A METHODOLOGY FOR THE SEGMENTATION AND IDENTIFICATION OF MUSIC WORKS

Riccardo Miotto and Nicola Orio
University of Padova
Department of Information Engineering

ABSTRACT

The identification of unknown recordings is a challenging problem that has several applications. In this paper, we focus on the identification of alternative releases of a given music work. To this end, a statistical model of the possible performances of a given score is built from the recording of a single performance. The methodology is based on the automatic segmentation of audio recordings, exploiting a technique that has been proposed for text segmentation. The segmentation is followed by the automatic extraction of a set of relevant audio features from each segment. Identification is then carried out using an application of hidden Markov models. The approach has been tested with a collection of orchestral music, showing good results in the identification of acoustic performances.

1 INTRODUCTION

The automatic identification of music works has a number of applications, that range from digital right management, to automatic metadata extraction, and to music access and retrieval. Given the amount of music recordings that are continuously released, manual identification of music works is an unfeasible task.

A common approach to music identification is to extract, directly from a recording in digital format, its *audio fingerprint*, which is a unique set of features that allows for the identification of digital copies even in presence of noise, distortion, and compression. It can be seen as a content-based signature that summarizes an audio recording. A comprehensive tutorial about audio fingerprinting techniques and applications can be found in [3]. Audio fingerprinting systems are normally designed to identify at the same time the music score, which is the symbolic notation of the music events, and the particular recording of a performance, which is an audio signal as captured by one or more microphones [15]. On the other hand, the identification of a music work may be carried out also without linking the process to a particular performance. There are some cases where this approach may be required. Music identification of broadcasted live performances may not benefit from the fingerprints of other performances, because most of the acoustic parameters may be different. In the case of classical music, the same works may have hundreds of different recordings, and it is not feasible to collect all of them to create a different fingerprint for each recording.

An alternative approach to music identification is *audio watermarking*. In this case, research on psychoacoustics is exploited to embed an arbitrary message, the watermark, in a digital recording without altering the human perception of the sound [2]. The message can provide metadata about the recording (such as title, author, performers), the copyright owner, and the user that purchases the digital item [6]. Similarly to fingerprints, audio watermarks should be robust to distortions, additional noise, A/D and D/A conversions, and compressions. On the other hand, watermarking techniques require that the message is embedded in the recording before its distribution, a situation that can be applied only on newly released material.

This paper reports a novel methodology for automatic identification of music works from the recording of a performance, yet independently from the particular performance. Unknown music works are identified through a collection of indexed audio recordings, ideally stored in a music digital library. The approach can be considered a generalization of audio fingerprinting, because the relevant features used for identification are not linked to a particular performance of a music work. This work extends previous work on music identification based on audio to score matching [11], where performances were modeled starting from the corresponding music scores. Also in this case, identification is based on hidden Markov models (HMMs). The application scenario is the automatic labeling of performances of tonal Western music through a match with pre-labeled recordings that are already part of an incremental music collection. Audio to audio matching has been proposed in [9, 5] for classical music audio to audio matching and audio to audio alignment respectively, and in [7] for pop music.

2 HIGH LEVEL DESCRIPTION OF MUSIC PERFORMANCES

The identification of music performances is based on a *audio to audio* matching process, which goal is to retrieve all the audio recordings from a database that represents the same musical content as the audio query. This is typically the case when the same piece of music is available in several interpretations and arrangements.

The basic idea of the proposed approach is that, even if two different performances of the same music work may dramatically differ in terms of acoustic features, it is still possible to generalize the music content of a recording to model the acoustic features of other, alternative, performances of the same music work. A recording can thus be used to statistically model other recordings, providing that they are all performed from the same score.

With the aim of creating a statistical model of the score directly from the analysis of a performance, the proposed methodology is based on a number of different steps. In a first step, *segmentation* extracts audio subsequences that have a coherent acoustic content. Audio segments are likely to be correlated to stable parts in a music score, where there is no change in the number of different voices in a polyphony. Coherent segments of audio are analyzed through a second step, called *parameter extraction*, which aims at computing a set of acoustic parameters that are general enough to match different performances of the same music work. In a final step described in Section 3, *modeling*, a HMM is automatically built from segmentation and parametrization to model music production as a stochastic process. At matching time, an unknown recording of a performance is preprocessed in order to extract the features modeled by the HMMs. All the models are ranked according to the probability of having generated the acoustic features of the unknown performance.

