A FAST PROCEDURE FOR THE COMPUTATION OF SIMILARITIES BETWEEN GAUSSIAN HMMS

Ling Chen, Hong Man

ECE Department, Stevens Institute of Technology
Hoboken, NJ 07030, USA
Email: {lchen, hman}@stevens.edu

ABSTRACT

An appropriate definition and efficient computation of similarity (or distance) measures between stochastic models are of theoretical and practical interest. In this work a similarity measure for Gaussian hidden Markov models is introduced based on the generalized probability product kernel. An efficient scheme for computing the similarity measure is presented. The out of precision problem, which is a significant implementation issue, is considered and a scaling procedure is provided. The effectiveness of the proposed method has been evaluated on texture classification and preliminary experimental results are presented.

1. INTRODUCTION

Hidden Markov model (HMM) is widely adopted in many application areas. An interesting question about HMM is: given model parameters of two HMMs, how to define an appropriate similarity measure between the two models [1].

A solution to this question is of theoretical interest and an appropriate similarity measure between two HMMs can be useful in certain application areas such as computational biology [2] and image retrieval [3]. There have been some research efforts on this subject and several techniques have been proposed in the literature [1][2][3][4]. Recently, a *generalized probability product kernel* (GPPK) between distributions, which represents the similarity between two probability distributions p and p' is proposed by Jebara et.al. [5]:

$$K_\rho(p,p') = \int_\Omega p(x)^\rho p'(x)^\rho dx, \tag{1}$$

where normally $\rho \in \{1/2, 1, 2, 3, ...\}$. The motivation behind this method is to explore the fusion of discriminative and generative estimation to exploit their complementary advantages. In [5], the *expected likelihood kernel*, i.e., $K_1(p,p')$ is used to derive the similarity measure between Gaussian mixture models and HMMs. Some of the advantages of the GPPK include it is positive definite, symmetric and capable to handle a variety of generative models (including HMMs) in closed form. As a comparison, Kullback-Leibler distance is not positive definite, is asymmetric and in many cases, it can only be approximated by an upper bound [3].

However to evaluate GPPK between two Gaussian HMMs can be prohibitively computational intensive, e.g., the computational complexity is $O(3T(NN')^{T+1})$, where T is the number of transitions and N and N' are number of states of two HMMs. In this work we propose a fast procedure for the evaluation which decreases the computational complexity to $O(3T(NN')^2)$. Meanwhile, because the computation of similarity between HMMs will exceed the precision limit of almost any machines if T is large, we formulate a scaling procedure that can prevent the computation from going beyond the precision range as well as guarantee the exact value of the similarity measure can be evaluated.

In below, section 2 introduces the similarity measure of Gaussian HMMs and presents the fast procedure for the computation of similarity measure. Section 3 details the scaling procedure. Section 4 gives the preliminary experimental results. We conclude our work in section 5.

2. SIMILARITY MEASURE OF GAUSSIAN HMMS AND FAST PROCEDURE

The core of computing the similarity measure of Gaussian HMMs is to compute the similarity measure of Gaussian distributions. Based on Eq. (1), for D dimensional Gaussian $p(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $p'(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$, the GPPK is:

$$\begin{aligned} & K_\rho(p,p') \\ =& \int_{\mathbb{R}^D} p(\mathbf{x})^\rho p'(\mathbf{x})^\rho d\mathbf{x} \\ =& (2\pi)^{(1-2\rho)D/2} |\boldsymbol{\Sigma}^\dagger|^{1/2} |\boldsymbol{\Sigma}|^{-\rho/2} |\boldsymbol{\Sigma}'|^{-\rho/2} \\ & \times \exp\left(-\frac{\rho}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{\rho}{2}\boldsymbol{\mu}'^{\mathrm{T}}\boldsymbol{\Sigma}'^{-1}\boldsymbol{\mu}' + \frac{1}{2}\boldsymbol{\mu}^{\dagger\mathrm{T}}\boldsymbol{\Sigma}^\dagger\boldsymbol{\mu}^\dagger\right), \end{aligned} \tag{2}$$

where $\boldsymbol{\Sigma}^\dagger = (\rho\boldsymbol{\Sigma}^{-1} + \rho\boldsymbol{\Sigma}'^{-1})^{-1}$ and $\boldsymbol{\mu}^\dagger = \rho\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \rho\boldsymbol{\Sigma}'^{-1}\boldsymbol{\mu}'$.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

