect to diacritics and Lampung script. Then, Section 3 provides an introduction to the script particularly about characters, diacritics and syllables. Section 4 illustrates the novel recognition approach, followed by the experiments and results. The final section contains the conclusion of this research.

II. Related Works

In the field of handwritten character recognition, the characters and diacritics are generally considered as two different separate entities. Therefore, a document is commonly segmented into two parts: The characters and the diacritics [1], [6]. However, for analysis, the character and its diacritics should be handled as one. In fact, a separated character and its diacritics have to be associated with each other. The composition of a character with diacritics which can be called a compound character, will increase the complexity of the recognition task. Extra efforts should be proposed to address them correctly.

Work on characters and diacritics has been done before for French handwriting [6]. In this work, the character and

diacritic were separated and handled as a single piece of the component. Each of them was recognized independently. Then, to form compound characters, each vowel in the test set was combined with diacritics (if applicable). For recognition, the authors did not build a specific system, but they realized it based on handwritten character recognition for lower case characters and a recognizer for the four diacritics. The recognition rate in their experiments depends on lower case recognition rate, diacritic recognition rate and segmentation of characters and diacritics. Applying their approach, they acquired 93.5% recognition rate for their artificial database and 92.7% for their local collection. In general, the recognition rate of compound characters is almost the same as the recognition rate of the lower case characters.

Similar work has been conducted on Vietnamese [7] that also uses Roman scripts. Four diacritics are used for creating additional sounds and five others to control the tone of the word. The tonal diacritics will guide the voice of a speaker like the low, high, sharp, fall, rise in tone and it effects to the words meaning. The idea of the work is the usage of a Modified Optimized Cosine Descriptor (MOCD) with appropriate sampling algorithms to represent multiple strokes of a character in a single feature set. Then this MOCD was fed into a handwritten character recognition system which is a three layers processing classifier. The first layer is intended to classify the main character, the second is used for classifying circumflex diacritics, and the last one is for recognizing tonal diacritics. The average recognition rates of main characters is 87.79%, circumflex, $93.37 - 99.91\%$ and tonal diacritics 93.88% respectively.

Pure diacritic processing work has been carried out on Arabic handwriting [1]. Instead of utilizing the main character, the authors only relied on the features of the diacritics for identifying the writer of a document. Each diacritic was read as input data and then a Linear Binary Pattern (LBP) histogram was calculated. All diacritic LBPs with respect to each writer were concatenated in order to obtain a complete feature representation for the associated writer. For identifying the writer, a distance metric between the LBP histograms of the unknown writer and LBP histograms of the writers in the database was evaluated. The decision was made after examining the minimum distance between the LBP histograms. Testing on the IFN/ENIT database [8] showed that their approach could achieve 97.56% accuracy for 287 different writers.

In our previous work, particular research on the Lampung handwritten character recognition has been addressed for semi-automatic labeling [9] and recognition [10]. In the first work, we manually assigned labels to only 0.5% of the training data, the rest of the labels were inferred automatically by the proposed method. Then a classifier was built on top of the inferred labels for recognizing the test set. For recognition, we proposed a water reservoir based feature set in order to recognize Lampung handwritten characters and achieved 94.27% accuracy for 18 classes.

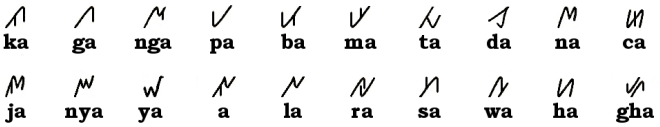


Fig. 2. Examples for all 20 characters of the Lampung script.

III. Overview of the Lampung writing

As one might see in Fig. 1, the Lampung script is not a cursive script. Both characters and diacritics can be visually distinguished. They can stick closely and form a compound character. These can be found in various configurations, from a simple configuration i.e. a base character without any diacritics, to a crowded configuration i.e. a compound character built by a character and a set of diacritics. In order to understand the concept of the Lampung script, the characters as well as the diacritics and and their usage for composing a syllable are explained in the following.

