FINGERPRINT MATCHING BASED ON DISTANCE METRIC LEARNING

Dalwon Jang and Chang D. Yoo

Div. of EE, School of EECS, KAIST, Yuseong Gu, Daejeon 305-701, Korea
dal1@kaist.ac.kr and cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper considers a method for learning a distance metric in a fingerprinting system which identifies a query content by measuring the distance between its fingerprint and a fingerprint stored in a database. A metric having a general form of the Mahalanobis distance is learned with the goal that the distance between fingerprints extracted from perceptually similar contents should be smaller than the distance between fingerprints extracted from perceptually dissimilar contents. The metric is learned by minimizing a cost function designed to achieve the goal. The cost function is convex, and the global minimum can be obtained using convex optimization. In our experiment, the distance metric learning is applied in an audio fingerprinting system, and it is experimentally shown that the learned distance metric improves the identification performance.

Index Terms— Fingerprinting, Identification, Distance measurement

1. INTRODUCTION

There is a growing demand for protecting, managing, and indexing digital content, and as a viable solution, fingerprinting is receiving increased attention. Fingerprinting is a technique that identifies an unknown content using a short feature vector called fingerprint. In recent years, various audio/video/image fingerprinting systems have been proposed [1]-[7].

A fingerprinting system for content identification generally consists of three essential components: fingerprint extraction, database (DB) search, and fingerprint matching [4]. In the fingerprint extraction process, a query fingerprint is extracted from a query content. In the DB search process, a set of candidate fingerprints from a DB close to the query fingerprint are obtained. In the fingerprint matching process, the distances between the candidate fingerprints and the query fingerprint are computed based on a distance metric. The fingerprinting system provides the meta-data associated with the closest candidate fingerprint.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2008-314- D00309)

The fingerprint extraction and matching processes influence the identification performance more than the DB search process which determines the computational efficiency of the system. The identification performance depends highly on the distance metric used in fingerprint matching process.

In this paper, a method for learning a distance metric in fingerprint matching is considered [8, 9, 10, 11]. In recent years, various literatures have shown that distance metric learning can improve classification and clustering performances [11]. The distance metric used in previous fingerprinting systems, which is not determined by learning, may not be suitable to the fingerprint used in the fingerprinting system and the distortions, thus the identification performance can be improved by metric learning.

By learning a distance metric from training data consisting of original and distorted contents, the identification performance can be improved. Fingerprints of original contents are assumed to be fingerprints stored in a DB, and fingerprints of distorted contents are assumed to be the query fingerprints. For correct identification, the distance of the fingerprint of a distorted content to the fingerprint of the original content from which the distorted content was obtained - called hereafter corresponding content - should be smaller than the distance to fingerprints of other original contents - called hereafter non-corresponding contents. A large distance margin should be established between fingerprints of the distorted and non-corresponding contents [10]. This is the goal of the distance metric learning considered in this paper, and specifically a distance metric having a general form of the Mahalanobis distance is considered. A cost function to be minimized is designed so that the cost increases when the fingerprint of the distorted content is further away from the fingerprint of the corresponding content than from fingerprints of non-corresponding contents. The parameter of the distance metric is determined by minimizing the cost function by convex optimization. We assume that the fingerprint is real valued, thus the distance metric learning considered in this paper is effective only for the real-valued fingerprint.

The remainder of this paper is organized as follows. Section 2 explains the distance metric, and Section 3 explains the cost function used to learn the distance metric. Section 4 presents the experimental results, and Section 5 concludes the paper.

2. DISTANCE METRIC

The considered distance metric measures the distance between two N-dimensional fingerprints $\mathbf{x}$ and $\mathbf{x}'$ as $||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2$ where $\phi : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is a mapping function. This paper considers a linear projection, thus $\phi(\cdot)$ is chosen as $\phi(\mathbf{x}) = \mathbf{W}\mathbf{x}$ where $\mathbf{W}$ is $N \times N$-dimensional matrix. The distance is computed as

$$||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2 = (\mathbf{x} - \mathbf{x}')^T \mathbf{A} (\mathbf{x} - \mathbf{x}') \quad (1)$$

where $\mathbf{A} = \mathbf{W}^T\mathbf{W}$. Thus, the distance metric has a general form of the Mahalanobis distance. Hereafter, the distance is denoted as $D_{\mathbf{A}}(\mathbf{x}, \mathbf{x}') = ||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2$. To learn the distance metric means to determine the matrix $\mathbf{A}$, which is the parameter of the distance metric. If $\mathbf{A}$ is the identity matrix, $D_{\mathbf{A}}(\mathbf{x}, \mathbf{x}')$ is the Euclidean distance.

