Int J Multimed Info Retr (2016) 5:105–115
DOI 10.1007/s13735-016-0096-5

REGULAR PAPER

Automatic environmental sound concepts discovery for video retrieval

Issam Feki[1] · **Anis Ben Ammar**[1] · **Adel M. Alimi**[1]

Received: 12 February 2016 / Revised: 27 February 2016 / Accepted: 1 March 2016 / Published online: 14 March 2016

Abstract This paper characterizes a new method for video–soundtrack retrieval based on environmental sounds. Actually, a set of 26 semantic audio concepts is employed. This set is chosen for its helpfulness to the users in terms of video browsing. Additionally, a set of 2000 videos has been annotated with these concepts. To enhance a new signal processing, we start with the separation of the audio sources. In addition, using a fundamental representation of the audio signal as a sequence of Mel Frequency Cepstral Coefficient, we can carry out experiments with three signal representations: the Support Vector machines, the Gaussian Mixture Model and the Hidden Markov Model. Throughout the experiment synthesis, we maintain the Gaussian Mixture Model classifier based on the Kullback–Leibler distance measure. As a matter of fact, we preserve this audio concept classification to integrate a video retrieval system. Hence, the obtained results mirror the effectiveness of our approaches in distinguishing environmental sound and researching video.

Keywords Concepts · Environmental sound · Sound indexing · Sound query · Video retrieval

✉ Issam Feki
feki_issam@yahoo.fr

Anis Ben Ammar
anis.benammar.tn@ieee.org

Adel M. Alimi
adel.alimi@ieee.org

[1] REGIM: Research Group on Intelligent Machines, University of Sfax, ENIS, BP 1173, 3038 Sfax, Tunisia

1 Introduction

At the outset, the strength of a video document lies in its capability to transmit a rich semantic presentation through the audio sync, text and visual presentations over time. In the early Content-Based Video Retrieval research, most of the efforts were made to extend the systems and algorithms from images and texts. Although this research trend has certainly added a degree of success, it is not satisfactory for all the applications because video contains other components that can enrich its semantic meaning. Despite its neglect, the audio component within video remains an important source which is worth being explored and exploited in the CBVR. Therefore, the CBVR system should satisfy the various needs of the users. For instance, image-based indexing usually fails to meet the goals of the users in their search for a particular event in video such as finding the event of an explosion in a video clip. Unfortunately, in some video clips, this event is produced just through sound effects but without any visual effect. To meet most of the users' requirements, many techniques have been proposed to bridge the semantic gap between the features that can be automatically extracted to upgrade the quality of the users' queries in video retrieval. Among these techniques, we can point up the approach that makes use of the distinctive characteristics of an audio to design the most useful tools for the content extraction and indexing as well as the related browsing and retrieval. In this paper, the emphasis is laid on audio as it is an attraction for a universal audience. Moreover, such a novel use of video would establish a strong foundation for the use of the CBVR system. Compared to other video components, audio causes a number of exclusive challenges. Among these challenges, we can cite its dependence on the context such as music and particular sound effects. In addition, it is produced with diverse modes of limited effects depending on the need to

describe some effects. More importantly, it can be appealing for diverse purposes such as speech, music analysis and environmental sound. This work takes in hand these challenges using the acoustic information of the video soundtrack to spot which descriptors can be dependably extracted from this modality. If the visual information in video is deemed to be rich, audio is also a valuable and complementary source of information. This paper is organized in the following way: Sect. 2 presents an overview of video retrieval based on audio modality. Section 3 details the suggested framework for video retrieval system. The assessment and discussion of experimental results are presented in Sect. 4.