2.1 Segmentation of the Audio Signal

The audio recording of a performance is a continuous flow of acoustic features, which depends on the characteristics of the music notes – pitch, amplitude, and timbre – that vary with time according to the music score and to the choices of the musicians. In order to be structured, the audio information has to undergo a *segmentation* process. According to [1], the word segmentation can have two different meanings: one is related to musicology and is normally used in symbolic music processing, whereas the other one follows the signal processing point of view and it is used when dealing with acoustic signals. This second aspect of segmentation is the one addressed in this paper.

In this case, the aim of segmentation is to divide a musical signal into subsequences that are bounded by the presence of music events. An event, in this context, occurs whenever the current pattern of a musical piece is modified. Such modifications can be due to one or more new notes being played or stopped. This approach to segmentation is motivated by the central role that pitch plays in music language. In fact the segmentation of the acoustic flow can be considered the process of highlighting audio excerpts with a stable pitch.

The first step of the approach is based on the computation of the *similarity* of the audio frames. This is computed as the cosine of the angle between the frequency representations of two audio frames. Thus, given X and Y the Fourier transforms of two frames:

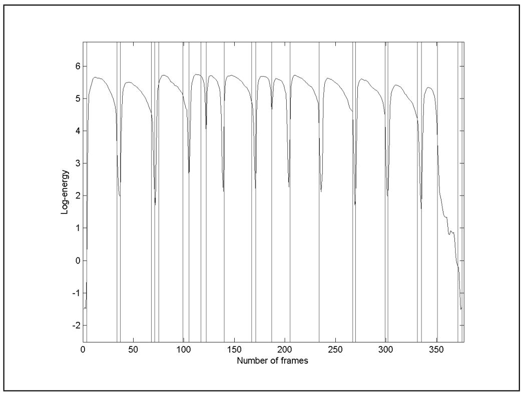

Figure 1. Example of segmentation of a monophonic audio recording, represented by its energy envelope

$$sim(X,Y) = \frac{X \cdot Y}{|X| \cdot |Y|} \qquad (1)$$

High correlation is expected between frames where the same notes are playing, while a drop in correlation between two subsequent frames is related to a change in the active notes. Similarity between different parts of an audio recording can be represented with a symmetric matrix where high values of the elements correspond to high similarity.

Pure similarity values based on correlation may not be completely reliable for a segmentation task, as it has been shown for text segmentation, because changes in the local correlation could be more relevant to its absolute value. For this reason, segmentation has been carried out according to the methodology proposed in [4] for text segmentation. The basic idea is that, in non-parametric statistical analysis, one compares the rank of data sets when qualitative behavior is similar but the absolute quantities are unreliable. Thus, for each couple of frames $\{X, Y\}$ that represents an element of the similarity matrix, the similarity value is substituted by its *rank*, which is defined as the number of neighbors elements which similarity is less than $sim(X, Y)$. That is

$$r(X,Y) = ||\{A,B\}|| : sim(A,B) < sim(X,Y) \qquad (2)$$

where matrix elements $\{A, B\}$ represent the neighbor elements of element $\{X, Y\}$ and the operator $|| \cdot ||$ computes the number of elements.

Once the rank is computed for each couple of frames, hierarchical clustering on the similarity matrix is exploited to segment a sequence of features in coherent passages. The clustering step computes the location of boundaries using Reynar's maximization algorithm [14], a method to find the segmentation that maximizes the inside density of the segments. A preliminary analysis of the segmentation step allowed us to set a threshold for the optimal termination of the hierarchical clustering. It is interesting to note that it is possible to tune the termination of hierarchical clustering, in order to obtain different levels