3. LEARNING USING THE COST FUNCTION

3.1. Training data

To learn a distance metric, a set of training data consisting of the fingerprints of the original and the distorted contents, which are respectively denoted as $\mathbf{x}_i$ and $\mathbf{x}_{i,j}$ ($i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$), is required. The fingerprint $\mathbf{x}_{i,j}$ is extracted from the jth distorted version of the ith original content from which $\mathbf{x}_i$ is extracted. In the learning procedure, $\mathbf{x}_i$ represents the fingerprint stored in a DB, and $\mathbf{x}_{i,j}$ represents a query fingerprint. The pair $(\mathbf{x}_i, \mathbf{x}_{i,j})$ is a matching fingerprint pair, and the pair $(\mathbf{x}_k, \mathbf{x}_{i,j})$ $(k \neq i)$ is a non-matching pair. In this paper, distortions that often occur in real application are considered.

3.2. Cost function

The parameter of the distance metric $\mathbf{A}$ is determined so that a cost function is minimized. The cost function is minimized when $D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j})$ is smaller than $D_{\mathbf{A}}(\mathbf{x}_k, \mathbf{x}_{i,j})$ for $k \neq i$. To correctly identify a query content which is assumed to be a distorted version of the original content, $\mathbf{x}_{i,j}$ should be closer to $\mathbf{x}_i$ than to any $\mathbf{x}_k$ for $k \neq i$. The cost function considered in this paper is given by

$$\varepsilon(\mathbf{A}) = \sum_{i,j} [M + D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j}) - D_{\mathbf{A}}(\mathbf{x}_{\xi(i,j)}, \mathbf{x}_{i,j})]_+ \quad (2)$$

where $[\cdot]_+$ [1], M, and $\mathbf{x}_{\xi(i,j)}$ denote respectively the standard hinge loss function, margin, and the non-corresponding fingerprint closest to $\mathbf{x}_{i,j}$. The index $\xi(i,j)$ is mathematically expressed as

$$\xi(i,j) = \arg_{k,k\neq i} \min D_{\mathbf{A}}(\mathbf{x}_k, \mathbf{x}_{i,j}). \quad (3)$$

[1] $[z]_+ = \max(z, 0)$

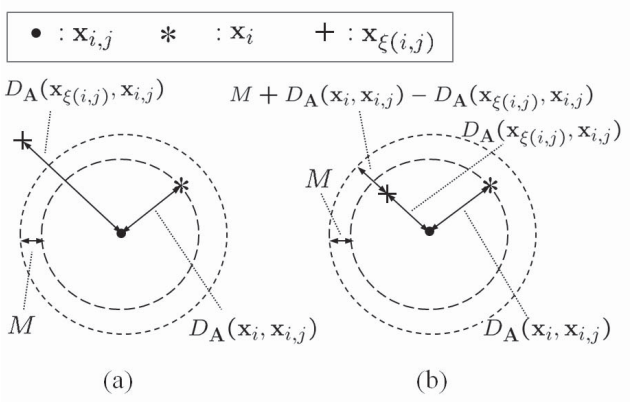

Fig. 1. Due to the hinge loss function, the cost function does not increase for the case (a) $M + D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j}) \leq D_{\mathbf{A}}(\mathbf{x}_{\xi(i,j)}, \mathbf{x}_{i,j})$, and the cost is added to the cost function for the case (b) $M + D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j}) > D_{\mathbf{A}}(\mathbf{x}_{\xi(i,j)}, \mathbf{x}_{i,j})$

