On-Line Signature Verification With Two-Stage Statistical Models

Liang Wan
The Chinese University of Hong Kong
lwan@cse.cuhk.edu.hk

Bin Wan
Huazhong University of Science and Technique
wanbin@gmail.com

Zhou-Chen Lin
Microsoft Research, Asia
zclin@microsoft.com

Abstract

Signature verification is a challenging task, because only a small set of genuine samples can be acquired and usually no forgeries are available in real application. In this paper, we propose a new two-stage statistical system for automatic on-line signature verification. Our system is composed of a simplified GMM model for global signature features, and a discrete HMM model for local signature features. To be practical, we introduce specific simplification strategies for model building and training. Our system requires only 5 genuine samples for new users and relies on only 3 global parameters for quick and efficient system tuning. Experiments are conducted to verify the effectiveness of our system.

1. Introduction

Signature verification as a technique to authenticate the signer's identity has a long history in human society. Similar to other biometrical information, such as speech and fingerprint, signature conveys distinguishable characteristics that indicate personal identity. Signature verification aims at using such properties for making reliable authentication. However, automatic signature verification is a challenging task due to practical constraints. For example, we cannot collect large amount of signatures for training purpose, while handwriting or speech recognition does. And negative samples cannot be captured when the system is used. On the other hand, there are increasing security demands on identity verification in the field of finance, security, etc. Therefore, automatic signature verification deserves deep investigation.

The existing methods of signature verification can be classified into two categories: off-line and on-line. Off-line methods acquire data by scanning signatures and process them as static images. On-line methods capture signature tracks in time-variable sequences, such as positions, pressure, and pen tilt. On-line methods usually achieve higher accuracy than off-line ones do because they can make use of dynamic information (speed, pressure, etc.) that is missing in static images.

For on-line signature verification [4], dynamic time warping (DTW) [10][17][6][8] is one widely-used method to find the similarity between the input signature patterns and the stored templates. The signature pattern is usually represented by a sequence of feature vectors defined on every sample point of the signature. Huang [5] built a structure graph for each signature considering stroke correlation, and did segment-to-segment matching by DTW. Hidden Markov Model (HMM) is another common technique for signature verification in recent years, because it has been successful in modeling time-variable sequences for speech and on-line handwriting recognition. Yang [18] evaluated different HMM models over the same signature dataset. In [3][7][9][11][14] various segmental features are proposed for HMM verification. Richiardi [17] applied Gaussian Mixture Model (GMM) for signature distribution estimation. These models (DTW, HMM, and GMM) focus on local properties of signatures such as local moving direction and shape curvature. Global features like writing time or signature length may be incorporated to speed up verification procedure and improve accuracy.

In the practical applications, users may feel uncomfortable if required to provide lots of samples. Only small set of signature samples can be acquired, say $5 \sim 6$ genuine signatures [14][19]. This demands the ideal signature verification system to be simple because complex system usually needs large training sample set. On the other hand, many existing systems require false samples, i.e., forgery signatures, for system training. But it is actually hard to collect forgery signatures for every user in practice. So the practical system would be more applicable if it uses forgery sam-

ples as few as possible.

Due to the above-mentioned two considerations, we propose a two-stage statistical system for on-line signature verification. Our system is composed of a simplified GMM model built on global signature properties and a left-to-right HMM model based on segmental features. Note that the general GMM model and HMM model are complex for this specific application, so we introduce specific strategies to do model simplification and initialization. Our system depends on three global parameters to control its performance. We estimate those parameters globally for all users such that forgeries are only needed for system tuning. When put into real use, for any new user, our system needs only 5 genuine signatures and no forgery samples. Our system achieves good performance considering that it uses few training samples.

The rest of the paper is organized as follows. Section 2 gives an overview of the proposed system and its main components. Section 3 describes how to build a simplified GMM model to estimate the distribution of global features. Section 4 discusses the HMM model estimation based on segmental features. In Section 5, the experiments are reported. Finally, Section 6 concludes the paper and discusses the future work.