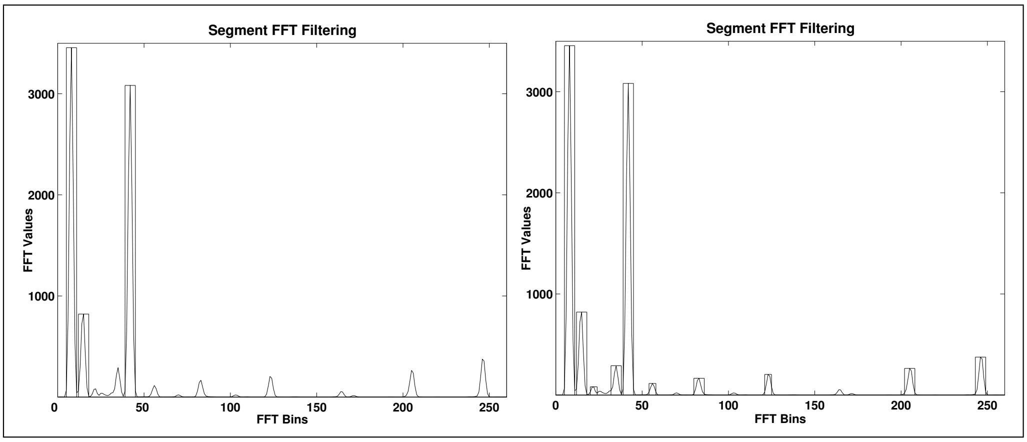

Figure 2. Parameters extraction considering 70% (left) and 95% (right) of the overall energy

of cluster granularity, for instance at note level or according to different sources or audio classes. Figure 1 depicts the computed segments over the energy trend of an audio recording

2.2 Feature Extraction from Segments

In order to obtain a general representation of an acoustic performance, each segment needs to be described by a compact set of features that are automatically extracted. In line with the approach to segmentation, also parameter extraction is based on the idea that pitch information is the most relevant for a music identification task. Because pitch is related to the presence of peaks in the frequency representation of an audio frame, the parameter extraction step is based on the computation of local maxima in the Fourier transform of each segment, averaged over all the frames in the segment.

The positions of local maxima are likely to be related to the positions along the frequency axis of fundamental frequency and the first harmonics of the notes that are played in each frame. In principle it could be expected that all the different performances of a given music work will have similar spectra. Yet this assumption does not hold for real performances, because of differences in performing styles, timbre, room acoustics, recording equipment, and audio post processing. A more general assumption is that alternative performances will have at least similar local maxima in the frequency representations, that is the dominant pitches will be in close positions.

When comparing the local maxima of the frequency representation, it has to be considered that Fourier analysis is biased by the windowing of a signal, which depends on the type and of the length of the window. These effects are expected both on the reference performances and on the performance to be recognized. Moreover, small variances on the peaks positions are likely to appear between different performances of the same music work, because of imprecise tuning and different reference frequency. For these reasons, the features are computed by averaging the FFT values of all the frames in a segment, by selecting the positions of the local maxima, and by associating to each maximum a frequency interval with the size of a quarter tone. Figure 2 exemplifies the approach: the light lines depict the average FFT of a segment, while the darker rectangles show the selected intervals.

The number of intervals is computed automatically, by requiring that the sum of the energy components that fall within the selected intervals is above a given threshold. Figure 2 depicts two possible sets of relevant intervals, depending on the percentage of the overall energy required: 70% on the left and 95% on the right. It can be noted that a small threshold may exclude some of the peaks, which are thus not used as content descriptors.

Feature extraction of the performance to be recognized is carried out by computing, for the set of intervals of each segment, the amount of energy that falls within the frequency intervals. Thus feature extraction for the unknown performance is driven by the feature extraction of the performances in the database, that is the approach is based on the expected distribution of the energy along the frequency axis.

3 PERFORMANCE MODELING AND IDENTIFICATION

Each music work is modeled by a hidden Markov model, which parameters are computed from an indexed performance. HMMs are stochastic finite-state automata, where transitions between states are ruled by probability functions [12]. At each transition, the new state emits a random vector with a given probability density function. A HMM λ, made of a set of N states $Q = \{q_1, \dots, q_N\}$, is completely defined by: a probability distribution for state transitions, that is the probability to go from state q_i to state q_j; a probability distribution for observations, that is the probability to observe the features r when in state q_j.

Music works can be modeled with a HMM providing that states are labeled with events in the audio recording, transitions model the temporal evolution of the audio recording, and observations are related to the audio features previously extracted that help distinguishing different events. The model is hidden because only the audio features can be observed and it is Markovian because transitions and observations are assumed to depend only on the actual state.