Including both a constant M and the hinge loss function in Equation (2) and minimizing $\varepsilon(\mathbf{A})$ enforce $D_{\mathbf{A}}(\mathbf{x}_{\xi(i,j)}, \mathbf{x}_{i,j}) \geq M + D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j})$ [10]. Thus, the distance metric is learned so that the distances between query fingerprint $\mathbf{x}_{i,j}$ and its non-corresponding fingerprints are at least larger than $(M + D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j}))$. The summand of $\varepsilon(\mathbf{A})$ equals 0 when $\mathbf{x}_{\xi(i,j)}$ lies outside the ball centered at $\mathbf{x}_{i,j}$ with the radius of $(M + D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j}))$ as shown in Fig. 1 (a). But, the cost of $(M + D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j}) - D_{\mathbf{A}}(\mathbf{x}_{\xi(i,j)}, \mathbf{x}_{i,j}))$ is added to the cost function when $\mathbf{x}_{\xi(i,j)}$ lies within the ball as shown in Fig. 1 (b). Without loss of generality, we set $M = 1$ since $\mathbf{A}$ can be scaled by M.

3.3. Convexity of the cost function

The cost function is a convex function in $\mathbf{A}$, thus the global minimum can be obtained. To prove convexity, $\varepsilon(\mathbf{A})$ is rewritten as

$$\varepsilon(\mathbf{A}) = \sum_{i,j} [K(\mathbf{A}, i, j)]_+ \quad (4)$$

where $K(\mathbf{A}, i, j)$ is defined by

$$K(\mathbf{A}, i, j) = M + D_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_{i,j}) - D_{\mathbf{A}}(\mathbf{x}_{\xi(i,j)}, \mathbf{x}_{i,j}). \quad (5)$$

If $[K(\mathbf{A}, i, j)]_+$ is convex, then $\varepsilon(\mathbf{A})$ is also convex since a sum of convex functions is also convex. If the function $K(\mathbf{A}, i, j)$ is convex, then $[K(\mathbf{A}, i, j)]_+$ is also convex. Since $K(\mathbf{A}, i, j)$ is a sum of a constant and two linear functions, $K(\mathbf{A}, i, j)$ is linear with $\mathbf{A}$. Thus $K(\mathbf{A}, i, j)$ is convex, and $\varepsilon(\mathbf{A})$ is also convex.

3.4. Optimization

To find matrix $\mathbf{A}$, the projected gradient method is used [12]. The distance metric should be non-negative and satisfy the

Iterate

$$\mathbf{A} := \mathbf{A} - \beta \nabla_A \varepsilon(\mathbf{A})$$
$$\mathbf{A} := \arg_{\mathbf{A}'} \min\{||\mathbf{A}' - \mathbf{A}||_F : \mathbf{A}' \succeq 0\}$$

until $\mathbf{A}$ converges

Fig. 2. Procedure to find $\mathbf{A}$. Here, β is the step size, and $||\cdot||_F$ is the Frobenius norm ($||\mathbf{A}||_F = \sum_{i,j} A_{i,j}^2)^{1/2}$).

triangle inequality, $\mathbf{A}$ must be positive semidefinite [8]. The projected gradient method is performed in two steps. First, gradient decent method is used for unconstraint minimization problem. Then, $\mathbf{A}$ is projected to a space of positive semidefinite. The procedure to find $\mathbf{A}$ is shown in Fig. 2. The projection is performed by a semidefinite programming [12].

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

The performance improvement due to the the distance metric learning is presented in terms of the performance improvement in its application to an audio fingerprint system [4]. In the system proposed in [4], the 16-dimensional fingerprint is extracted in 371.5ms frame whose shift is 185.7ms, and the Euclidean distance is used in fingerprint matching. The fingerprint matching is performed using the fingerprint from 5 or 10s audio clip (27 or 54 frames), thus $N = 432$ or $N = 864$. Thus, in our experiment, the performance obtained by learned distance metric is compared with that obtained by the Euclidean distance when $N = 432$ or $N = 864$. In the experiment, the DB search process in fingerprinting system is excluded since only the performance of fingerprint matching is our concern.

