Distribution-Based Similarity for Multi-Represented Multimedia Objects

Hans-Peter Kriegel, Peter Kunath, Alexey Pryakhin, Matthias Schubert

Institute for Informatics, Ludwig-Maximilians University, Munich, Germany
{kriegel,kunath,pryakhin,schubert}@dbs.ifi.lmu.de

Abstract. In modern multimedia databases, objects can be represented by a large variety of feature representations. In order to employ all available information in a best possible way, a joint statement about object similarity must be derived. In this paper, we present a novel technique for multi-represented similarity estimation which is based on probability distributions modeling the connection between the distance value and object similarity. To tune these distribution functions to model the similarity in each representation, we propose a bootstrapping approach maximizing the agreement between the distributions. Thus, we capture the general notion of similarity which is implicitly given by the distance relationships in the available feature representations. Thus, our approach does not need any training examples. In our experimental evaluation, we demonstrate that our new approach offers superior precision and recall compared to standard similarity measures on a real world audio data set.

1 Introduction

Similarity search and content-based object retrieval are important topics when handling multimedia data like sound files or music titles. In recent years, the research community introduced a large variety of feature transformations for all different types of sound data and music titles [1]. Systems like Muscle Fish [2] employ multiple representations like loudness, pitch, or harmonicity for the retrieval of audio data. Additionally, multiple representations are a common setting in various other areas of multimedia data, such as image or video data. Since most of the feature transformations focus on different aspects of the objects, it is beneficial to use more than one feature representation when processing similarity queries. Thus, the most suitable feature representations need to be selected or combined.

In this paper, we introduce a novel technique for combining distance values being observed in multiple representations. In our context, a representation is a method for calculating a distance value between two objects. Our approach assumes that there exist two probability distributions modeling the likelihood that a user would state that two objects displaying a certain distance value in some representation are similar or dissimilar. These distributions are influenced by the degree of similarity between the objects as well as the complete set of objects displaying the same given distance value. Furthermore, we assume that both distributions are connected by a third distribution modeling the ambiguity of the distance value, e.g. the likelihood that both distributions display the same distance value. Now, we can derive the likelihood that a given distance

value has a definite meaning, i.e. the likelihood for an ambiguous meaning is rather small. Since small distances correspond to similar objects, we can assume that distance values being smaller than all distances having an ambiguous meaning indicate object similarity. Based on the local likelihood for similarity in each representation, we can now calculate a joint probability score indicating object similarity.

An additional contribution of this paper is a method for explicitly describing the local similarity likelihoods as explicit functions. Furthermore, to fit the distributions to the given application, we propose an iterative algorithm which is based on maximizing the agreement between representations. To run this algorithm, it is not necessary to manually label pairs of similar objects. Instead our new approach is based on the fact that the feature representations are usually selected to be useful for the given application. Under the condition of meaningful feature transformations, very small distance values will imply similarity and very large distance values will imply dissimilarity. Exploiting this observation, our new algorithm finds a meaningful parametrization within a few iterations.

The rest of the paper is organized as follows. In Section 2, we survey related work in the area of multi-represented similarity search. Our new theory about mapping distances to similarity statements is introduced in Section 3. Section 4 presents solutions for calculating the similarity likelihood and unsupervised parameter tuning. Our experimental evaluation is presented in section 5. The paper concludes in Section 6 with a summary and some directions for future work.

2 Related Work

Considering objects with multiple representations has attracted more and more attention in the multimedia research community. The approaches proposed over the last years can be divided in two categories, namely supervised and unsupervised. The supervised techniques either revert to user feedback or assume that labeled data with known affiliation of objects to some classes are available. For both categories, a common framework is the use of a weighted linear combination.

Supervised by user feedback. Various approaches have been proposed to compute the weights with the help of user feedback. To give feedback, a user has to label if an object being retrieved by a similarity query is really similar to the given query object. In general, the weighted average of the distances observed in all representations was reported to provide good results in several publications (e.g., [3]). For instance, the approaches in [4, 5] compute the weights based on the idea of relevance feedback. The authors of [6] suggest another relevance feedback based technique. This technique implements a weighted distance approach that uses standard deviations of the features. Another example is logistic regression w.r.t. user feedback [7]. From the user's point of view, it is rather inconvenient to provide feedback several times to get the result. In contrast, our method can operate without any training objects.