2. System overview

Figure 1 illustrates the structure of our system. The right part (indicated by dashed lines) shows the training procedure. For each signer, two models are processed separately, corresponding to global and local signature information. In global modeling, we use a Gaussian mixture to estimate the distribution of global features, such as time duration and average speed. In local modeling, we build an HMM model based on both piecewise information and structural relation between strokes.

The left part (indicated by solid lines) shows the verification procedure. The input signature is first fed into the GMM classifier. If the confidence is below a threshold, the signature is rejected as a forgery. For signatures that pass the first test, we extract segmental feature sequences and feed them into the HMM classifier. The signature is accepted as genuine when it also passes the HMM verification test.

We list the highlights of our system in the following:

- Given the well-established system, it only uses few genuine signatures as training data for a new user. No forgeries are needed in the training stage.

- Discriminative features are proposed at global and local levels, respectively.

- Our system adopts a two-stage statistical structure, where the global level features can rule out obvious forgeries quickly.
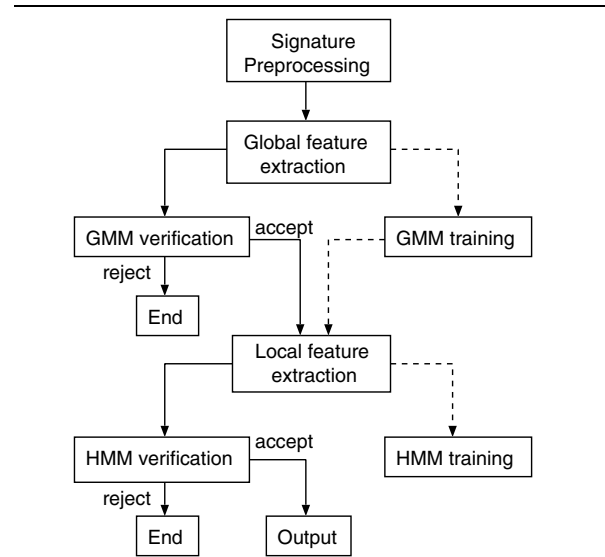

Figure 1. Block diagram of our signature verification system

- The system can be easily tuned since there are only three global parameters involved.

3. First stage: GMM for global verification

It is reasonable to assume that signatures of the same writer are very similar. In other words, global information of signatures, such as width and height, are consistent for the same signer. Global verification aims at ruling out obvious forgeries quickly using the consistent global information. In this section, we first introduce global features and then discuss how to build a simplified Gaussian Mixture Model (GMM) over reference samples.

3.1. Global feature extraction

We extract two kinds of global features, spatial and dynamic features. The spatial features include: (i) width and height; (ii) total length of signature strokes, (iii) stroke count and number of self-intersection points; (iv) segment count and (v) total curvature. Spatial features are usually enough for detecting random forgery, i.e., signatures from other signers. However, they may fail for skilled forgeries, which result from purposely imitation and may have a similar shape to the signer's genuine signatures.

Dynamic features, on the other hand, have good discriminative power beyond shape features, because they are hard to be imitated by observing the signature shape only. We propose a set of dynamic features: (i) the average speed and

the maximum speed; (ii) the average pressure and the maximum pressure difference between two sample points, (iii) total duration time, and (iv) ratio of pen-down time to total writing time.

This group of spatial and dynamic features has low correlation, which makes model simplification possible as will be discussed later.

3.2. GMM modeling

GMM model is a weighted combination of multiple Gaussians. A distinct property of GMM model is its power to approximate any probability density. Denote an M-component GMM model by the notation $\Theta = \{(\alpha_l, \mu_l, \Sigma_l) : l = 1, \ldots, M\}$. It has the general form as equation 1,

$$p(x|\Theta) = \sum_{l=1}^{M} \alpha_l \cdot p(x|\mu_l, \Sigma_l) \qquad (1)$$

where x is an n-dimensional feature vector, α_l is the coefficient with $\alpha_l = P(l|\Theta)$, $\Sigma \alpha_l = 1$, and μ_l and Σ_l are the mean vector and the covariance matrix of the l-th Gaussian component, respectively. Each component's distribution is an n-variable Gaussian function.