Learning of an $N \times N$-dimensional matrix is computationally intractable since N is too large. Thus, in our experiment, an $M \times M$-dimensional matrix (denoted as $\mathbf{A_s}$) was learned instead of $\mathbf{A}$ ($M < N$). Using the M-dimensional fingerprint (fingerprint of $M/16$ frames) obtained by dividing N-dimensional fingerprint into M-dimension, $\mathbf{A_s}$ is determined. The distance between $\mathbf{x}$ and $\mathbf{x}'$ is computed as

$$\tilde{D_{\mathbf{A}}}(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{N/M} (\mathbf{x_s}^{(k)} - \mathbf{x_s}'^{(k)})^T \mathbf{A_s} (\mathbf{x_s}^{(k)} - \mathbf{x_s}'^{(k)}) \quad (6)$$

where $\mathbf{x_s}^{(k)}$ and $\mathbf{x_s}'^{(k)}$ denote the M-dimensional fingerprint obtained by dividing respectively $\mathbf{x}$ and $\mathbf{x}'$. In our experiment, $M = 48$, thus the summand in Equation (6) is the distance between fingerprints of 3 frames.

4.2. Training set

A set of training data from 100 different songs is used for distance metric learning. In our experiment, $I = 8000$ and $J = 4$. The list of audio distortions considered in our experiment is as follows [2]:

- **L1** Octave band equalization (**EQ1**): Adjacent band attenuations set to -6dB and +6dB in an alternating fashion.
- **L2** Echo (**E**): Filter-emulating old time radio.
- **L3** Band-pass filtering (**BPF**): 0.4-4kHz band pass filtering.
- **L4** WMA encoding (**WMA**): 64kbps WMA encoding.

For every distortion, 96kbps MP3 encoding (**MP3**) is followed.

4.3. Comparative test

For performance evaluation, 100 different songs completely separate from the training set were used as a test set. In the evaluation, 7 distortion were considered: EQ1, E, BPF, WMA, and the following 3.

- **T1** Time delay (**TD**): 92.9ms shift.
- **T2** Sampling rate change (**SR**): Down-sampling to 16kHz and up-sampling to 44.1kHz.
- **T3** 1/3 octave band equalization (**EQ2**): 30-band pop equalization.

The test sets of 3 combined distortions were also considered. Each combined distortion includes the distortions considered in the learning and not considered in the learning.

Fig. 3 compares the fingerprinting performances using the learned distance metric with those using the Euclidean distance by showing the receiver operating characteristic (ROC) curve. The ROC curve plots the false negative (FN) rate versus the false positive (FP) rate. The FN rate is defined as the rate that matching fingerprint pairs are determined as non-matching pairs, and the FP rate is defined as the rate that non-matching fingerprint pairs are determined as matching pairs. For each experiment, $60,000$ matching and $100,000,000$ non-matching fingerprint pairs were used. Fig. 3 (a)-(d) show the performance against the distortions considered in the learning, and Fig. 3 (e)-(g) show the performance against the distortions not considered in the learning. Fig. 3 (h)-(j) show the performance against combined distortions. As shown in the figure, the performances obtained using the learned distance metric are better than or comparable to the performances obtained using the Euclidean distance. The metric learning extremely improves the performance against the distortions of E and BPF which are the more serious distortions among the 4 distortions used in the learning. The identification performances are also extremely improved against the combined distortions which include E and BPF (Fig. 3 (i) and (j)). The metric learning does not degrade the performance against the distortions which were not considered in the learning process.

5. CONCLUSION AND FURTHER WORKS

In this paper, the method to improve the fingerprint matching process through distance metric learning is considered. By minimizing the cost function which concerns identification performance, the distance metric is learned. The cost function is designed to decrease when the query content is correctly identified. In our experiment using an audio fingerprinting system, it is shown that the distance metric learning improves the fingerprinting performance. The followings are left as further works. To confirm the improvement induced by the distance metric learning in fingerprinting, it is necessary to apply the distance metric learning to other fingerprinting systems. The distance metric learning for binary and index-valued fingerprints is also a further work.

6. REFERENCES

[1] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," *Proc. Int. Conf. Music Information Retrieval,*, 2002

[2] E. Allamanche, J. Herre, O. Helmuth, B. Frba, T Kasten, and M Cremer, "Content-based identification of audio material using MPEG-7 low level description," *Proc. Int. Symposium of Music Information Retrieval*, 2001

[3] C. Burges, J. Plat, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 3, pp. 165-174, May, 2003.