The interactive search fusion method [8] provides a set of fusion functions, e.g. min, max, sum and product function that can be used for combining different representations in order to improve the effectiveness of similarity search. This method supports a manual and an interactive search that is supervised by the user's assistance or by a

user-defined query. In addition, Boolean operators on aggregation functions are supported, e.g. "AND" can be applied to the product aggregation function. This technique requires strong interaction with the user. This is not always desirable because in order to use this method the user has to understand its concepts first.

Supervised by labeled data. The authors of [9] introduce two methods for improving the effectiveness in a retrieval system that operates on multiple representations of 3D objects. The proposed techniques are based on the entropy impurity measure. The first method chooses the best representation w.r.t. a given query object. The second method performs a so-called dynamic weighting of the available representations that is computed at query time, and that depends on entropy impurity in the local neighborhood of a query object. This work also presents encouraging experimental results that demonstrate a significant improvement in effectiveness of the similarity search for both proposed techniques. The methods described in [9] need a set of labeled data in order to measure entropy impurity.

Unsupervised. An unsupervised way to determine the weights for a linear combination is counting the number of representations for which a similarity larger than zero is observed [3]. This method is strongly dependent on the occurrence of zero distances. An unsupervised technique for the weighted combination of multiple representations for similarity search in multimedia databases was proposed in [10]. This technique exploits the fact that it is often beneficial to summarize multimedia data, like e.g. videos, in order to achieve higher efficiency during query processing. Compared to our new method, this method has the general drawback that it is only applicable together with summarization. However, object summarization is not a necessary element of general multi-represented similarity search. In [11], a template matching method based on the time warping distance is presented. This approach can measure the temporal edit similarity in order to process audio-visual similarity queries. However, temporal order is not necessary in many applications.

3 A Distribution-Based approach for Similarity Estimation

In this section, we will introduce our new method for estimating the similarity between two multimedia objects based on multiple feature representations. We can formalize a multi-represented object as follows:

Definition 1 (Multi-Represented Object).
Let $R = \{R_1, \ldots, R_n\}$ be a set of feature spaces. For each feature space R_i there exists a distance function $d_i : R_i \times R_i \rightarrow \mathbb{R}_0^+$. A multi-represented object o over the representations R is given by the n tuple $o = (r_1, \ldots, r_n) \in R_1 \times \ldots \times R_n$.

To compare two multi-represented objects o_1 and o_2, we have to combine the distances that can be derived from each representation. In general, most systems combine the distances using the weighted average over all distances and all representation spaces [3, 9, 10]. Though this standard approach is adjustable by varying the weights, it considers that the dissimilarity is linearly decreasing in all representations. However, we argue that very small and very large distances indicate a more clear statement about object similarity and dissimilarity, respectively, than medium distance values and thus, have to be treated in a different way.

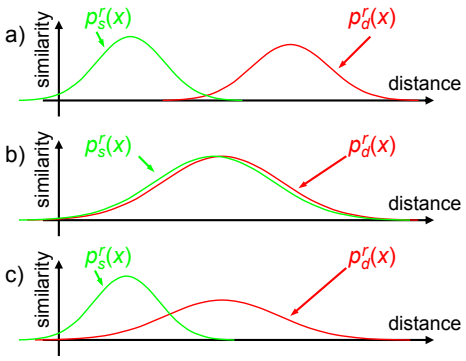


Fig. 1. Three examples for the distributions of similar and dissimilar distances.

In our new model, we want to quantify similarity as the likelihood that a user would assign the label "similar" to the comparison of o_k and o_l given the distance value $d_i(o_k, o_l)$ in representation R_i. Let us note that in this model the complementary label to "similar" is not "dissimilar". Instead the user would assign a label "unknown" in the case if it is not clear whether the object should be considered as similar or not. Correspondingly, we can consider the likelihood that a user would assign the label "dissimilar" to both objects. For this distribution, not labeling two objects as similar would also indicate "unknown". As a conclusion, we can distinguish two types of distributions mapping a distance value to "similar", "dissimilar" or "unknown". The first distribution describes the likelihood that a user would consider two objects clearly as similar and the second distribution describes the likelihood that a user would consider two objects clearly as dissimilar. Thus, both distributions are not necessarily complementary because a user could be ambiguous about object similarity as well. However, both distributions are implicitly connected by a third distribution modeling the ambiguity of meaning of a given distance value.