A. The character and diacritic

Lampung script contains 20 characters and 7 basic diacritics. In terms of geometric shape, all the characters have at least one curve as a dominant shape and some of them may have a short strip attached to the curve (see Fig. 2). The diacritics are much simpler in shape compared to the character.

If we observe them in more detail, three diacritic shapes are very similar to three characters, *ga* ( ), *pa* ( ), and *ha* ( ), respectively. The only distinction among them is their size. In general, the diacritics size is much smaller than the characters size, so that they can easily be discriminated by size-based filtering. For a visual illustration of the shape, the diacritics in various position are depicted in Fig. 3.

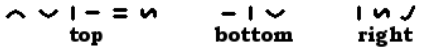


Fig. 3. The Lampung diacritics with their possible positions in the script.

B. Syllable construction

All characters of the Lampung script end with an 'a' sound. Therefore, a character without any diacritics will always generate a syllable with the ending sound 'a'. In order to change the pronounciation, the diacritics can be placed around the character. They can be positioned on the top, the bottom or the right of the character which then produces a specific pronounciation. Hence, the sound can be modified from a raw character syllable sound into various syllable sounds.

For example in Fig. 4a, a character *ba* ( ) without a diacritic will produce the syllable *ba*. It can become a syllable *bar* if an *s-like* diacritic is on the top of the character, or it becomes a syllable *bu* if there is only a *horizontal-line* diacritic on the bottom of the character. However with two diacritics as we can see in Fig. 4a, the syllable changes to *bur*.

A complete word is constructed by some syllables created by characters and diacritics. Thus, incorporating diacritics in the handwriting character recognition is very important for understanding the true meaning of a word.

IV. Proposed Approach

The main concern of our work is to analyse the assignment of a diacritic to the correct character among several characters nearby. Concerning that goal, in this section we describe the representation of a compound character in terms of diacritic-character pairs and their feature representation. We then introduce the statistical model for associating diacritics with characters.

A. Feature representation of diacritic-character pairs

The objective of our approach is to associate a diacritic with a character. Hence, the view of the data is no longer character-centered but diacritic-centered. For further analysis of compound characters, a diacritic and a corresponding character is considered as one pair. If there are more diacritics associated with one character each association is treated as an independent diacritic-character pair. A character without any diacritics will not be considered in the pairing.

Each diacritic-character pair is expressed in terms of a feature vector using the following steps: First, the character is located and its geometric center is computed. This center serves as the origin of a local Cartesian system for the compound character (see Fig. 4).

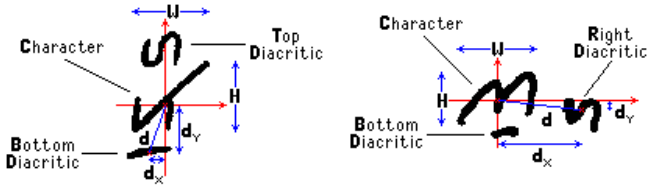


Fig. 4. Examples of Lampung syllables consisting of a basic character and diacritics: (a) the character *ba* with top and bottom diacritics generates the syllable *bur* and (b) the character *na* with bottom and right diacritics generates the syllable *nuh*.

We then examine the relation of a diacritic and its corresponding character by computing the distance between the origin and the gravity center of the diacritic. This distance is projected along the X axis (d_x) and Y axis (d_y). We normalize the projected distance measure by dividing it by the width (W) or height (H) of the associated character respectively, as defined by:

$$x = \frac{d_x}{W}, \qquad y = \frac{d_y}{H} \tag{1}$$

Both values can be rewritten as an ordered pair $v = [x, y]$, which serves as a feature vector of a diacritic and its corresponding character. The entity v represents a normalized position of a diacritic relative to the character as the central viewpoint. In more specific term, this entitiy symbolizes the relative position of the diacritic on the top, the bottom or the right side of the character.

