Signature and Lexicon Pruning Techniques

Srinivas Palla, Hansheng Lei, Venu Govindaraju
Centre for Unified Biometrics and Sensors
University at Buffalo
{spalla2, hlei, govind}@cedar.buffalo.edu

Abstract

Handwritten word recognition and Signature identification are important areas in machine vision that require extensive exploration for improved results. The performance of such systems tend to degrade when the number of choices to be dealt with increase. In case of a lexicon-driven handwritten word recognizer, the performance degrades when the lexicon size increases [8]. Similarly, the matching process is tedious when the number of reference templates increase in case of online signature identification. For better performance, both in terms of recognition rates and response time, interactive models are suggested, which involve feedback process to further enhance the systems. Interactivity is attained by choice pruning, which filter out useless entries, thus providing the system with a smaller set for further detailed investigation. The paper mainly identifies the necessity for choice pruning and deals with two specific cases – signature pruning and lexicon pruning.

1. Interactivity and choice pruning

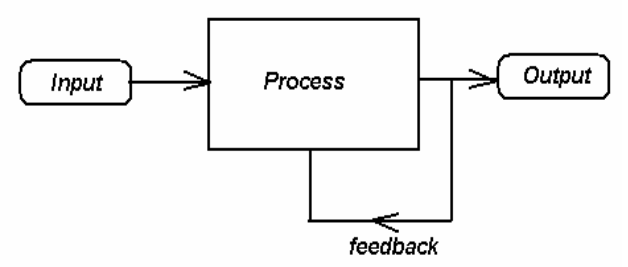

Figure 1. Generic interactive model

Interactive Process is the one, which uses a feedback to enhance the process performance, and thus generates a better output. The whole process is based on choice excitation wherein certain choices are excited and certain others are hindered based on the feedback given. The next iteration considers only the excited choices and filters out the unexcited ones. The number of levels of interactivity can vary depending on the application. The model has been described in [2] as interactive activation model. Choice pruning is innately built into the interactive activation model. Choice pruning should allow non-dismissal of correct choices. The filtering process has several advantages. In case of handwritten word recognition, the pruning process results in a smaller lexicon and thus a detailed one-to-one matching of the image with word models is quicker and more efficient. In case of signature recognition too, the number of one-to-one matches reduces. Another advantage of pruning signatures occurs in the case wherein a detailed one-to-one finger print matching is done only on the fingerprint templates corresponding to pruned signature templates. Such a system increases the efficiency of fingerprint matching based on signature pruning. The paper deals with these issues in the following manner. Section 2 deals with a novel signature pruning method based on regression of quantile samples. Section 3 deals with lexicon pruning based on regular expression matching. Section 4 concludes the paper with a glimpse on future work.

2. Signature Pruning

Online Signature is a time series data and therefore regression analysis [6] can be done to study the properties of online data. The similarity between two signatures can be related to a distance between them, where the distance is the squared error distance, which can be found from simple linear regression. Regression methods are used to quantify the relationship between two variables. Mean regression analysis does not take the whole data distribution into consideration and therefore there is information loss when such methods are used. To

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

properly quantify the whole relationship between two data distributions, it has been shown in literature that quantile regression can yield better results [4,5].

2.1. Quantile samples

Quantile samples are values that split the data population into different proportions. To summarize a vector of numeric values, the general approach is to represent the vector using the mean value. However, the mean does not always provide the best representation of the whole vector. If the data values are skewed such that there are very few high values but a large number of small values, the mean is sensitive to high values. Such a data can be summarized by using not just one mean but by using different quantiles.