The GMM complexity is controlled by component count M and covariance matrix Σ. Larger M and full covariance matrix Σ define a complex GMM model. Obviously one complex GMM can approximate a sample distribution more accurately. However, the complex model has many unknown parameters to estimate, which needs many training samples. We have to simplify the model to meet the number constraint of signature samples.

The first simplification is to represent the distribution variance by a diagonal matrix. The diagonal covariance matrix compared to a full covariance matrix may lose some information. When feature vectors are not considerably correlated, the lost information can be neglected. As mentioned previously, the global features we selected have low correlation to each other. Therefore we may use diagonal matrix safely. Another advantage of using a diagonal matrix is great reduction in the computational load.

The further simplification is to choose fewer Gaussian components. Note that only 5 genuine samples are collected as the training set. Such a set size is far from enough for a complex GMM model. Therefore, we choose a 2-component GMM in practice.

In order to avoid GMM learning from being stuck into local extrema, K-means clustering is usually applied for "good" initialization. Because our simplified GMM model contains just 2 components, we propose a simple method to initialize the GMM. There are 10 different combinations for 5 samples being separated into 2 classes. We may choose

the best combination with small intra-class distance and large inter-class distance by exhaustive search. Note that the feature vector does not have the same unit in each dimension. We should normalize each dimension before computing the distance matric between two feature vectors.

Then EM algorithm [2] is then applied for GMM model learning. Given an incomplete data set $X = \{x_1, \ldots, x_K\}$, where K is the number of samples, and x_k is the global feature vector, the GMM log-likelihood is computed as,

$$log(p(X|\Theta)) = \sum_{k} \log(p(x_k|\Theta)) \qquad (2)$$

The EM algorithm updates the initial model and guarantees that the log-likelihood of the new model increases. The new model is in turn used for the next iteration. This iterative process continues until the log-likelihood of the new model converges.

4. Second stage: HMM for local verification

Compensating for global properties, local information offers signature details, spatial or dynamic, discriminative for different signers [17]. Those local features are incorporated with DTW or HMM techniques for signature verification. DTW-based systems may resample the signature into an equal-distant point sequence before string matching. But this resampling process may result in loss of important local details. Our system uses non-resampled signature data in experiments.

4.1. Local feature extraction

Local features in the system are extracted at segment level. We first adopt Sklansky's algorithm [15] to segment the input signature at points of high curvature. From each segment a set of local features are computed including (i) count of sample points, (ii) segment length and direction, (iii) average pressure and maximum pressure difference, (iv) time duration, and (v) relative position/direction with respect to other segments.

Denote the segmental features extracted from the k-th signature as Y^k.

$$\begin{aligned} Y^k &= (y_1^k, y_2^k, \ldots, y_{Tk}^k), \\ \text{and} \quad y_t^k &= (f_{t1}^k, f_{t2}^k, \ldots, f_{tD}^k), \end{aligned}$$

where T_k is the number of segments, f_{td}^k is the d-th feature for the t-th segment's feature vector y_t^k and D is the dimension of segmental feature vector. After local features, i.e. $\{Y^k, k = 1, \ldots, K\}$, from all training samples are obtained, we start to train a discrete left-to-right HMM model.

4.2. HMM modeling

A discrete HMM λ [13] has five elements: the state number N, the number M of distinct observation symbols per state, the states transition probability distribution $A = \{a_{ij}\}$, the observation symbol probability distribution $B = b_i(k)$ in state i, and the initial state distribution $\pi = \{\pi_i\}$.

Several factors have impact on the complexity of the HMM model, such as the underlying model structure, distribution form (continuous or discrete), and the number N or M. Given a discrete HMM, the underlying model structure determines the number of unknown parameters. The more complex the structure is, the more parameters there are, and the more training samples are required. Considering the small size of our training set, we employ a first-order left-to-right no-skip DHMM model with $N = 6$ and $M = 16$.