For Gaussian HMM, given the observation sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T)$ and the model $\lambda = (A, B, \pi)$, where A is *state transition distribution*, B is *observation probability distribution*, and π is *initial state distribution*, the likelihood is:

$$P(\mathbf{O}|\lambda) = \sum_{s_0,...,s_T=1}^{N} \pi_{s_0} b(\mathbf{o}_0|s_0) \prod_{t=1}^{T} b(\mathbf{o}_t|s_t) a_{s_t|s_{t-1}}.$$

When $\rho = 1$, the similarity measure of two Gaussian HMMs is:

$$\begin{aligned} & K_\rho(\lambda, \lambda') \\ &= \int_{\mathbb{R}^D \times ... \times \mathbb{R}^D} P(\mathbf{O}|\lambda) P(\mathbf{O}|\lambda') d\mathbf{O} \\ &= \sum_{s_0,...,s_T=1}^{N} \sum_{s'_0,...,s'_T=1}^{N'} \pi_{s_0} \pi_{s'_0} \psi_{s_0,s'_0} \prod_{t=1}^{T} a_{s_t|s_{t-1}} a_{s'_t|s'_{t-1}} \psi_{s_t,s'_t}, \end{aligned} \tag{3}$$

where $\psi_{s_t,s'_t}, t = 0, 1, 2, ..., T$ is the GPPK of two Gaussian corresponding to states s_t and s'_t of two HMMs. Note that if ρ is set as $1/2$, the corresponding similarity measure of two Gaussian is actually the Bhattacharyya's measure of affinity between Gaussian distributions [7], which is of interest to be evaluated. But if ρ take values other than $\rho = 1$, it is difficult to compute the GPPK of two HMMs in closed form based on the definition of GPPK. Heuristically, we can just change the value of ρ for the computing of ψ_{s_t,s'_t} in Eq. (3).

The brute force computation of similarity measure between two Gaussian HMM, however, is prohibitively intensive. The computational complexity in the evaluation of the similarity measure under Eq. (3) is $O(3T(NN')^{T+1})$. Precisely speaking, there will be $(NN')^{T+1} - 1$ additions and $(NN')^{T+1}(3T - 1)$ multiplications. Clearly a more computational efficient procedure is needed.

We define the forward similarity measure of two Gaussian HMMs as

$$\begin{aligned} & \alpha_\tau(i, j) \\ &= K_\rho(\lambda, \lambda', s_\tau = i, s'_\tau = j) \\ &= \sum_{s_0,\cdots,s_{\tau-1}} \sum_{s'_0,\cdots,s'_{\tau-1}} \Big(\pi_{s_0} \pi_{s'_0} \psi_{s_0,s'_0} \\ & \prod_{t=1}^{\tau-1} a_{s_t|s_{t-1}} a_{s'_t|s'_{t-1}} \psi_{s_t,s'_t} a_{i|s_{\tau-1}} a_{j|s'_{\tau-1}} \psi_{i,j} \Big), \end{aligned} \tag{4}$$

that is, the similarity measure of two Gaussian HMMs when only $0 \le t \le \tau$ is considered and $s_\tau = i, s'_\tau = j$. Then $\alpha_\tau(i, j)$ can be inductively computed as following forward procedure:

1. Initialization

$$\alpha_0(i, j) = \pi_i \pi'_j \psi_{i,j}, \quad 1 \le i \le N,\ 1 \le j \le N'. \tag{5}$$

2. Induction

$$\begin{aligned} & \alpha_\tau(i, j) = \sum_m \sum_n \alpha_{\tau-1}(m, n) a_{i|m} a_{j|n} \psi_{i,j}, \\ & 1 \le \tau \le T,\ 1 \le i \le N,\ 1 \le j \le N'. \end{aligned} \tag{6}$$

3. Termination

$$K_\rho(\lambda, \lambda') = \sum_i \sum_j \alpha_T(i, j). \tag{7}$$

The computational complexity of the above fast procedure is $O(3T(NN')^2)$. To be precise, there will be $NN'(2 + 3NN'T)$ multiplications and $NN'(1 + (NN' - 1)T)$ additions. Comparing to the brute force computation, the computational complexity of the fast procedure is much lower especially when T is large.