In the following, we will refer to the probability density function over the distances in representation R_i implying similarity as $p_s^i(x)$ and the corresponding density function modeling the distribution of dissimilarity as $p_d^i(x)$. Figure 1 displays three examples of both distributions over the distance values in a data space. For this example, we assume Gaussian distributions. In Subfigure a), $p_s^i(x)$ and $p_d^i(x)$ are rather well separated. Therefore, a ranking query in this representation would most likely retrieve similar objects first and afterwards objects being dissimilar to the given query object. Subfigure b) displays the other extreme. Both distributions are rather identical. Thus, the distance in this representation is rather uncorrelated with object similarity. Therefore, similar as well as dissimilar objects are encountered for any given distance value. Finally, subfigure c) illustrates a more realistic distribution. $p_s^i(x)$ is much more dense for small distances and displays a small density for large distances. $p_d^i(x)$ displays medium densities for a large spectrum of small and large distance values. A corresponding ranking, would start with similar as well as with dissimilar objects. However, after a certain distance is reached, the ranking would contain mostly dissimilar objects.

Based on this observation, we can measure the suitability of a given representation R_i to distinguish similar from dissimilar objects. To measure how meaningful observing

a certain distance δ between two objects o_k and o_l is, we can consider the likelihood that the distance δ is observed for similar as well as for dissimilar objects. If δ is common in both distributions $p_s^i(x)$ and $p_d^i(x)$ the meaning is rather ambiguous. However, if it is much likelier that δ is observed for similar objects, the meaning of δ is rather definite.

Formally, the complete likelihood that both distributions display the same distance in representation R_i can be expressed as follows:

$$p^i_{ambiguous} = \int_{-\infty}^{\infty} p_s^i(x) p_d^i(x) dx$$

The product of densities corresponds to the event that both distributions display the same distance and the integral sums up this joint density for all possible distance values. Since $p^i_{ambiguous}$ describes the complete amount of ambiguity in representation i, it can be considered as a measure of how far a given representation is suitable to model similarity. If $p^i_{ambiguous}$ is rather large, the implication of most distance values in R_i on similarity will be rather small. To calculate the probability that a given interval of distance values $[a, b]$ contains more ambiguous distances than the rest, we can determine the portion of the error relative to $p^i_{ambiguous}$:

$$P^i_{ambiguous}(a, b) = \frac{\int_a^b p_s^i(x) p_d^i(x) dx}{p^i_{ambiguous}}$$

To determine the probability that a distance $x < \delta$ has a definite meaning, we can now determine the probability of the ambiguity of distances which are larger than δ:

$$P^i_{definite}(x < \delta) = P^i_{ambiguous}(x \geq \delta) = \frac{\int_\delta^\infty p_s^i(x) \cdot p_d^i(x) dx}{p^i_{ambiguous}}$$

Since distances between similar objects are naturally smaller than distances between dissimilar objects, we can conclude that $p^i_{definite}(x < \delta)$ corresponds to the likelihood that two objects having a distance of δ or smaller are similar. Formally, we can define the local similarity likelihood in representation R_i as follows:

Definition 2 (Local Similarity Likelihood in R_i). *Let o_k and o_l be multi-represented objects over the representations $R = \{R_1, \ldots, R_n\}$. Then the similarity Likelihood for the comparison of o_k and o_l in representation R_i is defined as follows:*

$$L^i_S(o_k, o_l) = p^i_{definite}(x < d_i(o_k, o_l))$$

After having formalized the local similarity likelihood for representation R_i, we can now define the complete similarity likelihood over all representations:

Definition 3 (Similarity Likelihood).
Let o_k and o_l be multi-represented objects over the feature spaces or representations $R = \{R_1, \ldots, R_n\}$. For each representation R_i, we consider the similarity likelihood $L^i_S(x)$. The similarity likelihood P_{SIM} between o_k and o_l is defined as follows:

$$P_{\text{SIM}}(o_k, o_l) = \prod_{R_i \in R} L^i_S(o_k, o_l),$$

Let us note that the similarity likelihood assumes independence between the representations. This is a valid assumption because the benefit of combining multiple representations strongly depends on the use of independent object representations. Thus, when using a set of representations which always have the same implication on object similarity, any combination rule will follow this implication anyway.

4 Efficient Calculation and Parameter Fitting

After describing the general model, we will now turn to calculating the similarity likelihood. Therefore, we need to explicitly describe our distribution functions and fit the parameters of the distributions to the data objects in the given application.