The number of states in the model is proportional to the number of segments in the performance. In particular, experiments have been carried out using a fixed number of n states for each segment, where states can either perform a self-transition or a forward-transitions. As described in [13], if all the states in a given segment have the same self-transition probability p, the probability of having a given segment duration is a negative binomial. Once chosen the value of n, it is possible to compute p on the basis of the duration of the segments, in order to statistically model the expected duration of the events of the performance to be recognized.

A preliminary evaluation with synthetic performances where durations have been artifically modified, showed that this modeling is robust to large timing variations between the performances used to build the models and the performances to be recognized. Identification rate was not substantially affected even when tempo was twice as fast or twice as slow.

Figure 3 depicts an excerpt of an HMM, representing two states and their transition probabilities. Each state in the HMM is labeled to a given segment and, accordingly with the parameter extraction step, emits the probability that a relevant fraction of the overall energy is carried by the frequency intervals computed at the previous step. The modeling of emission probabilities build upon an approach to score following and alignment that has been presented in [10] using dynamic time warping and is similar to the one presented in [11] using HMM and, in fact, one of the goals of this work was to create a common framework where an unknown performance could be recognized from either its score or an alternative performance.

3.1 Identification

Recognition, or identification, is probably the most common application of HMMs. The identification problem may be stated as follows:

> given an unknown audio recording, described by a sequence of features $R = \{r(1), \cdots, r(T)\}$ and given a set of competing models λ_i: *find the model λ that more likely generated R*

The most common approach to HMM-based identification, is to compute the probability that λ_i generates R regardless of the state sequence. This can be expressed by equation

$$\lambda = \arg\max_i P(R|\lambda_i) \qquad (3)$$

where the conditional probability is computed over all the possible state sequences of a model. The probability can be computed efficiently using the *forward probabilities*.

Even if this approach is the common practise for speech and gesture recognition, it may be argued that also paths that have no relationship with the actual performance give a positive contribution to the final probability. For instance, a possible path, which contributes to the overall computation of the forward probabilities, may consist in the first state of the HMM that continuously performs self-transitions. These considerations motivated the testing of two additional approaches, with the aim of taking into account only the optimal path that aligns the two performances. Preliminary tests showed that the classical approach based on the forward probabilities outperforms the other approaches that take into account either the global or the local optimal alignment. The results are not reported in this paper and can be found in [8].

3.2 Computational Complexity

It is known in the literature that the computation of the forward probabilities requires $\mathbf{O}(DTN^2)$ time, where D is the number of competing models, T is the duration of the audio sequence in analysis frames, and N is the average number of states of the competing HMMs. Considering that, as described in Section 3, each state may perform a maximum of two transitions, it can be shown that complexity becomes $\mathbf{O}(DTN)$. In order to increase efficiency, the length of the unknown sequence should be small, that is the method should give good results also with short audio excerpts.

An important parameter for computational complexity is the number of states N. A first approach to reduce N is to compute a coarse segmentations, which corresponds to a smaller number of group of states. On the other hand, a coarse segmentation may give poor results in terms of emission probabilities, because a single segment could represent parts of the performance with a low internal coherence. Another approach to reduce the computational complexity is to use a small number of states n for each segment, and model the durations with higher values of the self-transition probabilities p. As previously mentioned, in our experiments we found that setting $n = 4$ for each segment gave a good compromise.

4 EXPERIMENTAL EVALUATION

The methodology has been evaluated with real acoustic data from original recordings taken from the personal collection of the authors. Tonal Western music repertoire has been used as a test-bed because it is common practice that musicians interpret a music work without altering pitch information, which is the main feature used for identification. The audio performances used to create the models were 206 incipits of orchestral works of well known composers of Baroque, Classical, and Romantic periods. All the incipits used for the modeling had a fixed length of 10 seconds. The audio files were all polyphonic recordings, with a sampling rate of 44.1 kHz, and they have been divided in frames of 2048 samples, applying a hamming window, with an overlap of 1024 samples. With these parameters, a new observation is computed every 23.2 mil-

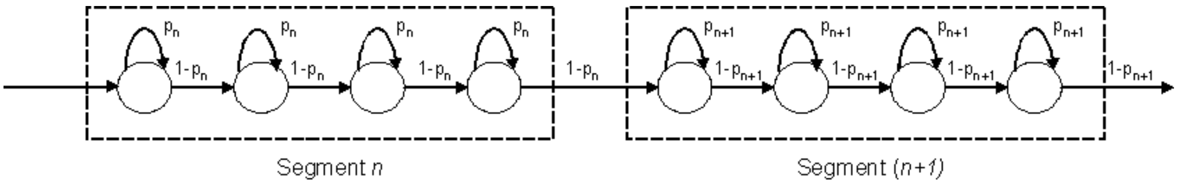

Figure 3. Graphical representation of HMM corresponding to two general segments.