3. SCALING PROCEDURE

In Eq. (4), the initial state distribution π, the state transition probability distribution a are less than 1. It is apparent that when τ gets big, each term of the sum in Eq. (4) goes to zero and rapidly the dynamic range of $\alpha_\tau(i, j)$ will go beyond the precision range of any machine. Then a scaling is needed to maintain the value of $\alpha_\tau(i, j)$ within the dynamic range of the machine as well as guarantee the exact value of the similarity measure can be realized.

We denote $\alpha_\tau(i, j)$ as the unscaled αs, $\hat{\alpha}_\tau(i, j)$ as the scaled αs and $\hat{\hat{\alpha}}_\tau(i, j)$ as the temporary variable for the computation of $\hat{\alpha}_\tau(i, j)$. Below is the fast procedure embedded with the scaling procedure.

1. Initialization
   Let $\hat{\hat{\alpha}}_0(i, j) = \alpha_0(i, j)$. Define the scaling coefficient c_0 as $c_0 = (\sum_{i,j} \hat{\hat{\alpha}}_0(i, j))^{-1}$. Let $\hat{\alpha}_0(i, j) = c_0 \hat{\hat{\alpha}}_0(i, j)$.

2. Induction
   Let $\hat{\hat{\alpha}}_\tau(i, j) = \sum_m \sum_n \hat{\alpha}_{\tau-1}(m, n) a_{i|m} a_{j|n} \psi_{i,j}$, and $c_\tau = (\sum_{i,j} \hat{\hat{\alpha}}_\tau(i, j))^{-1}$, then $\hat{\alpha}_\tau(i, j) = c_\tau \hat{\hat{\alpha}}_\tau(i, j)$.

3. Termination
   From the induction step, it can be found that

$$\hat{\alpha}_\tau(i, j) = c_\tau \hat{\hat{\alpha}}_\tau(i, j) = \cdots = c_\tau c_{\tau-1} ... c_0 \alpha_\tau(i, j). \tag{8}$$

   Then

$$\begin{aligned} K_\rho(\lambda, \lambda') &= \sum_i \sum_j \alpha_T(i, j) \\ &= \frac{1}{c_T c_{T-1} ... c_0} \sum_i \sum_j \hat{\alpha}_T(i, j). \end{aligned} \tag{9}$$

Fig. 1. 13 categories of texture images. From top to bottom and from left to right, they are Bark, Brick, Bubbles, Grass, Leather, Pigskin, Raffia, Sand, Straw, Water, Weave, Wood, and Wool.

Because $K_\rho(\lambda, \lambda')$ and $c_T c_{T-1} ... c_0$ may also go beyond the dynamic range of the machine, we take the logarithm:

$$\log(K_\rho(\lambda, \lambda')) = \log(\sum_i \sum_j \hat{\alpha}_T(i,j)) - \sum_{t=0}^{T} \log c_t. \tag{10}$$

From the scaling procedure, for each $1 \leq \tau \leq T$, the values of the scaled αs, $\hat{\alpha}_\tau(i,j)$, are kept within the dynamic range of the computer by multiplying a scaling coefficient c_τ. By exploit the relationship between αs and $\hat{\alpha}$s, the exact logarithm value of $K_\rho(\lambda, \lambda')$ is realized.

4. PRELIMINARY EXPERIMENTAL RESULTS

The method of similarity measure of Gaussian HMMs is tested on texture classification. 13 texture images of Brodatz texture images [6] are used for classification, see Fig. 1. All 13 texture images are monochrome with the size as 512×512. Each texture images are divided into 16 128×128 non-overlapping sub texture images.