2 Overview

The previous work on audio analysis has classically focused on distinguishing between two categories such as speech, music, silence, noise, or applause. The request field has relatively produced sources, such as movie soundtracks or audio broadcasts. The authors in [1] offered a method of speech/music discrimination based on the zero-crossing rate and short time energy features and used the Gaussian classifier for radio broadcasts. This research reported an accuracy rate of 98 %. In [2], the authors used the same approach to distinguish between speech and music to classify 50 different phone sounds and achieved the same results. In [3], the authors classified 2.4 segments of radio-data broadcasts as speech and music. This research used temporal and spectral features pursue by the Gaussian mixture model (GMM) classifier. The results indicate an error rate of 1.4 %. The authors, in [4], implanted a speech/music distinguishing framework based on the hidden Markov model (HMM) classification. The entropy based on posterior probabilities of speech classes is used as features. In [5], the authors presented a system to distinguish between additional classes other than speech and music, such as song, speech with music background, environmental sound, and so on from TV programs or movies. The heuristic rule-based classifier using the zero-crossing rate, spectral peak tracks, energy and pitch features, indicates an accuracy rate of 90 %.

For environmental sounds l, the research community has examined problems such as the content-based retrieval. A well-liked audio framework is to analyze, cluster, and classify the environmental sound into concepts like applause, whistle, airplane, breeze glasses and ringtones with evaluation of a small amount of data. The authors in [6] used SVM to classify the Muscle Fish audio data classifiers with a collection of perceptual and cepstral features. This research work provides approximately the same errors rate in [7]. To show the efficiency of the linear auto encoding neural networks and the Gaussian mixture model (GMM), the authors in [8] suggest a classification of the environments such as outdoor and office. The arrangement of these two techniques showed a better performance. The authors in [9] presented research advancements in the environmental sound classification field on a handy device. They employed the MFCC classified features and the HMM speech recognition. The results showed more than 90 % accuracy for 11 environments distinguishing. A comparison of the nearest neighbor (NN), GMM, and SVM classifiers with a large collection of features on a five conduct classification assignment is made in [10]. The use of the scheme integrating SVM classifiers and a subset of features presented the best performance. In [11], the authors suggest an environmental sound identification system. They exploit the local discriminate bases method for feature extraction progression and HMM as a classifier. The experimentation results show that 21 audio concepts were distinguished, but the average recognition accuracy did not exceed 81 % for the test set. But when the scene includes more than one audio source, the averages of the accuracy decrease to 28.6 %.

The research work that goes along with the present paper is [12] in which the authors studied the recognition of 13 audio concepts such as explosion, automobile, helicopter, water, wind, and rain. By comparing different collections of features and classifiers, they could realize an excellent efficiency with an uncomplicated approach of distinguishing based on SVM and Audio Spectrum Flatness, Centroid, Spread, and Audio Harmonicity (ASFCS-H) features. The assessment gives the best performance with an average measure value of 80.6 %.

Actually, none of the previous works has immediately attended to the CBVR by their soundtracks. This research field witnessed an amount of novel issues that are dealt with for the first time in this paper. First, we are interested in 26 concepts which outnumber the concepts mentioned in [12]. Second, our concepts stem from the user's study of video retrieval. Accordingly, they replicate concrete types of queries that the users asked more willingly rather than the simple retrieval that we suppose to be apparent in the video data. Our approach is stimulated by a related work in CBVR which has a number of visual concepts [13]. Third, our data set is different from any earlier reported data in the field of environmental sounds; in other words, it is composed of 2000 soundtracks taken from the Daily Motion and the YouTube videos. These soundtracks are not considered useful in the research area as they are very rich noise. Unlike the previous works, however, the working procedure in this paper is more demanding.

Not only does this paper discuss the originality of the problem, but it also presents some new specific techniques. On the one hand, we expound the practicability of retrieval based only on the environmental sounds because the video, in its visual aspect, does not reflect sound. For example, the search for video shots containing the sound of a helicopter concept results in video with its sound and image unrelated to helicopter. On the other hand, we prove the technical solu-