In order to select a suitable distribution function, we have to examine co-occurrence between distance values and object similarity. Let us note that our technique can work with an arbitrary probability distribution function. In our application, we observed that Gaussian distribution functions seem to be a suitable description of the density of object pairs being labeled as similar or dissimilar. Furthermore, when modeling $p_s^i(x)$ and $p_d^i(x)$ as Gaussian distribution, it can be shown that $p_{ambiguous}^i(x)$ again follows a Gaussian distribution.

A Gaussian distribution function is characterized by the mean value and the standard deviation which can be calculated in a straight-forward way.

In order to calculate $L_S^i(o_k, o_l)$, we would need to integrate over a Gaussian distribution function modeling $p_{ambiguous}^i(x)$. Unfortunately, there is no known antiderivative for the Gaussian density function and thus, we need to employ an approximation in order to calculate our similarity likelihoods. To solve this problem, we employ the sigmoid function which is quite often used to approximate the integral from $-\infty$ to a or from a to ∞ over a Gaussian distribution. The sigmoid function is defined as follows:

$$sig_{\alpha,\beta}(x) = \frac{1}{1 + \exp(\alpha \cdot x + \beta)}$$

The local similarity likelihood is modeled by a sigmoid function having a negative α-value. To determine the sigmoid function approximating the cumulative density for a given Gaussian, there are various methods. In our system, we employed the following: We derived a set of sample points by cumulating the values of the Gaussian. Afterwards we employed the method of Levenberg and Marquadt [12] for fitting a sigmoid function to the sample points.

Finally, to employ the similarity likelihood, we have to find suitable function parameters for each representation. Our method is based on distinguishing the distributions of distance values implying similarity and dissimilarity. Thus, we need to find a way for approximating the likelihood that two objects are similar or dissimilar. Our solution to this problem is based on the following observation. If the distances observed in all representations for a given pair of objects are rather small, it is very likely that the objects are considered as similar. Correspondingly, if the distances between two objects are rather large in all representations, we can assume that the objects are dissimilar. To apply this observation for determining a good parametrization, we have to find a way

```
FUNCTION unsupervisedParameterTuning()
  D = generateDistanceVectors()
  FOR EACH representation R_i ∈ R DO
    initSimilarityLikelihoods(R_i,D)
  END FOR
  DO
    L_S^AVG = calculateAVGSIM(D)
    FOR EACH representation R_i ∈ R DO
      calcDistributions(D, L_S^AVG)
      p^i_ambiguous = calcAmbiguity(D, L_S^AVG)
      approximateSimilarityLikelihoods(D, L_S^AVG,p^i_ambiguous)
    END FOR
  WHILE(old parameters ≠ new parameters)
```

Fig. 2. The algorithm for unsupervised parameter tuning.

to maximize the agreement between the similarity likelihoods. In other words, the density functions modeling the ambiguity of a distance in each representation should be synchronized in a way that $d_i(o_k, o_l)$ has a comparable probability density in representation R_i as $d_j(o_k, o_l)$ in representation R_j. In order to avoid a costly computation of the distances between all database objects, we sample a small example data set S. We observed in our experiments that using about 100 - 200 objects is sufficient to compute a good approximation. Formally, we can capture the agreement between the similarity likelihoods in all representations by the average variance of estimate values on a sample data set S:

Definition 4. *Let $R = \{R_1, \ldots, R_n\}$ be a set of representations and let $L_S^i(o_k, o_l)$ be the similarity for R_i with $1 \leq i \leq n$. Then the average similarity $L_S^{\mathrm{AVG}}(o_k, o_l)$ for the comparison of two multi-represented objects o_k, o_l is defined as :*

$$L_S^{\mathrm{AVG}}(o_k, o_l) = \frac{\sum_{R_i \in R} L_S^i(o_k, o_l)}{|R|}$$

Consequently, the average variance of a given example set S consisting of multi-represented objects is given as:

$$Var(S) = \sum_{o_n, o_m \in S} \sum_{R_i \in R} (L_S^i(o_n, o_m) - L_S^{\mathrm{AVG}}(o_n, o_m))^2$$

After providing a measure of the agreement between the distributions in each representation on a given example set S, we now introduce an iterative method to find estimate parameters minimizing $Var(S)$. Our method iteratively minimizes a target function, i.e. $Var(S)$, and updates the parameters in each representation to better resemble the similarity value induced by $L_S^{\mathrm{AVG}}(o_k, o_l)$.