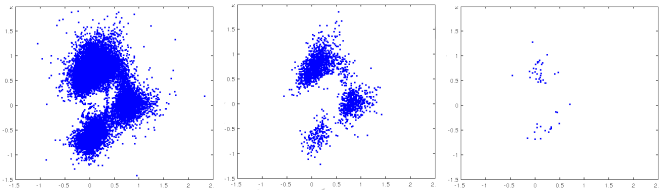


Fig. 5. Distribution of normalized diacritic positions: (a) for all characters in the training set, (b) for the character with the most frequent diacritics, and (c) for the character with the second least frequent diacritics.

B. Statistical model

As we represented the diacritic-character pair in form of a feature vector v, we can use this representation for creating a statistical model of these relations. All characters around a diacritic are considered as candidates for pairing (see Fig. 6). Note that for every pairing a different feature vector is computed, as the feature representation is based on the local coordinate system of the associated character.

For each of the nearby Lampung characters c_j the probability of a pairing can be estimated by a Gaussian mixture model:

$$P(v|c_j) = \sum_{i=1}^{k_j} w_{i,j} \mathcal{N}(v|\mu_{i,j}, \Sigma_{i,j}) \tag{2}$$

Where:
k_j: is the number of components for character c_j
$w_{i,j}$: the weight of component i
$\mathcal{N}$: the Gaussian normal distribution
$\mu_{i,j}$: the mean of the component i
$\Sigma_{i,j}$: the covariance of the component i.

Note that from the training data different means and covariances can be estimated for each character. We applied k-Means clustering [11] on the complete training data for computing an initial model. Then we used the EM-algorithm [12] to optimize the model parameters with respect to the different character specific distributions. Examples for different distributions of the feature vectors can be seen in Fig. 5.

For associating a diacritic with a character the s closest characters are considered. The diacritic-character pairing with the maximum conditional likelihood over all possible pairings is assumed to be the correct one:

$$s = \arg\max_{s} \left(P(v_s|c_s)\right) \tag{3}$$

Since there are Lampung characters that rarely occur in association with any diacritics, the estimation of the model parameters for these will be less reliable than for characters with a high number of samples. Hence, in a more general case the association between diacritics and characters can also be made based on a Gaussian mixture model that is estimated on

the complete dataset. This can be formulated as the marginal densitiy of $P(v_j, c_j)$ or approximated by estimating the model parameters character independently:

$$P(v) = \sum_{j=1}^{|c|} P(c_j)P(v|c_j) \approx \sum_{i=1}^{n} w_i \mathcal{N}(v|\mu_i, \Sigma_i) \quad (4)$$

Here n denotes the the number of mixture components computed on the complete training set and $|c|$ denotes the set of characters.

V. Experiments

The following subsection outlines the dataset and the experiments. A brief discussion is also provided in order to emphasize the important aspects of the experiments.

A. Dataset

For this experiment, we used the same dataset as descibed in [9], [10] but with a different composition. We did not work on a class-wise manner but rather document-wise. The complete collection was separated in training and test set, containing 62 and 20 documents, respectively.

Among the training set, there are 17476 diacritic-character pairings. There are also 8039 characters without any diacritics around them, which are of no concern for our purposes. Similarly, the test set consists of 6058 such pairings and 2522 characters that should not be associated with any diacritics.

The complete dataset was manually annotated so that the ground truth of all diacritic-character pairs is available. The pairs of the training set are also arranged in a character-based representation. Since there are 20 different characters, the overall set consists of 20 groups of pairs. Note that the groups are not significantly balanced in number of pairs.

B. Experimental setup

The first experiment was executed for supplying the baseline model. This was carried out by computing the closest distance of the diacritic and the character as the association criteria.

In the following we evaluated our approach, as described in Section IV. For these experiments we considered only the six characters that are closest to the diacritic in question for further investigation (see Fig. 6).

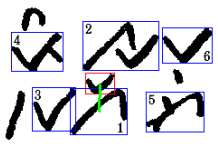


Fig. 6. The six characters closest to the diacritic are evaluated by our statistical approach.

For the evaluation, we needed to estimate the mixture model parameters in terms of the mean and covariance. Therefore, we designed four different experiment runs concerning the determination of these parameters.