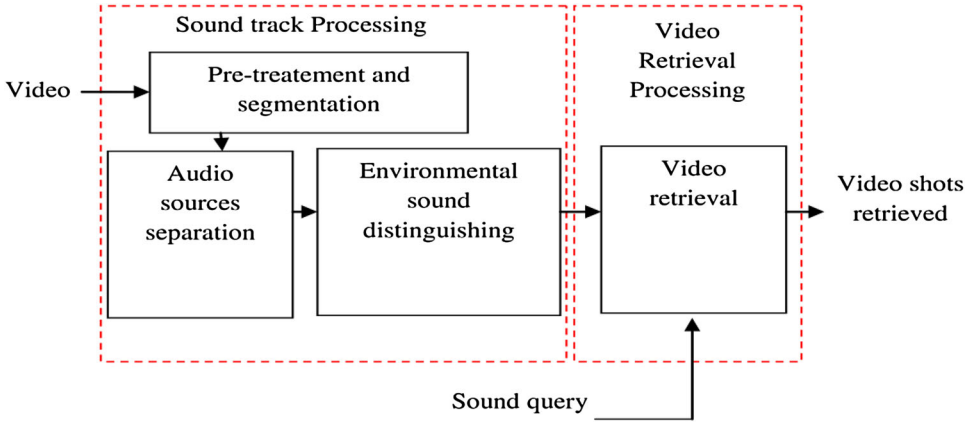

Fig. 1 System overview

tion to the problem of overlapping sounds through a new soundtrack segmentation process. Our audio concepts have a different distinctiveness in terms of the frequent needs of the user to video retrieval. For example, the users always seek the exciting moments in the video. These moments are usually expressed by the concept "explosion", "glass breaking" and "police alarm", etc.

3 Proposed system overview

In the proposed system, there are two major steps as shown in Fig. 1. The first one, named Sound track Processing (StP), supports the acoustic treatment. The second, untitled Video Retrieval Processing (VRP) handles the video shot retrieval. In the StP step, we start by extracting the soundtracks from video. The pre-treatment and segmentation module supports the removal of the silence fragments and the accomplishment of the proposed segmentation strategy. In addition, audio sources separation, as a secondary step, puts in appearance, speech, music and environmental sound signals. Then, the environmental sound distinguishing module applies the classifiers trained on the annotations of the audio concepts to categorize the environmental sound signals. Finally, VRP calculates the link between the sound query vectors and the environmental sound vectors and extracts the video shots from the corresponding audio sample.

3.1 Sound track processing

We start by constructing only one video file (F) from all the video portions and sequences in the database. This merger of files is usually recommended due to the operating principle of our system. This fusion is based on a continuous time axis. All the videos used for research are bound and spread on a single timeline. This alternative is explained by the need of audio concept localization by the system. It is noticed that the main goal is to search an audio concept in a large variety of video data. After the merger of video sequences, our system extracts automatically the sound track of the video and produces an audio stream (AS). The obtained AS can be in any audio format.

3.1.1 Pre-treatment and segmentation

It is obvious that data analysis usually requires a pre-treatment process. Just like the image, the audio track needs a different treatment ready to use in the analysis process. Some actions, like segmentation and filtering, must be imperatively executed to the audio track.

Regarding segmentation, our main goal is to discover an entity of sound sources. The elementary subject in the temporal segmentation of sound tracks is the one that precisely represents the source. This phenomenon is not only explained by the overlap of sound sources but also by the existence of silence between the different audiences. The previous work in audio segmentation has paid attention to speech [14] and music [15]. The additional approaches deal with the segmentation of continuous environmental sound signals [16–18]. A number of methods [16,17] attend to segmentation essentially in terms of the semantic context (e.g., "shopping") despite the fact that these approaches are helpful in indexing.

In this paper, the acoustic sources are a distinct and, structurally, significant entity of speech, music and environmental sounds. Therefore, noise and silence have never been considered as sound sources. For that reason, a new audio segmentation strategy is now presented to ensure the filtering of sound tracks and the separation of sound sources

• *Sound track: first segmentation and silence detection*
As shown in Fig. 2, the system makes the first segmentation for the wave AS. The sound track is segmented into frames (Si) of length (L) equal to 3 seconds with 1 second overlapping with the previous ones. Each clip is then divided into frames that are 512 samples long and are shifted by 256 samples from the previous frames. The idea behind this proposal is to detect and remove long runs of silence from the audio

