An Effective and Efficient Method for Query by Humming System Based on Multi-Similarity Measurement Fusion

Lei Wang[1], Shen Huang[1], Sheng Hu[1], Jiaen Liang[1] and Bo Xu[1,2]
[1]*Digital Content Technology Research Center, Institute of Automation,*
[2]*National Lab of Pattern Recognition, Institute of Automation,*
Chinese Academy of Sciences, Beijing 100080, China
{Leiwang, shenhuang, Shu, jeliang, xubo}@hitic.ia.ac.cn

Abstract

Since it is the most natural way for people to search a specific melody in large music database, query by humming/singing is attracting more and more researchers' attention in the field of content-based music information retrieval. In this task, note-based and frame-based similarity measures are two commonly used methods. However, in previous works, researchers always focus on one of the two methods alone. In this paper, we propose a novel scheme taking advantage of two different similarity measurements to improve not only the retrieval accuracy but also the retrieving speed. First, Earth Mover's Distance (EMD), which is note-based and much faster, is adopted to eliminate most unlikely candidate. Then, Dynamic Time Warping (DTW), which is frame-based and more accurate, is executed on these surviving candidates. Finally, fusion strategies of these two similarity measurements are employed to improve the performance of whole system. Experiments show our approach can achieve 92.9% accuracy on the database used in MIREX 2006 QBH contest, which is better than those systems participated in that task

1. Introduction

In recent years with the increasing of music data on the internet and music databases, there is a strong demand for better procedures for automatic classifying, indexing and retrieving the massive music information. There are two main approaches in indexing and retrieving music files in large scale database: (a) keyword-based retrieval by matching the keyboard-input queries and human annotations stored in the keyword-database; (b) content-based retrieval by matching several seconds acoustic-input and the music score database. The latter one is more attracting since it could provide a natural and convenient way for customers to search in large music database, but also offer an extra interface when users forget the song's name but only remember a part of the melody. This field has been focused by a large group of researchers, and there is a contest hold by MIREX every year since 2006 [1]. In this task, note-based and frame-based similarity measures are two commonly used methods. Jang proposed a frame-based template matching strategy by calculating time series similarity with high precision [2], but this method is very time-consuming when the template database growing larger. Typke used transportation distance (EMD), which is a variation of note-based measurement, to achieve satisfying retrieval speed comparing to the frame-based method but loss of precision in some degrees [3].

In this paper, we propose a multiple similarity measurements fusion method to achieve better precision and lower computational cost by fusing frame-based and note-based similarity measurement together. Multiple similarity measurement fusion, a method to integrate multiple information sources together to improve recognition accuracy, has been tested effective and successfully adopted in many areas. The reason why this mechanism can improve performance is quite simple: Neither of the similarity measurement is perfect; each of the individual similarity measurement may produce some errors. Well-selected information sources that are complementary to each other could capture additional information that a single one could not gain. Therefore, combination of such multiple similarities would reduce overall error rate and emphasize correct outputs [4].

The outline of this paper is organized as follows. Section 2 offers a brief overview of our QBH system. In section 3, two similarity measures and their fusion method are introduced in detail. Experiment and result are given in section 4. Finally, conclusions are drawn in section 5.

978-1-4244-1724-7/08/$25.00 ©2008 IEEE

2. Overview of our QBH system

As shown in figure 1, the system we adopted is functionally divided into 3 parts: 1) music score processing module, taking music score file to generate song template database; 2) Acoustic processing module, taking user's humming/singing as acoustic input queries to generate frame-based and note-based features; 3) template matching module, computing the similarity between template and user's query using certain measurement and produce the final ranked candidates sorted by similarities.