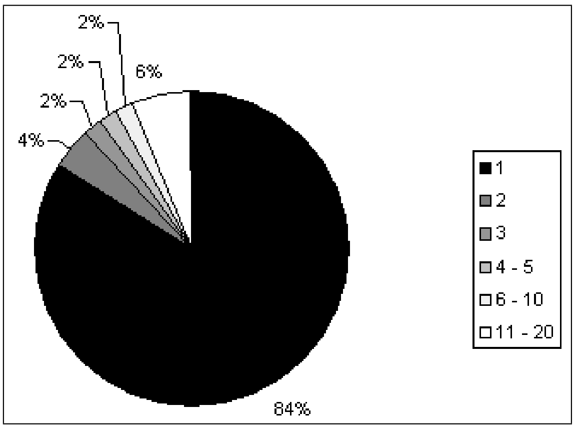

Figure 4. Rank distributions of correct matches

liseconds.

The recordings to be recognized were 50 different performances of a subset of the music works used to build the models. Also in this case they were the incipit of the music works, with a length of 8 seconds. The goal was to have a high likelihood that each unknown performances was included in the corresponding performance in the database, even in the case where the two performances had a different tempo. All the other parameters of the audio files were the same. The 50 audio excerpts have been considered as unknown sequences to be identified, using the approach presented in Section 3.1. Figure 4 shows the percentages at which the correct audio recording was ranked as the most similar one, and when it was ranked within the first two, three, five, ten and twenty positions. As it can be seen, 42 out of 50 queries (84%) were correctly identified, while 45 queries (90%) returned a correct match among top 3 models. Moreover, only 3 queries (6%) returned the correct match after the first 10 positions and none after the first 20 positions. The Mean Reciprocal Rank (MRR) for all the 50 recordings was 87.78.

4.1 Effects of Lossy Compression

In order to test the robustness of the methodology, we applied a lossy compression algorithm to the experimental setup. In particular, the compression has been applied only to the audio excerpts to be recognized, simulating a real situation where the elements of the database are of high quality while there is no control about the quality of the excerpts submitted by the users.

We compressed the performances using MP3 encoding, at three different bitrates: 32, 64 and 128 kbps. In this way we could compare different levels of quality, from two poor bitrates up to a common one among the digital music that travels through the Web, which is usually considered of satisfactory quality. The results are shown in Table 1, which reports the percentage of performances that have been ranked within different thresholds together with the MRR. As it can be seen, the approach is robust to lossy compression, because the three bitrates gave almost the same results, with a decrease in effectiveness of about 2% compared to the results without compression.

Compression	**32 kbps**	**64 kbps**	**128 kbps**
= 1	82	82	82
≤ 3	88	88	88
≤ 5	90	90	90
≤ 10	92	94	94
≤ 20	98	100	100
MRR	85.86	85.91	85.95

Table 1. Identification rates in presence of lossy-compression, in terms of percentage of being ranked within given thresholds and of Mean Reciprocal Rank

4.2 Robustness to Additional Noise

In the last test, we verified the robustness of the algorithm with noisy recordings. To this end, we added a component of white noise to the audio elements to be recognized. Also in this case, we figured out a situation in which the database contains high quality elements whereas the unknown audio excerpts could be disturbed or damaged. A pink noise has been added with an energy of -12, -18, and -24 dB in respect to the maximum peak of the signal of the query, which was set to 0 dB. Clearly, high values of noise are not realistic, because at least for -12 dB, some parts of the query were almost inaudible. On the other hand, a noise of -24 dB seemed to be a good approximation of the typical noise of old analog tapes.