For the first experiment, the mixture model parameters of the training set were obtained by estimating the densities regardless of the character information (see Fig. 5a). This distribution mainly has three separate modes, one for each diacritic position. Based on that, we proceed the estimation of parameters for three density components. Later, we also evaluated more components i.e. 5, 10 and 20 in order to have a comparable output. Second, we used the EM-algorithm to optimize the likelihood of the character independent model. All density parameters accounted in this stage are called global parameters.

The next experiment run uses the character specific model parameters, as described in Section IV-B, in order to capture more detailed information about the diacritic-character relation. We refer to this model as the local one.

The idea of the last experiment is simple. When a specific character has only a small number of training pairs, then the clustering process may not have enough information to succeed. This situation eventually could distort the estimation of parameters. To cope with this problem, a replacement of the local with the global parameters could lower this risk because the global parameters were determined from a number of data which were definitely reliable. To accomplish this, first, both the global and local parameters were calculated. Then we sorted each character and its local model parameters according to the number of samples in an ascending order. The mechanism of the replacement has been administered in such a way that it was started by a single replacement of the parameters of the lowest number of samples. The next experiment would be two replacements of the parameters of the first and the second lowest number of samples and so forth. At the end, we replaced all 20 local models with the global one.

C. Results and discussion

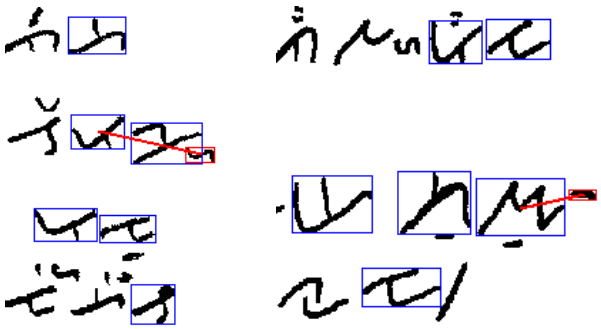
As we defined the nearest distance for assigning a diacritic to a character, we achieve an association accuracy 90.5% for the baseline model. This means that among 6058 diacritics being probed, 5481 of them were correctly associated to the characters.

The first task of the experiment was the usage of global parameters. We evaluated experiments with mixture models using 3, 5, 10, and 20 densities using k-Means clustering and later k-Means with EM. Although the number of densities is increased, the association rate of the approach does not significantly increase. As shown in Table I, the best results of 92.1% could be achieved using 20 densities and optimizing them with EM. The best trade-off between model complexity and association rate could be achieved with only five densities and EM-optimization.

While for local models for each character, the best association rate is 91.7% with three clusters (see the first row of Table II), we can see that the accuracy of the local parameters rate is not significantly different from the association rate of the global parameter experiments. The reason is that for some characters the number of diacritic associations was not enough

to form three or more clusters. Consequently, the association criteria of the diacritics to a character was deficient due to the bias of the parameters.

Finally the output of the last experiment is shown in Table II. The second row with 19 character specific models indicates the experiment with only one character model being replaced by the global one. The next shows the experiment with two characters parameters being replaced by the global parameters and keeping 18 local parameters and so forth. If we compare the association rates in those tables and the baseline indicator, we can infer that the majority of the association rates are significantly improved from the baseline indicator. The rates fluctuate around $91.0 - 92.2\%$ with the best result being achieved with five densities and 12 or $14-18$ character specific models.



(a) (b)

Fig. 7. Incorrect association of a diacritic to the character: (a) due to a lack of data, (b) due to a misclassified diacritic.

During our experiments we also analyzed some cases of incorrect associations. Fig. 7 shows two examples of them from the last experiment with five densities. In Fig. 7a, the source of incorrect association is a lack of data. The character *wa* ( ) has the diacritics with the third lowest rank. In this step it became a character candidate. However, when the Gaussian mixture model was used for exploring the density elements, there were not enough data samples. Hence, the probability density function was biased and then assigned to the wrong character (*ba* or).

Whereas in the second example, the association step is systematically correct. However, the diacritic of such a shape as seen in Fig. 7b may not occur on the right side of any character. Due to the fact that the diacritic classification on the test set is fully automatic, the classifier incorrectly assigned it as being a diacritic of the character.