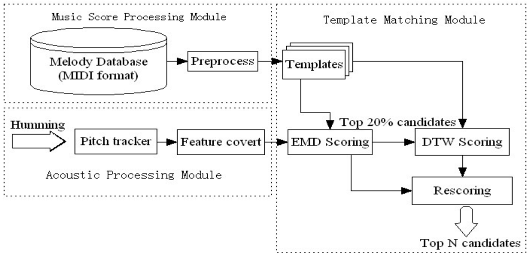

Fig.1. Overview of our QBH system

2.1. Music score processing module

Since it contains the exact information of music scores and it is prevailing on internet, we take MIDI files as our music score source to build up melody template database. MIDI files usually contain main melody track and other accompany tracks. Only main melody track is useful to build melody database. To save human's laborious work, we first use an automatic main melody track extraction method based on pattern recognition approach [5] to separate melody from other accompany track of each MIDI file. Each main melody extracted by this method is associated a confidence measure. Then, to make the main melody extraction result more credible, manually check by human-being is performed on these with low confidence.

2.2. Acoustic processing module

Comparing to the automatic speech recognition system, the query by humming/singing system only concern with these parameters that is conveying music information. Therefore, the arrangements of music notes and durations are enough to identify a unique song. This information could be generated from the pitch sequence extracted from the direct acoustic input.

We do not make any constraint on input methods, users could sing lyrics or humming a tune of a specific melody or even could query with meaningless syllables. When the acoustic input is obtained, the pitch tracker divides the acoustic input into 50 milliseconds frames and calculate fundamental frequency candidate using autocorrelation function in each frame. Then, dynamic programming method is adopted for finding a best path within these pitches candidates [6].

Since the midi scores are in semitone domain, after obtained the pitch sequence, we need to map pitch's frequency value to the semitone domain. The following Formula is used to transform frequency into the semitone [2]

$$Semitone = \log_2 \left(\frac{PitchFrequence}{440} \right) \times 12 + 69$$

Since no existing pitch extraction method is perfect, errors might be introduced during this stage. There are mainly two kinds of pitch errors: 1) double or half of the normal frequency; 2) small pitch fluctuation caused by jitter. To mitigate the impact caused by the two error source, a five-point median filter is adopted to generate smoother pitch sequence. This kind of smoothing method could not only alleviate double or half frequency error, but also remove jitter in singing voice.

2.3. Template matching module

The template matching module is in a multi-stage scheme. The first stage is based on Earth Mover's Distance, which is note-based and much faster, to eliminate most of unlikely candidates; the second stage is based on Dynamic Time Warping distance, which is frame-based and more accurate, to rescore these survival candidates to generate final decision. These two template matching method and their fusion method are introduced in detail in section 3.

3. Multiple distance rescoring

3.1. Note-based matching

Suppose one point set as "supplier" which is a number of hillocks, with which weight represented as the volume occupied. Another set as "demander" can be imaged as a mass of holes with a certain amount of capacities. EMD measures the minimum flow work needed to transport earth of these hillocks to fill the total amount of holes, with which the cost considering both capacities and two dimensional positions. Such kinds of problem can be transformed into the following linear programming [3, 7].

EMD has robust properties to errors brought about in the front end. Note segmentation mistakes to fragment note into two separate parts. In string matching, this action will induce an insertion cost and otherwise a deletion cost from query to template. But in EMD, Such error is trivial because EMD is in a topdown scale of view, the contribution of query notes is approximate equal to the notes of template, no matter how much fragmented or consolidated errors segmentation brings. As for insertion or deletion error, the influence of string matching is two units cost, but EMD is little, either. Moreover, a small altered weight or position of one note, taken as replacement error, also cause an infinitesimal change in the global EMD measure. Many note-based matching algorithms used in QBH systems are lack of using both pitch and tempo information. Except that, DP based matching works in a bottom-up fashion. However, Long time information such as tempo and duration exist in a global view and are difficult to capture by bottom-up methods. But EMD measure works with optimization of both global pitch and tempo information in implicitly.

3.2. Frame-based matching

Traditional similarity measure method, such as Euclidean distance, is very sensitive to missing a data point, or having additional noise points. This is particularly the case for the query by humming system. The user's input may be performed faster or slower comparing to the template stored in melody database, thus, the Euclidean distance would produce significant error. To overcome this problem, dynamic time warping (DTW) algorithm is adopted to fill the gap caused by tempo variation between each individual user's acoustic input and music score template.