The results are shown in Table 2, which reports the percentage of performances that have been ranked within different thresholds together with the MRR. The results show that the recognition rate is sensible to the presence of additional noise, even if it is unlikely that such poor quality recordings will be of interest for the end user. The results could probably be improved by applying some noise removal tool to corrupted recordings. Yet, in order to be meaningful, this combined approach has to be applied to real noisy recordings, which will be collected and added to the test collection in the future.

Noise level	**-12 dB**	**-18 dB**	**-24 dB**
$= \mathbf{1}$	32	54	64
$\leq \mathbf{3}$	46	66	74
$\leq \mathbf{5}$	52	72	84
$\leq \mathbf{10}$	64	84	88
$\leq \mathbf{20}$	74	90	96
MRR	42.62	62.90	71.92

Table 2. Identification results in presence of white noise, in terms of percentage of being ranked within given thresholds and of Mean Reciprocal Rank

5 CONCLUSIONS

A methodology for automatic music identification based on HMMs has been proposed. The methodology has been tested on a collection of digital acoustic performances. Experimental results showed that, at least for tonal Western music, it is possible to achieve a good identification rate that was about 84% with the optimal configuration of the parameters.

These results suggest that the approach can be successfully exploited for a retrieval task, where the user queries the system through an acoustic recording of a music work. The automatic identification of unknown recordings can be exploited as a tool for supervised manual labeling: the user is presented with a ranked list of candidate music works, from which he can choose the correct one. In this way, the task can be carried out also by non expert users, because they will be able to directly compare the recordings of the unknown and of the reference performances through direct listening. Once that the unknown recording has been correctly recognized, it can be indexed and joint to the musical digital library, allowing us to increment the information stored inside it.

Future works will involve the extension to other music genres, in particular pop and rock music, for which preliminary results on a small collection have been already obtained. Current works regard the realization of a distributed prototype, with the aims of increasing the scalability of the approach.

6 REFERENCES

[1] Aucouturier, J. *Segmentation of Music Signals and Applications to the Analysis of Musical Structure*. Master Thesis, King's College, University of London, UK, 2001.

[2] Boney, L., Tewfik, A., Hamdy, K. ”Digital watermarks for audio signals”, *IEEE Proceedings Multimedia*, 473–480, 1996.

[3] Cano, P., Batlle, E., Kalker, T., Haitsma, J. ”A review of audio fingerprinting”, *Journal of VLSI Signal Processing*, **41**, 271–284, 2005.

[4] Choi, F. ”Advances in domain independent linear text segmentation”, *Proceedings of the Conference on North American chapter of the Association for Computational Linguistics*, 26–33, 2000.

[5] Dixon, S., Widmer, G. ”MATCH: a music alignment tool chest”, *Proceedings of the International Conference of Music Information Retrieval*, 492–497, 2005.

[6] Haitsma, J., van der Veen, M., Kalker, T., Bruekers, F. ”Audio watermarking for monitoring and copy protection”, *Proceedings of the ACM workshops on Multimedia*, 119–122, 2000.

[7] Hu, N., Dannenberg, R., Tzanetakis, G. ”Polyphonic audio matching and alignment for music retrieval”, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 185–188, 2003.

[8] Miotto, R., Orio, N. ”Recognition of music performances through audio matching”, *Proceedings of the Italian Research Conference on Digital Library Management Systems*, in press, 2007.

[9] Müller, M., Kurth, F., Clausen, M. ”Audio matching via chroma-based statistical features”, *Proceedings of the International Conference of Music Information Retrieval*, 288–295, 2005.

[10] Orio, N., Schwarz, D. ”Alignment of monophonic and polyphonic music to a score”, *Proceedings of the International Computer Music Conference*, 129–132, 2001.

[11] Orio, N. ”Automatic recognition of audio recordings”, *Proceedings of the Italian Research Conference on Digital Library Management Systems*, 15–20, 2006.

[12] Rabiner, L., Juang, B. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[13] Raphael, C. ”Automatic segmentation of acoustic musical signals using hidden markov models”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**, 360–370, 1999.

[14] Reynar, J. *Topic Segmentation: Algorithms and Applications*. PhD Thesis, Computer and Information Science, University of Pennsylvania, 1998.

[15] Suga, Y., Kosugi, N., Morimoto, M. ”Real-time background music monitoring based on content-based retrieval”, *Proceedings of the ACM International Conference on Multimedia*, 120–127, 2004.