Before HMM learning [13], we should convert feature sequence Y^k to an HMM's observation sequence O^k. Such conversion can be realized by vector quantizer. We adopt a modified K-means clustering algorithm [12] to group segmental feature vectors into M (the number of HMM observation symbols) clusters, and take the cluster centers as the codebook. Accordingly the sequence Y^k can be encoded into a series of code index or symbol:

$$O^k = (o_1^k, \ldots, o_{Tk}^k)$$

where o_t^k is the class label for the t-th segment of the k-th signature. In turn, O^k is taken as the observation sequence for HMM learning.

Similar to GMM, HMM model may reach rapid and proper convergence in iterative learning process if starting from a good initial estimate of the parameters. For our DHMM model, $A = \{a_{ij}\}$ can be selected randomly, but $\pi = \{\pi_i\}$ must be fixed as $\pi_1 = 1$, $\pi_i = 0$, $i \neq 1$. As for $B = \{b_i(k)\}$, we employ a segmental K-means training procedure [13] to optimally search the initial values. The EM algorithm is then applied to train the HMM model.

5. Experiment results and discussions

Our system is built on the Tablet PC, and also works for other electronic writing devices, such as PDA. The signature sample has two kinds of raw data, position (x, y) and pressure p. We apply a Gaussian filter to smooth $x/y/p$ respectively. This technique eliminates some unexpected noise from original signature data. Special attention should be paid to rotation. In practice, the user may sign his/her signature in an arbitrary orientation. Therefore, the system detects writing direction, the average estimation from two methods, principal component analysis [16] and linear regression, and rotates the signature to horizontal. Signature

samples undergo the above preprocessing steps before training or verification.

A test signature x is identified as genuine if it passes both global and local verification tests. Specifically, we first evaluate its GMM likelihood as $p(x|\Theta)$. It passes the global verification and undergoes local verification if

$$p(x|\Theta) > c_g \min(p(x_k|\Theta)) \tag{3}$$

where x_k is k-th training signature and c_g is a global coefficient.

Denote the observation sequence corresponding to the test signature x by O, and the joint log-likelihood of sequence O against the HMM by $P(O|\lambda)$. The test sample passes HMM test when it satisfies

$$P(O|\lambda) > c_h \min(p(O_k|\lambda)), \tag{4}$$

where c_h is another global coefficient and O_k is the observation sequence of the k-th sample signature.

The HMM probability $P(O|\lambda)$ already takes the sequence length into account. However, we found that observation sequences of short length may have large likelihoods, though the corresponding signatures may be forgery. It is partly because short sequences enjoy smaller penalty than long sequences. Therefore, we impose an additional constraint on sequence length $L(O)$:

$$L(O) > c_l \min(L(O_k)) \tag{5}$$

where $L(O_k)$ is the length of the observation sequence O_k and c_l is a global coefficient. If the condition 5 is not satisfied, the test signature x is rejected as a forgery.

The three coefficients (c_g, c_h, c_l) affect the system performance. We can find the optimal values for each signer individually or for all signers globally. In other words, they are different for various signers in the former case, and the same in the latter case. In our experiment, we find the global optimal (c_g, c_h, c_l) by exhaustive search for all registered signers. Two sets of coefficients (c_g, c_h, c_l) are determined for verification with or without pressure, respectively.

The signature data in test contains 25 volunteers. Each of them provided 15 genuine signatures, and they were asked to imitate signatures of another three subjects, each in 5 times as skilled forgeries. We also randomly selected 15 signatures from others for each signer as random forgeries. In total, each signer has 15 genuine signatures, 15 random and 15 skilled forgeries. Although no public signature database is available for testing, our database is comparable to that used in Signature Verification Competition [1].

The signature model for one signer gets trained by 5 genuine signatures randomly selected. The rest samples are used for parameter tuning. Currently we search for the optimal parameters using the testing samples, and the verification results are actually ones after threshold tuning. With

pressure information, the accuracy is 93.3%; without pressure information, the accuracy is 89.7%. These numbers are not the highest among those reported in the literature. However, one should be reminded that our system requires much less samples than many other systems do.

For further testing generalization performance of the system, we would like to collect more signature data from other signers. At that time, the global system coefficients (c_g, c_h, c_l) are fixed, and only genuine samples are required for new signature model training.