The DTW algorithm is widely used for comparing time series with non-linear distortions in the time axis. This method exhibits good performance in isolate word speech recognition and query by humming [2]. As shown in following Formula,

$$D[i,j] = d[i,j] + \min \begin{cases} D[i-2, j-1] \\ D[i-1, j-1] \\ D[i-1, j-2] \end{cases}$$

D[i, j] is the minimum distance starting from the beginning of the DTW table to the current position [i, j], and d[i, j] is the distance between sample i of querying signal and sample j of template. DTW algorithm searches the path with the least global distance from the beginning D[0, 0] to the ending D[M, N]. The best path is the one with the least global distance, which is the sum of cells alone the path.

3.3. Multi-Similarity measurement fusion

Since no existing similarity measurement is perfect, and it is hard to develop a better new one, it is necessary to fusion the EMD and DTW similarity output together to make the final decision. Actually, the combination of different similarity measurement is a way of uncertainty reduction. There are many ways to fusion these multiple information together to get better results, such as dynamic information selection, multi-information structuring and grouping, Hierarchical mixture of experts, candidate set reduction, parallel voting strategy and so on [4]. Among these methods, candidate set reduction and voting fusion methods are the most simple and straight way to take advantage of the multi-source information.

To take advantage of different information sources, it is reasonable to try to reduce the candidate set by one measurement and rescore by another one. Since EMD is much faster than DTW, we first use EMD similarity measurement to rank the whole set of candidates; then a threshold is taken to remove these candidates which is decided unlikely to be true by this measurement; finally, DTW algorithm is used to sort these surviving candidates. This method use the faster mechanism to filter out these candidates which is obvious not the true answer, and the uncertainty for the DTW is reduced to a minimum level so that the error rate would be reduced.

A more sophisticated multi-source information fusion method is the parallel voting strategy. If only ranked candidate set are available, the voting fusion is often used as a decision level fusion. Nevertheless, in this method all the similarity measurements are considered equally important so that a less reliable one may disturb the final decision significantly. A possible way of overcoming this drawback is that of including a measure of the importance of each similarity measurements in the fusion rules. The similarity measurement with higher accuracy is more likely to be correct and should be given more voting weight. Thus, the weighted voting fusion can be expected a high performing solutions by assigning appropriate degrees of weight or each information source to vote. The weighted voting function is

$$score(t_j) = \sum_{i=1}^{M} weight(c_i) prob(c_i, t_j)$$

Where $weight(c_i)$ is the weight of similarity measurement c_i and $prob(c_i, t_j)$ is the similarity obtained when c_i runs on candidate t_j , and the candidate with the highest voting scoring is assigned as the winner:

$$\mathrm{t}_j^* = \arg\max_{\mathrm{j}} score(t_j)$$

Furthermore, this fusion mechanism does not need additional computation cost. As shown in figure 1, since the EMD runs much faster than DTW, we perform EMD at first, and reserve top N of candidate for the DTW input. The top N candidate's EMD distance is also saved for the fusion step. Therefore, when performing EMD measurement, we achieve two objectives with a single effort, which is not only the candidate set is reduced, but the similarity of EMD measurement is also saved for the next fusion stage.

4. Experiments

In our experiments, the dataset used in the *MIREX 2006 Query by Singing/Humming Contest* is taken to compare with the other researchers' results. This corpus [8] contains 48 monophonic MIDI files plus 2000 non-target files as the melody database, and 2797 sung queries from 118 persons as the acoustic input. The non-target files are selected from ESSAN collection [9]. Since the contest does not provide the exact file list of these non-target files, to make our result comparable, we select 2000 randomly from the ESSAN collection. These sung queries were humming/singing from the beginning of the song and last for exactly 8 seconds. The way for singers to query is various: some of these song queries are sung with lyrics, some are hummed nasal voice, even some are queried with nonsense syllables. All the querying files are digitized at 8 kHz, 8 Bit, PCM format.