For each sub texture image, an one dimensional Gaussian HMM is trained by the observation vector sequence generated from the sub texture image. The generation of the observation vector sequence can be summarized as using a 8×8 sized sliding window to scan a sub texture image with 75% (or 6 pixels) overlap between consecutive steps from left to right and from top to bottom. The windowed image blocks are normalized to zero mean and further transformed by an 8×8 DCT. Only the 3×3 lowest frequency coefficients in the DCT domain are used to form the 9-dimensional observation vectors. All consecutive observation vectors forms the observation vector sequence.

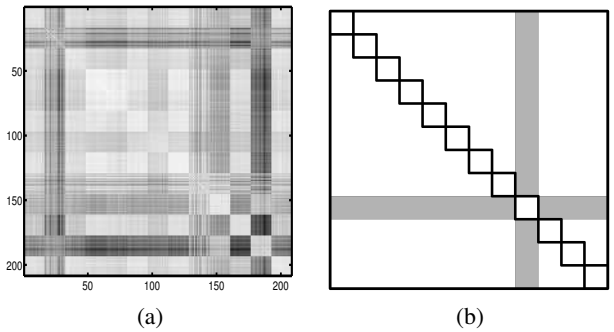

Fig. 2. (a) is similarity measure matrix of 13×16 Gaussian HMMs generated from 13×16 texture images. (b) illustrative plot of squares along the diagonal of the similarity measure matrix and the corresponding off-diagonal squares.

4.1. Experiment One

In this experiment, for each class of texture, all 16 trained HMMs are selected. Then totally there are $13 \times 16 = 208$ HMMs are selected. The similarity measures of all possible pair of HMMs among all selected 208 HMMs are computed with $T = 2^2$ and $\rho = 1/2$.[1] Denote $s^{i,j}_{m,n}$ as the log of the similarity measure between the mth HMM of class i and the nth HMM of class j. The similarity measures are arranged as following similarity measure matrix and depicted in Fig. 2(a):

$$\begin{matrix} s^{1,1}_{1,1} & s^{1,1}_{1,2} & \cdots & s^{1,1}_{1,16} & \cdots & s^{1,13}_{1,1} & s^{1,13}_{1,2} & \cdots & s^{1,13}_{1,16} \\ s^{1,1}_{2,1} & s^{1,1}_{2,2} & \cdots & s^{1,1}_{2,16} & \cdots & s^{1,13}_{2,1} & s^{1,13}_{2,2} & \cdots & s^{1,13}_{2,16} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s^{13,1}_{16,1} & s^{13,1}_{16,2} & \cdots & s^{13,1}_{16,16} & \cdots & s^{13,13}_{16,1} & s^{13,13}_{16,2} & \cdots & s^{13,13}_{16,16} \end{matrix}$$

The brightness of each pixel in Fig. 2(a) represents the value of similarity measure of the corresponding pair of Gaussian HMMs, i.e. the brighter the pixel, the greater the similarity measure.

It can be seen from Fig. 2(a), within-class similarity measures are normally higher than between-class similarity measures, e.g., the squares along the diagonal of the similarity measure matrix are generally brighter than the corresponding off-diagonal squares. To illustrate this, Fig. 2(b) provides an example of squares along the diagonal of the similarity measure matrix and the corresponding off-diagonal squares (the gray areas).

[1] We excluded the influence of the initial probability distribution π by substitute all π_is and π'_js with $1/N$s and $1/N'$s due to the limited training data (just 1 sub texture image is used in the training of HMM), the initial probability distribution is unreliable and should be excluded from the computation of similarity measure.

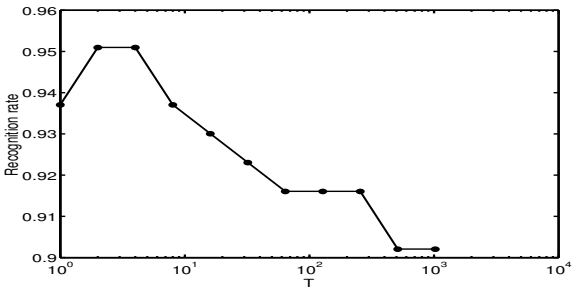

Fig. 3. Recognition rates when T is set as $2^0, 2^1, ..., 2^{10}$.