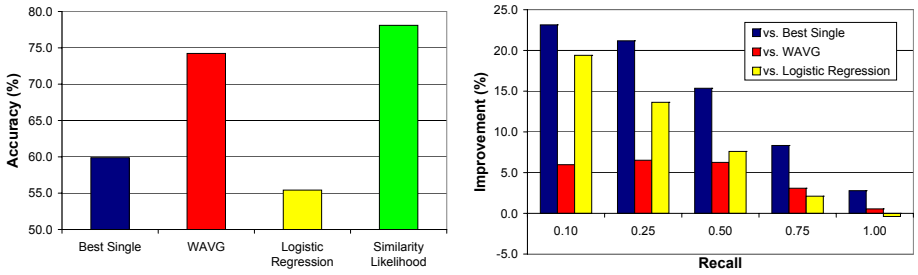
In the following, we describe the method in more detail. Figure 2 depicts the algorithm in pseudo code. In the initialization step, we calculate all distances in all rep-

resentations between two example objects. As a result, each object comparison is described by a distance vector $\overrightarrow{d_{m,n}}$ of dimensionality $|R|$ containing the distance in each representation. It makes sense to store the distance vectors for all pairs of objects to avoid recalculating the distances in each iteration. To derive an initial parametrization for the distribution function in representation R_i, all distance vectors are sorted w.r.t. the i-th component. The initial similarity likelihoods are now computed under the assumption that the s smallest observed distance values correspond to similarity. Based on this initial notion of similarity, the parameters of the underlying Gaussian distribution are derived. To get an initialization of dissimilarity, we assume that the d largest distances correspond to dissimilarity. After approximating both distributions, we can derive $p^i_{ambiguous}$ and fit a sigmoid to model $L^{\boldsymbol{i}}_S(o_k, o_l)$. Now the algorithm can enter the iteration loop. In a first step, we calculate $L^{\mathrm{AVG}}_S(o_k, o_l)$ for each distance vector and thus, receive a joint notion of similarity for each object comparison.

After building the current notion of similarity, we first of all can update the Gaussian distributions describing similar and dissimilar distances in each representation. Each distance is weighted with the current likelihood that the corresponding distance vector induces similarity or dissimilarity. After updating the distributions, we continue by calculating $p^i_{ambiguous}$ and transform both Gaussian distributions into our probability distribution describing ambiguity. Now the sigmoid functions describing $L^{\boldsymbol{i}}_S(o_k, o_l)$ can be updated to fit the joint notion of similarity in a better way. After having updated the similarity in each representation, we can check if any parameter value was indeed optimized. If this was the case, we proceed with an additional iteration. If the parameters did not change, the algorithm converges and we have found a suitable set of parameters maximizing the agreement between distributions. In our experiments, we observed that the algorithm usually terminates after 5 to 10 iterations.

5 Evaluation

All methods were implemented in Java 1.5, the experiments were performed on a workstation having 2GB main memory and an Intel Pentium IV (2.6 GHz) processor. We conducted our experiments on a music collection consisting of almost 500 songs which were taken from 15 different musical genres as the basis for the audio data set. We generated 6 different feature representations per song and depending on the representation, we extracted 30 to 300 features per second. Timbre features are derived from the frequency domain and were mainly developed for the purpose of speech recognition. The extraction of the timbral texture is performed by computing the short time fourier transform. We use the Mel-frequency cepstral coefficients (MFCCs), spectral flux, spectral deviation and spectral rolloff as timbral representations [1]. Rhythmic content features are useful for describing the beat frequency and beat strength of a piece of music. Features derived from beat histograms [1] are used for the description of the rhythmic content. Pitch extraction tries to model the human perception by simulating the behavior of the cochlea. Similar to the rhythmic content features, we derive pitch features from pitch histograms which were generated by a multipitch analysis model [13]. For each representation, we applied a vantage point based instance reduction [14] and transformed the corresponding feature vector set into one feature vector of about 500 dimensions.



(a) Accuracy of NN classification. (b) Precision/Recall improvement.

Fig. 3. Quality results on audio data set.

We compared our approach with three competitors, namely logistic regression [7], the weighted distances approach "WAVG", and the best single representation "Best Single", which provides the best single model for similarity and dissimilarity. The weights for each representation for WAVG are selected in the following way which is known as mean normalization. We normalize the distance with regard to the mean value μ_i^{orig} of the original distance $dist_i^{orig}$ in representation R_i, i.e. $dist_i^{\text{norm}}(o,q) = dist_i^{\text{orig}}(o,q)/\mu_i^{\text{orig}}$. The mean value can be approximated by sampling a small set of objects from the current representation R_i.