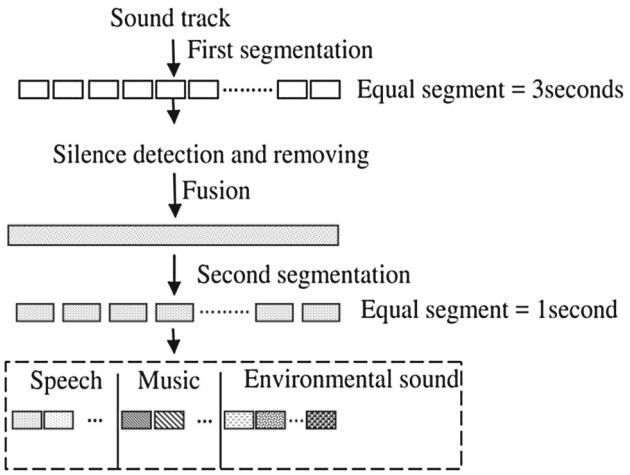

Fig. 2 The novel audio segmentation process

frame. Its significance as a construction block resides in its interest in solving two problems. The first one is lightness requirements of audio which are important; i.e., any reduction in terms of the total audio ground required by the system is a significant benefit. The second is the removal of silence which is essential in our research context; our goal is to index the audio concepts and develop a video retrieval system based on these indexing concepts. It is obvious that the users have no need for silence research in video. We use the STE (Short Time Energy) feature to detect silence [19]. The energy E of a distinct time signal $x(n)$ is computed by

$$E = \sum_{n=-\infty}^{\infty} x^2(n) \tag{1}$$

For lots of audio signals, such a measurement is less significant since it provides modest information about time-dependent characteristics of such signals. The amplitude of an audio signal shows a discrepancy with time. A suitable illustration that reproduces these amplitude variations is the STE of the signal. It is defined as follows:

$$E_m = \sum_{n} x(n)^2 h(m-n) \tag{2}$$

where $h(m) = w2(m)$. In the above expression $h(m)$ is deduced as the impulse response of a linear filter. The character of STE representation is determined by the alternative of the impulse response, $h(m)$. If the STE function is continuously lower than a certain set of thresholds (there may be durations in which the energy is higher than the threshold, but the durations should be short enough and far apart from each other), the segment is indexed as silence. Silence frames will be removed from the audio segments. For the remaining frame, we extract Short Time Energy (STE) value, the set that will be used to determine the silence frames. These frames are automatically eliminated because they are not part of our concepts.

• *Sound track: second segmentation*
After removing silence from all the frames, our system merges them and produces a new sound track. This audio tape is characteristically without silence. The system makes a second segmentation for AS. The sound track is segmented into frames (Si) of length (L) equal to 1 s with 0.5 second overlapping with the previous ones. Each clip is then divided into frames that are 512 samples long and are then shifted by 256 samples from the previous frames. Our audio segmentation strategy allows, at the same time, the correct filtering of the soundtrack and a beneficial specific treatment for the separation of sound sources. Although the source separation is not our focal point, our method does yield a rough idea of where a certain sound source can be distinguished [20].

3.1.2 Audio sources separation

Contrary to the previous studies, this paper aims to distinguish the appropriate class of environmental sounds. For a robust discrimination of these sound categories, we propose the following design using Low Short Time Energy Ratio (LSTER), Spectrum Flux (SF) and Band Periodicity (BP) features [21].

As shown in Fig. 3, a two-step scheme is proposed to classify audio clips into one of the three audio classes: speech, music, and environment sound.

• *Environmental sound distinguishing*
Our objective is to provide the users with the pertinent classification to browse their personal video collections; accordingly our concepts should reproduce their requests. The selected concepts are correctly labeled by humans as shown in Table 1. The majority of the concepts are essentially acoustic, although several concepts, such as explosion and helicopter, are first and foremost visual.

Our elementary frame-level feature is the MFCC regularly used in speech identification and additional auditory classification. The soundtrack of a video is first passed to 8 kHz. After that, a short time Fourier scale spectrum is calculated over 25-ms windows every 10 ms. The spectrum of every window is distorted by the Mel frequency scale, and the log of these acoustic spectra is intended for MFCCs by means of a discrete cosine transform.