4.2. Experiment Two

Again, in this experiment, for each class of texture, 5 trained HMMs are randomly selected. For the selected 5 HMMs of each texture class, they serve as class templates. For all the corresponding unselected 11 trained HMMs of each class, they serve as the testing data. When an arbitrary testing HMM is send to the classification system, its similarity measures towards all the class templates of every texture class are computed. Then the similarity of the testing HMM towards a particular texture class is computed as the mean value of the similarity measures of the testing HMM towards all the 5 templates of that texture class. The identity of the testing HMM is assigned to the texture class which has the highest similarity measure towards the testing HMM.

For the purpose of observing the convergence property of the similarity measure when T gets big. We tested the recognition rates (the rates that the testing HMMs are correctly classified) on $T = 0, 2^0, 2^1, 2^2, ..., 2^{10}$. And the ρ is set to be 1/2. When $T = 0$, the classification is actually based on the similarity measure of the observation distributions (Gaussian) of two HMMs and the state transition matrix is in fact excluded from the computation of the similarity measure. Obviously, when T gets big, the influence of the state transition matrix in the computation of similarity score gets big. The recognition rate when $T = 0$ is 0.8042. The recognition rates of other settings of T's values are depicted in Fig. 3. An interesting observation is that, in this experiment, the recognition rate jumps up from 0.8042 at $T = 0$ to 0.9510 at around $T = 2^2$. Then the recognition rate goes down and converges at around 0.90. This may be explained by noticing that all three components of model parameters of HMM (A, B, π) share influences on the determination of similarity measures. As stated in previous sub section, π is excluded from our experiment for its inaccuracy due to limited training data. And when $T = 0$, the similarity measure is solely determined by B; when $T \to \infty$, A becomes more and more influential in the computation of similarity score. Then for T, there may be some point between 0 and ∞ at where the combination of individual contribution from A and B get maximized.

5. CONCLUSION

In this work we introduced a similarity measure of Gaussian HMMs. The proposed similarity measure are positive definite, symmetric and capable to handle a variety of generative models (including HMMs) in closed form. A fast procedure and a scaling procedure for the efficient and effective computation of similarity measures were presented. The similarity measure is evaluated on texture classification and encouraging results testified the effectiveness of the proposed method for similarity comparison between Gaussian HMMs. The method can be further generalized for the comparison of mixture Gaussian HMMs and more complicated stochastic models and may find potential applications in other data analysis areas.

6. REFERENCES

[1] B.H. Juang and L. Rabiner, "A probabilistic distance measure for hidden Markov models", *AT&T Tech. J.*, 64(2): 391-408, Feb. 1985.

[2] R.B. Lyngsø, C.N.S. Pedersen, and H. Nielsen, "Measures on hidden Markov models", Technical Report RS-99-6, BRICS, 1999.

[3] M.N. Do and M. Vetterli, "Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden Markov models," *IEEE Trans. Multimedia*, vol. 4, pp. 517-527, Dec. 2002.

[4] C. Bahlmann, and H. Burkhardt, "Measuring HMM similarity with the Bayes probability of error and its application to online handwriting recognition," ICDAR'01-IEEE *International Conference on Document Analysis and Recognition*, Seattle, Washington, USA, pp. 406-411. Sep. 2001.

[5] T. Jebara, and R. Kondor, "Bhattacharyya and expected likelihood kernels," COLT2003 *Conference on Learning Theory*, Washington D.C., USA, Aug. 2003.

[6] The USC-SIPI Image Database, `http://sipi.usc.edu/services/database/Database.html`.

[7] F. Aherne, N. Thacker, and P. Rockett, "The Bhattacharyya metric as an absolute similarity measure for frequency coded data," *Kybernetika*, 32(4):1-7, 1997.