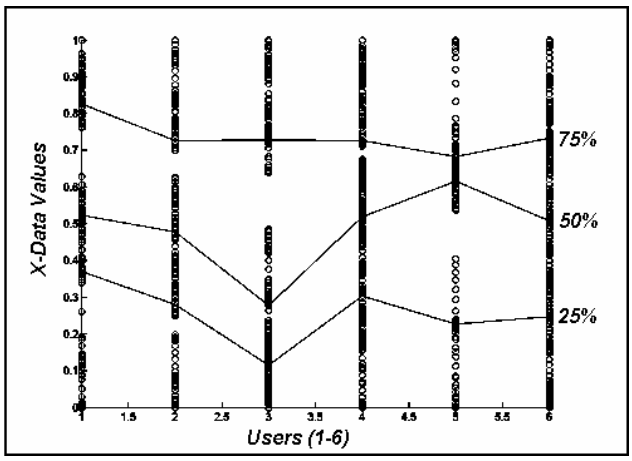

Figure 2. 25% - 50% - 75% Quantiles

A p% quantile is defined as the value that splits the data into proportions of p/100 and (100-p)/100. For example, a 25% sample quantile splits the data into ¼ and ¾ while a 50% quantile splits the data population into two halves. The latter is equal to the more popular median of the distribution. By taking into account all the quantiles instead of median, the innate skew ness in the data population can be accounted for. The figure 2 is a plot of the X-data values for various Users' signatures. The figure shows 25%, 50% and 75% quantiles of each X-data population.

2.2. Simple linear regression of quantiles

Quantile regression as described in [4] is an extension of classical least squares estimation of conditional mean models to the estimation of an ensemble of models for several conditional quantile functions. Data distribution similarity can be assessed by taking the quantile values from different data populations and then plotting the quantiles of one distribution against the other. Once the points are formed, a simple linear regression is done so as to get the best fitting line through these points. If the two data populations belong to the same distribution, then their quantile-quantile plot is a straight line. So the squared error distance gives how similar the two distributions are. The two distributions can also be time series data. Based on this idea, signature pruning is done as follows. For every user, a reference signature is taken. The reference signature is selected from the training samples (5 samples are taken during user registration). The reference signature is the one that has the least distance from all the other signatures. The "distance" here is the least square error distance found after simple linear regression through quantile samples. If only the x-y coordinates are taken to represent a user signature, the signature is first made scale and translation invariant. This is done as follows:

$$X_i = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \quad Y_i = \frac{Y_i - \min(Y_i)}{\max(Y_i) - \min(Y_i)}$$

Once the signatures are normalized, the distance between them can be found from the X-data and Y-data quantiles (fig 3). Let the quantile distance (qdist) be defined as the sum of the squared residuals after simple linear regression through the quantile Vs quantile plot, obtained by plotting the quantiles of one population against the other. Given two signatures $S_1 <X_1, Y_1>$ and $S_2 <X_2, Y_2>$, the distance between them is given by:

$$d_x = qdist(X_1, X_2) \qquad d_y = qdist(Y_1, Y_2)$$
$$d = d_x \times d_y$$

2.3. Pruning process

For every registered user, the reference signature is stored. When a test signature is given, the distance of the test signature with every reference signature is taken. The reference signatures are sorted according to their distance from the test signature and the top 'n' signatures are filtered for further consideration while other signatures are discarded. The choice of n is made based on the experiments with a set of test signatures pertaining to various users. The test dataset consisted of 40 users with 19 signatures per user for a total of 760 signatures. One reference signature per user is already set aside and does not constitute the test dataset.