[4] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, C. D. Yoo, Audio Fingerprinting Based on Normalized Spectral Subband Moments, *IEEE Signal Processing letters*, vol. 13, issue 4, pp. 209-212, Apr., 2006.

[5] J. Oostveen, T. Kalker, and J. Haitsma, "Feature extraction and a database strategy for video fingerprinting," *Proc. Int. Conf. on Visual Information and Information Systems*, pp. 117-128, 2002.

[6] S. Lee and C. D. Yoo, "Robust video fingerprinting for content-Based video identification," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 7, pp. 983-988, July 2008.

[7] C. Kim, "Content-based image copy detection," *Signal Processing: Image Communication*, Vol. 18 (3), pp. 169-184, March 2003.

[8] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance Metric Learning, with application to Clustering with side-information," *Proc. NIPS* 2003.

[9] A. Globerson and S. Roweis, "Metric learning by collapsing classes," *Proc. NIPS* 2006.

[10] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Proc. NIPS* 2006.

[11] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Technical report*, Department of Computer Science and Engineering, Michigan State University, 2006.

[12] S. Boyd and L. Vandenberghe, "Convex Optimization," *Cambridge University Press*, 2004

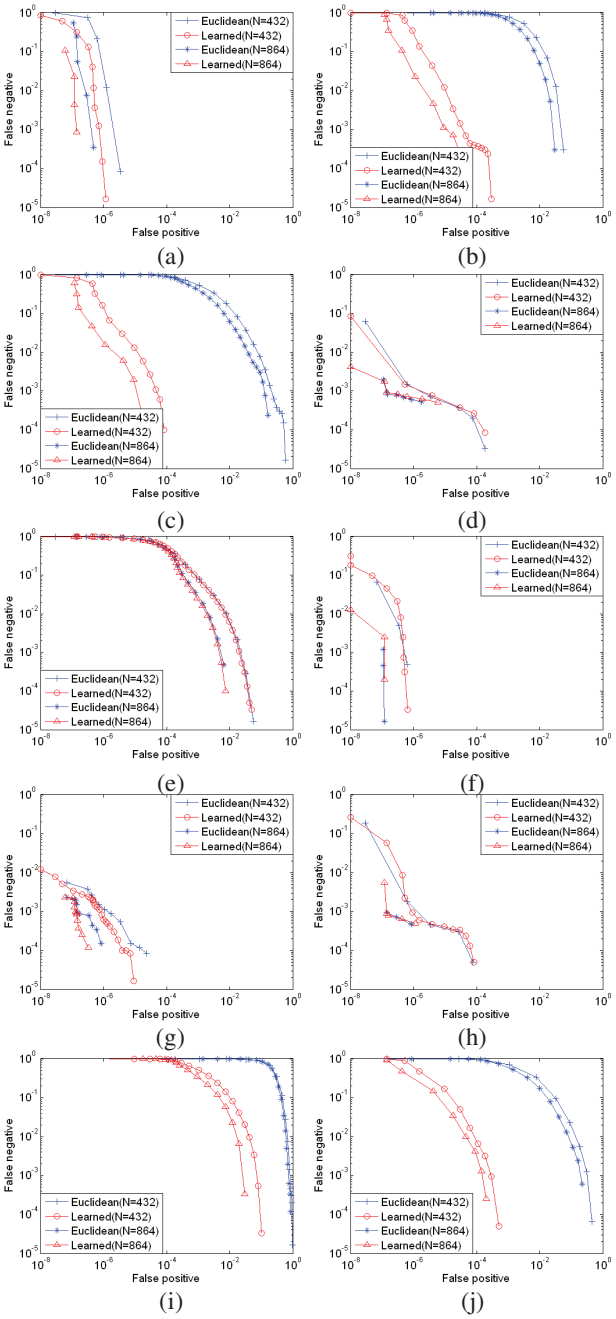

Fig. 3. ROC curves for various distortions: (a) EQ1+MP3, (b) E+MP3, (c) BPF+MP3, (d) WMA+MP3, (e) TD+MP3, (f) EQ2+MP3, (g) SR+MP3, (h) WMA+EQ2+SR+MP3, (i) TD+E+BPF+EQ2+MP3, and (j) EQ1+BPF+WMA+MP3.