Nearest Neighbor Classification. In this type of experiment, we employed nearest neighbor classification in combination with 10-fold cross validation to avoid overfitting. We employed our unsupervised method for parameter fitting and compared the classification accuracy of a nearest neighbor classifier. The results are depicted in Figure 3(a). The classification accuracy when using the similarity likelihood was higher than the accuracy of the classification employing logistic regression, WAVG, or any of the underlying representations. Thus, the similarity likelihood provided a even better notion of similarity than the supervised approach employing logistic regression.

Precision-Recall Results. A final type of experiment tries to capture precision and recall of the proposed method by constructing so-called precision-recall graphs. For this type of experiment a ranking query is posed for each object in the test set. For each query, we now measure the precision for 5 different levels of recall. Afterwards, the average precision for each recall level over all queries is computed. Thus, the 0.25 bin of the precision/recall graph corresponds to average precision observed in the result sets containing 25 % of the objects belonging the same class as the query object. We examined the performance of each representation separately and additionally, examined the performance of WAVG and logistic regression (cf. Figure 3(b)). For all measured recall values, the similarity likelihood achieves a significantly higher precision than both WAVG and the best single representation. Only logistic regression was able to slightly outperform our method for very high recall values.

6 Conclusions

In this paper, we proposed a novel method for multi-represented similarity search. Unlike previous methods which only focus on achieving comparability between the dis-

tances derived in each representation space, our new method additionally distinguishes the meaning of distance relationships in each representation. To capture this meaning, we define a so-called similarity likelihood which approximates the probability that two compared objects are truly similar. We propose an unsupervised method which finds parameters that resemble the current notion of similarity in a best possible way. This method employs the idea that the selection of representation spaces itself yields an implicit statement about the notion of similarity of the given application. Thus, our approach maximizes the agreement between the similarity likelihoods in the given representations in order to find a meaningful parameter setting. In our experimental evaluation, we demonstrate on a real world audio data set that our new likelihood based technique outperforms standard combination methods w.r.t precision and recall. For future work, we plan to investigate the use of other distribution functions to model distances between similar and dissimilar objects.

References

1. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE TSAP **10**(5) (2002) 293–302
2. Wold, E., Blum, T., Keislar, D., Wheaton, J.: Content-based classification, search, and retrieval of audio. IEEE Multimedia **3**(3) (1996) 27–36
3. Lee, J.: Analyses of multiple evidence combination. In: ACM SIGIR Conf. on Research and Development in Information Retrieval,Philadelphia,US. (1997) 267–276
4. Chua, T.S., Low, W.C., Chu, C.X.: Relevance feedback techniques for color-based image retrieval. In: Proc. MMM. (1998) 24
5. Rui, Y., Huang, T.S., Mehrotra, S.: Content-based image retrieval with relevance feedback in mars. In: Proc. ICIP. (1997) 815–818
6. Aksoy, S., Haralick, R.M., Cheikh, F.A., Gabbouj, M.: A weighted distance approach to relevance feedback. In: Proc. ICPR. (2000) 4812
7. Gey, F.: Inferring probability of relevance using the method of logistic regression. In: ACM SIGIR Conf. on Research and Development in Information Retrieval,Dublin,Ireland. (1994) 222–231
8. Smith, J.R., Jaimes, A., Lin, C.Y., Naphade, M., Natsev, A.P., Tseng, B.: Interactive search fusion methods for video database retrieval. In: Proc. ICIP. (2003) 741–744
9. Bustos, B., Keim, D.A., Saupe, D., Schreck, T., Vranic, D.V.: Using entropy impurity for improved 3d object similarity search. In: Proc. ICME. (2004) 1303–1306
10. Kriegel, H.P., Kröger, P., Kunath, P., Pryakhin, A.: Effective similarity search in multimedia databases using multiple representations. In: Proc. MMM. (2006) 389–393
11. Naphade, M., Wang, R., Huang, T.: Multimodal pattern matching for audio-visual query and retrieval. In: Proc. SPIE. (2001) 188–195
12. Gill, P.E., Murray, W.: Algorithms for the solution of the nonlinear least-squares problem. SIAM J. Num. Anal. **15**(5) (1978) 977–992
13. Tolonen, T., Karjalainen, M.: A computationally efficient multipitch analysis model. IEEE TSAP **8**(6) (2000) 708–716
14. Brecheisen, S., Kriegel, H.P., Kunath, P., Pryakhin, A.: Hierarchical genre classification for large music collections. In: ICME. (2006) 1385–1388