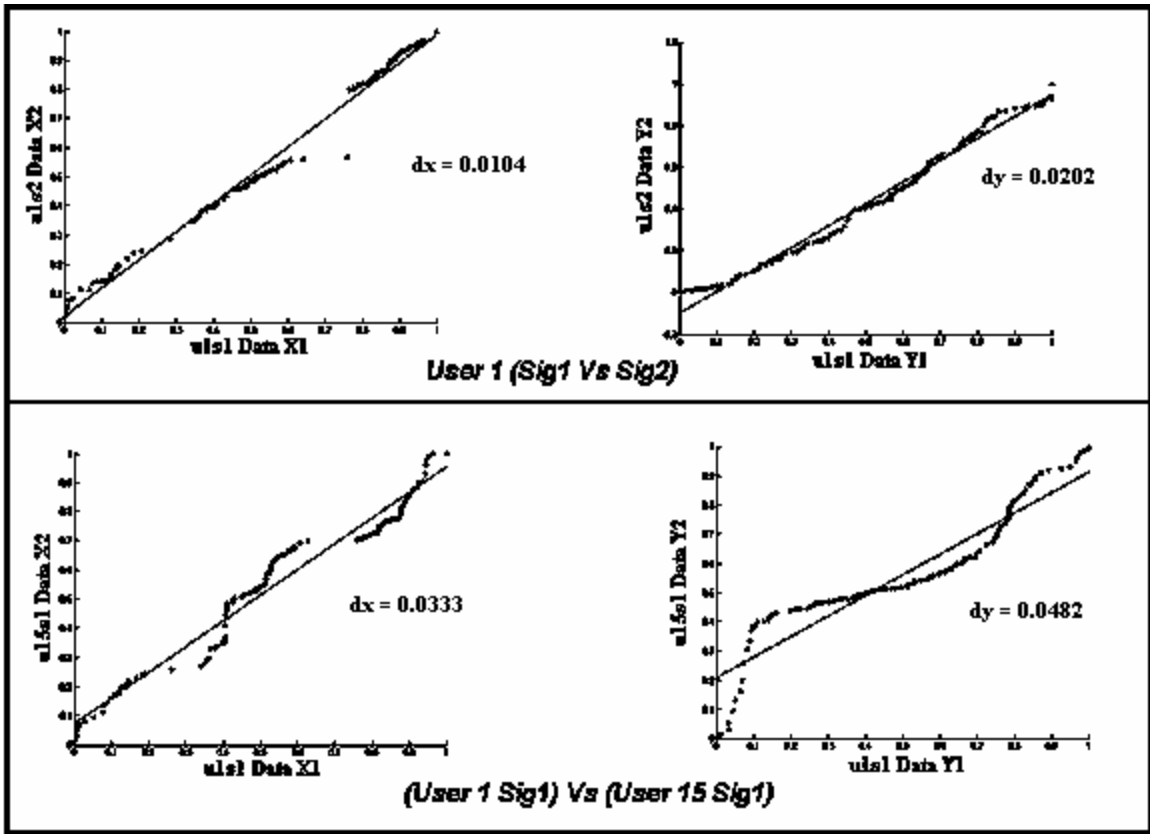

Figure 3. Quantile – Quantile plots

The pruning results are shown in table 1. The correct signature template is always within the pruned results when n=5. Also 86% of the time, the correct signature is the one with least quantile distance ($d = d_x \times d_y$). So by taking top 5 signatures, it is ensured that there are no false dismissals. No forgeries are considered here, since the main idea is to prune correct templates and pass it to the actual recognizer. So even if a forgery is given, it is for the later stage to deal with it. Pruning only ensures that false dismissals don't occur.

Table 1. Results of pruning on signature dataset

Total Signatures	Top 1	Top 2	Top 3	Top 4	Top 5
760	654	58	30	16	2

3. Lexicon Pruning

The word image can be considered as having three parts – main body, ascenders and descenders. Accordingly, we have four lines to segment the word image. These are the ascender line, the half line, the base line and the descender line. Once the image features are extracted, we can find whether the image contains ascenders or descenders. Based on this information, the lexicon can be pruned so that a one to one match is done only with the pruned lexicon entries. This will minimize the comparison time and at the same time increases accuracy.

3.1. Reference lines

Given a word image, the reference lines of the image are found. The reference lines are found according to the algorithm described in [3]. The basic idea is to construct horizontal runs from the image. Once the horizontal runs are found, the center of mass of the image is found. A regression line is drawn through all the horizontal runs below the center of mass, which have no neighbors below. This is an approximation for the base line. Now consider only those horizontal runs (with no neighbors below) within a small distance from the approximated baseline and do a simple linear regression to get the final baseline. The horizontal runs which are minima and maxima help determine the descender line and ascender line. The half line corresponds to maxima near to base line.

Figure 4. Original image

Figure 5. Reference lines

Once the reference lines are detected, the presence or absence of ascenders and descenders can be found. For this purpose the image is converted into a block adjacency graph from the horizontal runs. A graph with nodes and edges corresponding to the image is generated from the BAG. The graph representation along with the reference lines aid in the generation of a regular expression corresponding to the image.