TABLE I
EXPERIMENT OF MIXTURE MODEL WITH THE GLOBAL PARAMETERS

Number of density	Clustering method	Correct association (%)
3	K-Means	91.5
5	K-Means	91.5
10	K-Means	91.9
20	K-Means	91.8
3	K-Means with EM	91.6
5	K-Means with EM	92.0
10	K-Means with EM	91.9
20	K-Means with EM	92.1

TABLE II
EXPERIMENT OF MIXTURE MODEL WITH REPLACEMENTS THE LOCAL TO GLOBAL PARAMETERS

Number of character specific models	Association rate of 3 densities	Association rate of 5 densities
20	91.7	91.5
19	91.9	92.0
18	91.8	**92.2**
17	91.8	**92.2**
16	91.7	**92.2**
15	91.5	**92.2**
14	91.3	**92.2**
13	91.0	92.1
12	90.3	**92.2**
11	90.0	92.1
10	89.7	92.1
9	88.9	92.1
8	88.1	92.1
7	87.4	92.0
6	85.6	92.0
5	84.3	91.9
4	83.4	91.9
3	82.0	91.8
2	79.6	91.7
1	78.0	91.7
Global model	91.6	92.0

VI. CONCLUSION

In this paper, we proposed a technique for associating a diacritic to the correct character selected from a set of candidate characters in Lampung script. The association rule is built based on a statistical approach using a Gaussian mixture model. To achieve comparable results, we applied four different strategies in the determination of the mixture model parameters. From those strategies, the best association rate is 92.2% which is very promising for non cursive handwritten script like Lampung text. This achievement is also reflecting the effectiveness of our approach.

ACKNOWLEDGMENT

This work has been supported by the Directorate General of Higher Education, The Ministry of Education and Culture, Republic of Indonesia.

REFERENCES

[1] M. Lutf, X. You, and H. Li, "Offline Arabic handwriting identification using language diacritics," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, August 2010, pp. 1912 –1915.

[2] K. C. Leung and C. H. Leung, "Recognition of handwritten Chinese characters by critical region analysis," *Pattern Recognition*, vol. 43, pp. 949–961, March 2010.

[3] U. Bhattacharya and B. B. Chaudhuri, "Databases for research on recognition of handwritten characters of Indian scripts," in *International Conference on Document Analysis and Recognition*, vol. 2, 2005, pp. 789 – 793.

[4] S. Vajda and G. A. Fink, "Exploring pattern selection strategies for fast neural network training," in *International Conference on Pattern Recognition*, 2010, pp. 2913 –2916.

[5] S. Mozaffari and H. Soltanizadeh, "ICDAR2009 handwritten Farsi/Arabic character recognition competition," in *International Conference on Document Analysis and Recognition*, 2009, pp. 1413–1417.

[6] D. C. Tran, P. Franco, and J. Ogier, "Accented handwritten character recognition using SVM - application to French," in *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, November 2010, pp. 65 –71.

[7] D. K. Nguyen and T. D. Bui, "Recognizing Vietnamese online handwritten separated characters," in *Advanced Language Processing and Web Information Technology, International Conference on*, vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, 2008, pp. 279–284.

[8] M. Pechwitz, S. S. Maddouri, V. Maergner, N. Ellouze, and H. Amiri, "IFN/ENIT - database of handwritten Arabic words," in *In Proc. of CIFED 2002*, 2002, pp. 129–136.

[9] S. Vajda, A. Junaidi, and G. A. Fink, "A semi-supervised ensemble learning approach for character labeling with minimal human effort," in *International Conference on Document Analysis and Recognition*, IAPR. Beijing, China: IEEE Computer Society, September 2011, pp. 259–263.

[10] A. Junaidi, S. Vajda, and G. A. Fink, "Lampung - a new handwritten character benchmark: Database, labeling and recognition," in *International Workshop on Multilingual OCR (MOCR)*. Beijing, China: ACM, 2011, pp. 105–112.

[11] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1, 1967, pp. 281–296.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–22, 1977.