6. Conclusions and future work

In this paper, we propose a new signature verification framework, the two-stage statistical model. It requires very few user samples, yet achieves good performance. The system consists of a simplified GMM and a discrete left-to-right HMM. The two signature models work in a sequential order. We use GMM verification to quickly rule out obvious forgery signatures by computing global features in this stage. Then we evaluate the signatures that pass the global verification against HMM model. Our system works well for small training set.

There are still many open problems. We have noticed that a user's signatures may change along the time due to the change in physical and psychological conditions. It would be attractive if we could update the verification system dynamically based on previous verification results. Such a consideration can make automatic signature verification system more applicable. Moreover, by selecting discriminative features, we expect that the performance of our system will be even better.

References

[1] Results of first international signature verification competition (svc2004). http://www.cs.ust.hk/svc2004/results.html.

[2] J. A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, University of Berkeley, 1998.

[3] J. Dolfing, E. Aarts, and J. van Oosterhout. On-line signature verification using hidden markov models. In *14th International Conference on Pattern Recognition (ICPR'98)*, pages 1755–1757, 1998.

[4] G. Gupta and A. McCabe. A review of dynamic handwritten signature verification. Technical report, James Cook University of North Queenland, Computer Science Dept., 1997.

[5] K. Huang and H. Yan. Stability and style-variation modeling for on-line signature verification. *Pattern Recognition*, 36(10):2253–2270, October 2003.

[6] A. Jain, F. Griess, and S. Connell. On-line signature verification. *Pattern Recognition*, 35(12):2963–2972, December 2002.

[7] R. S. Kashi, J. Hu, W. L. Nelson, and W. Turin. On-line handwritten signature verification using hidden markov model features. In *4th International Conference Document Analysis and Recognition (ICDAR '97)*, pages 253–257, 1997.

[8] J. Lee, H. Yoon, J. Soh, B. Chun, and Y. Chung. Using geometric extrema for segment-to-segment characteristics comparison in online signature verification. *Pattern Recognition*, 37(1):93–103, January 2004.

[9] L. L. Lee, T. Berger, and E. Aviczer. Reliable on-line human signature verification systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(6):843–847, 1996.

[10] R. Martens and L. Claesen. On-line signature verification by dynamic time-warping. In *13th International Conference on Pattern Recognition (ICPR'96)*, pages 38–42, 1996.

[11] D. Muramatsu and T. Matsumoto. An hmm online signature verifier incorporating signature trajectories. In *7th International Conference Document Analysis and Recognition (ICDAR'03)*, pages 438–442, 2003.

[12] K. K. Paliwal. Comments on "modified k-means algorithm for vector quantizer design". *IEEE Tran. on Image Processing*, 9(11):1964–1966, 2000.

[13] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

[14] M. M. Shafiei and H. R. Rabiee. A new on-line signature verification algorithm using variable length segmentation and hidden markov models. In *7th International Conference Document Analysis and Recognition (ICDAR'03)*, pages 443–446, 2003.

[15] J. Sklansky and V. Gonzalez. Fast polygonal approximation of digitized curves. *Pattern Recognition*, 12(2):327–331, 1980.

[16] T. Steinherz. Skew detection via principal components analysis. In *5th International Conference Document Analysis and Recognition (ICDAR'99)*, pages 153–156, 1999.

[17] K. Tanabe, M. Yoshihara, H. Kameya, and S. Mori. Automatic signature verification based on the dynamic feature of pressure. In *6th International Conference Document Analysis and Recognition (ICDAR'01)*, pages 1045–1049, 2001.

[18] L. Yang, B. K. Widjaja, and R. Prasad. Application of hidden markov models for signature verification. *Pattern Recognition*, 28(2):161–170, February 1995.

[19] M. F. Zou, J. J. Tong, C. P. Liu, and L. Z. L. On-line signature verification using local shape analysis. In *in 7th International Conference Document Analysis and Recognition (ICDAR'03)*, pages 314–318, 2003.

IEEE
COMPUTER
SOCIETY