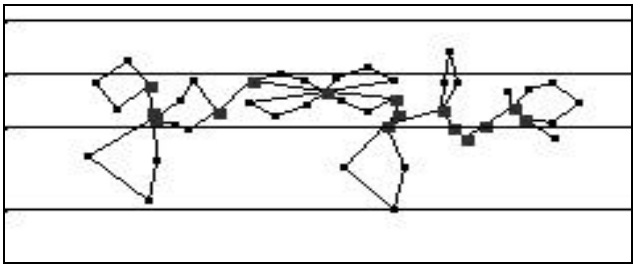

Figure 6. Graph Representation of Image

3.2. Regular expression match pruning

Given a word image, a regular expression pertaining to that word can be formed. Ascender-Descender detection is trivial, given a graph representation and reference lines. The regular expression is formed using three basic symbols <a, d, n>. Here 'a' stands for ascender, 'd' stands for descender and 'n' stands for any character. Now a regular expression is formed based on these symbols for a particular image and lexicon entries, which match the regular expression, are only considered for further matching. A Legal regular expression [7] contains symbols with their occurrence indicator. A "*" indicates that the symbol appears zero or more times. A "+" indicates that a symbol appears one or more times. For example, a regular expression "nd(n)*$(a)^+$(n)*a" indicates that the word starts with a normal character, followed by a descender, followed by zero or more occurrences of a normal character, followed by one or more occurrences of an ascender, followed by zero or more occurrences of a normal character, followed by an ascender. A word matching the regular expression is "symbol". As an other example, a regular expression "a(n)*$(d)^+$n*an" corresponds to "buffalo" but not "Amherst" since the regular expression indicates that there is at least one character which is a descender.

3.3. Pruning process

Given a word image, the reference lines are detected and the graph is generated. From this information, a regular expression corresponding to the word is generated. Every entry of the lexicon is checked against the regular expression based on a fast regular expression - matching algorithm. Only those words matching the regular expression are considered for further investigation. The pruning process is customized for a particular word recognizer [1] that uses the reference lines and graph representation for image feature extraction. So the pruning process is just a part of the image feature extraction process with only an additional regular expression matching [7] involved.

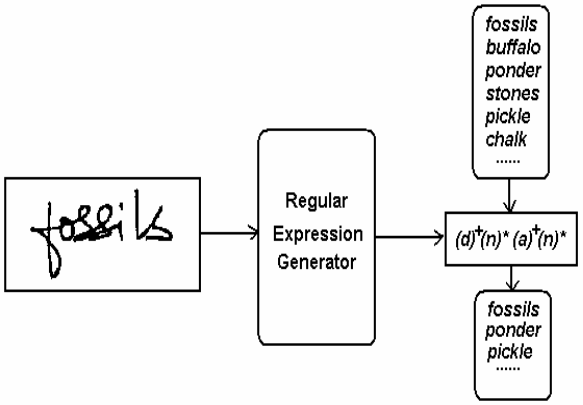

Figure 7. Regular expression based pruning

4. Conclusions

The signature pruning and lexicon pruning processes, when integrated with the actual recognizer make the whole system more robust. Any recognizer involving a one to one match mechanism will benefit from such a feedback.

The feedback can also be in a different form. An alternative approach involves binning where in the feedback is the bin identifier which contains templates to be matched. Signature binning can be quite useful in biometrics when a large fingerprint database is used. Signature identifies the bin and a fingerprint matching is done only with corresponding templates belonging to that bin. Study of various other kinds of feedback methods and faster pruning and binning techniques are promising areas where further work can be done.

References

[1] H. Xue, V Govindaraju. "Stochastic Model Combining Discrete Symbols and Continuous Attributes and its Application to Handwritten Recognition". *International Journal of Document Analysis and Recognition* [2003].

[2] Marcus Taft. "Reading and the Mental Lexicon", *Essays in Cognitive Psychology*.

[3] Slavik, V Govindaraju. "An Overview of Run-Length Encoding of Handwritten Word Images". *Buffalo tech Reports* [2000].

[4] Koenker Roger, Kevin F. Hallock. "Quantile Regression". *Journal of Economic Perspectives*, 15(4), Fall, 143-56.

[5] Keming Yu, Zudi Lu, Julian Stander. "Quantile Regression: Applications and Current Research Areas". *Journal of the Royal Statistical Society: Series D (The Statistician)*, Volume 52 Issue 3 [2003].

[6] F. Mosteller and J.W. Tukey. "Data Analysis and Regression – A Second Course in Statistics". *Addison-Wesley*, 1977.

[7] John E. Hopcroft, Rajeev Motwani, Jeffrey D. Ullman. "Introduction to Automata Theory, Languages and Computation". *Addison-Wesley* 2000.

[8] H. Xue, V Govindaraju. "On the dependence of handwritten word recognizers on lexicons". *IEEE transactions on pattern analysis and machine intelligence*, December 2002 (Vol 24 No 12).