We compared EMD, DTW and fusion of the two similarity measurement on the same corpus. The performance is evaluated in three measurements: Mean Reciprocal Rank (MRR), Top-1 rate, and Top-20 rate. MRR is calculated by the reciprocal of the rank of the first correctly identified cover for each query (1/rank) [1]. Top-1 rate is the percentage that the correct melody is ranked in the first place of all returns. Top-20 rate is the rate that the correct melody is ranked within the top 20 of all returns. Fig.2 illustrates the performance of different surviving rate of candidate set reduction fusion method.

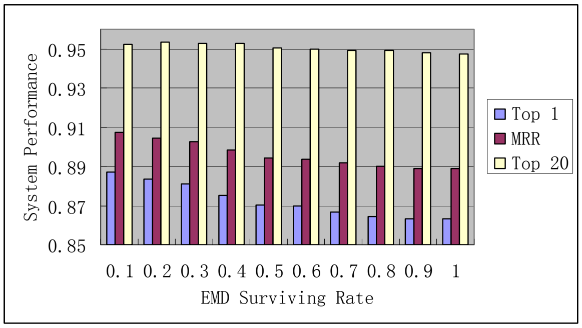

Fig.2. Evaluation of different survival rate of candidate set reduction fusion

It can be seen from this figure that the whole system performance get better when the survival rate of EMD stage get smaller. This could be explained as follow: the DTW similarity measurement may be disturbed by some wrong candidates, but most of these disturbing candidates could be easily eliminated by EMD method, a complementary measurement.

When evaluating EMD-DTW weighted rescoring method, we set the EMD survival rate to be 3%. The weight of EMD and DTW is set to be 40% and 60% respectively. From table 1, the rescoring of EMD and DTW reaches 92.90% of MRR. This result performs better than the systems participated in QBSH task, MIREX 2006 [1].

When evaluating the whole system speed, we test our system on the computer with Intel Pentium4 3.0GHZ, 1G memory. As shown in table 2, EMD runs much faster than DTW. Therefore, using EMD as the first stage filter could improve the searching speed significantly. We got 0.139 real time ratio when the EMD survival rate is set to be 3%.

Table 1 The performance of different methods

Method	TOP-1 (%)	TOP-20 (%)	MRR (%)
EMD (alone)	71.89%	80.10%	74.98%
DTW (alone)	86.35%	94.74%	88.88%
EMD+DTW (rescore)	91.49%	96.38%	92.90%

Table 2 The speed of different methods

Method	Real Time Ratio
EMD (alone)	0.094
DTW (alone)	1.51
EMD+DTW	0.139

5. Conclusion

In this paper, a multiple distance measurement fusion mechanism is adopted in query by humming/singing system to improve the retrieval accuracy and speed. The proposed method is evaluated on the dataset used in the *ISMIR2006 Query by Singing/Humming Contest*. Experimental results show that the performance of the proposed based method is better than using one distance measurement alone and achieving satisfying result comparing to other researchers.

6. References

[1] http://www.music-ir.org/mirex/2006/index.php/

[2] Roger Jang."Hierarchical Filtering Method for Content-based Music Retrieval via Acoustic Input", ACM-MC, 2001.

[3] R.Typke. "Using transportation distances for measuring melodic similarity". ISMIR. 2003.

[4] Ruta, D, Gabrys, B. "An overview of classifier fusion methods." Computing and Information Systems 7(1):1--10 (2000).

[5] David Rizo "A Pattern Recognition Approach for Melody Track Selection in MIDI Files". ISMIR, 2006.

[6] Paul Boersma "Accurate Short-Term Analysis Of The Fundamental Frequency And The Harmonics-To-Noise Ratio Of A Sampled Sound". ISMIR, 2006.

[7] Shen Huang "Query By Humming Via Multiscale Transportation Distance In Random Query Occurence Context", ICME 2008

[8] *http://neural.cs.nthu.edu.tw/jang2/dataSet/childSong4public/*QBSH-corpus

[9] http://www.esac-d*ata.org/*