Subsequent to the preliminary MFCC analysis, all soundtracks are represented as a set of 21-dimensional MFCC feature vectors. We experiment with more than one technique: the Support Vector Machines (SVM), the Gaussian mixture modeling (GMM), and the Hidden Markov Model (HMM). Each of them argued in more detail below. These representations are at that time measured up to one another by a number of distance measures: the cosine similarity [20], the

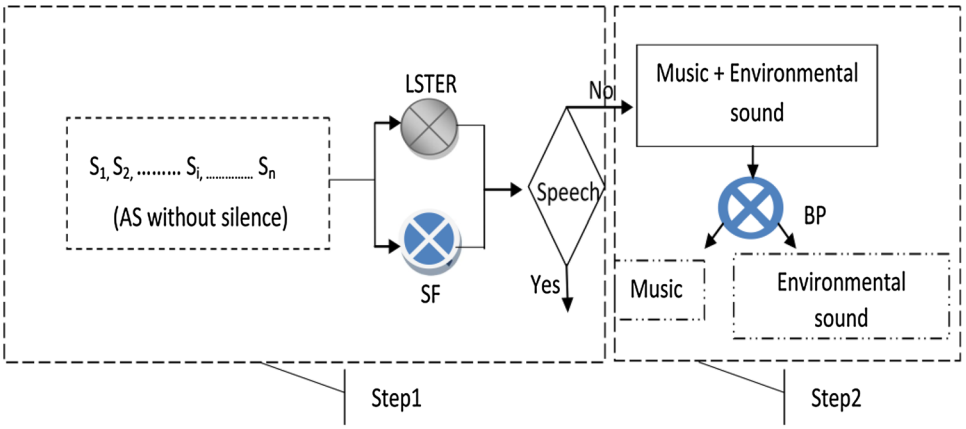

Fig. 3 Audio sources separation design

Table 1 Counts of manually labeled examples of each concept from 2000 videos

Concept	Designation	Examples
Ringtones	Phone	110
Train	The sound of the train horn	77
Motorcycle	The sound of the motorcycle engine	74
Explosion	Sudden increase	64
Helicopter	Airliner	59
Slamming door	Doors that close	71
Dog barking	Animal	88
Bird singing	Aves	68
Glass breaking	Glass giveway	92
Applause	One or more people applauding	51
Horse	The sound of a horse walking	49
Cat	Animal	66
Car	Engine sound of a car	118
Slot machine	Casino machines sound	63
Wind	The wind blowing	73
Plane	An aircraft in flight	89
Laugh	One or more people laughing	52
Police alarm	Police car	111
Whistle	Hiss, pipe, blow, boo	57
Car braking	Sudden and sharp braking car	119
Draws fire	Gun shot	327
Wolf	Animal	43
Fight	Kicking, slap	413
Rain	It is raining	95
Bell	Bell watches	74
Coin	Shekel	87
Total	26 concepts	2000

Kullback–Leibler divergence [22] and the Euclidian distance [23].

• *Support Vector Machines*
SVM is a two-class classifier constructed from sums of a kernel function K(.,.),

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b \tag{3}$$

where x is the vector (environmental sound signal) needed to classify and xi are the support vectors obtained from the training sets by an optimization process, yi is either 1 or −1 depending on the corresponding support vector belonging to class 0 or class 1,

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{4}$$

α_i $(i = 1, 2, \ldots, N)$ are constant terms and $\alpha_i > 0$.

A single SVM solves only two-class discrimination problems. For a multi-class discrimination, the one-against others strategy is usually used. It needs to train SVM for each class and we use a nonlinear support vector classifier to discriminate the various categories [24]. The classification parameters are calculated through the support vector machine learning.

• *Gaussian mixture model*
The mixture models are a kind of density model. They include a set of, usually Gaussian, component functions. These constituent functions are jointly used to provide a multi-modal density. A Gaussian mixture density is a biased sum of M component densities and is specified by the form:

$$P\left(\frac{x}{\lambda}\right) = \sum_{i=1}^{M} C_i b_i(x) \tag{5}$$

where x is a dimensional random vector, bi(x); $i = 1, \ldots, M$ is the component density and ci; $i = 1, \ldots, M$ is the mixture weight. Each density component is a varied Gaussian function of the form:

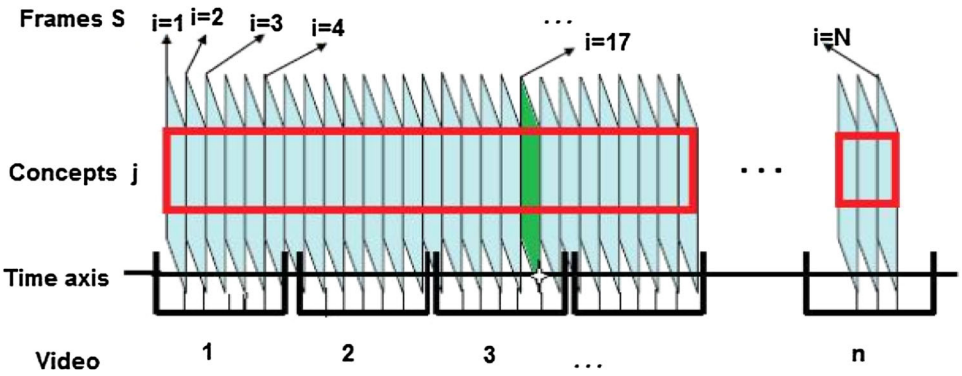

Fig. 4 Video shots research

$$b_i = \frac{\exp\left\{-1/2\,(x-u_i)^2\,(\sum_i)^{-1}\,(x-u_i)\right\}}{(2\pi)^{1/2}\left|\sum_i\right|^{1/2}} \tag{6}$$

with the mean vector u_i and the covariance matrix $\sum_i$. The mixture weights assure the restraint:

$$\sum_{i=1}^{M} C_i = 1 \tag{7}$$

The complete Gaussian Mixture density is parameterized by the signify vectors, the covariance matrices and the mixture weights from the entire constituent densities. These parameters are expressed together by the notation: $\lambda = \left\{C_i, U_i, \sum_i\right\}$ $i = 1, \ldots, M$. Each environmental sound signal is expressed by GMM and is referred to as a class. The GMM parameters are expected to use the standard Expectation Maximization algorithm. Then, by means of the logarithms and the independence between the observations, the feature identification system normalizes and calculates $p\left(\frac{x}{\lambda}\right)$ to create feature recognition decisions [25].

3.1.3 Hidden Markov model

A discrete HMM is determined by three groups of parameters: the state transition probability

$$A = \{\text{aij}\}, \text{ where } \{\text{aij}\} = p\,(\text{qt} + 1 = j|\text{qt} = i) \tag{8}$$

The observation symbol probability

$$B = \{\text{bj}, (k)\}, \quad \text{where bj}\,(k) = p\,(\text{ot} = \text{vk}|qt = j) \tag{9}$$

and the initial state distribution $\pi = \{\pi i\}$, where $\pi i = P(ql = i)$. Here, qt is the state at time t, vk is the distinct observation symbols in the observation space and ot is the observation vector in time t. For convenience, we use $\lambda = (A, B, \pi)$ to indicate the model parameters. In our case, the observation space is the feature space and we need to quantize it to a finite number of vectors before we use the discrete HMM. Here, we generate the codebook using a binary split algorithm described in [25]. HMM has been successfully applied in several large-scale laboratories and commercial speech recognition systems. In a traditional speech recognition system, a distinct HMM is trained for each word or phoneme, and the observation vector is computed in every frame (10–30 MS).

3.2 Video retrieval processing

In fact, the audio modality is a very suitable and efficient technique as it allows for a better discrimination between the concepts. Nevertheless, the retrieval results will be in the form of a whole sequence involving all the components. However, we propose a scheme to solve the indexing problem formulated on both modalities.

Since the audio stream segments are sequential i, in Si, i represents the essential part to get video shots correspondence. Therefore, a simple calculation is used to keep the retrieved shot. Since the segmentation length is L, $i \times L$ refers to the time of the segment whose order is i. As the audio time is synchronous with the input stream, the outcome $i \times L$ is the time location of Si in F. As shown in Fig. 4, the label of the frame S (results of query) is $i = 17$. Remember that the frame length is $L = 1$; consequently, the 17th visual frame is the start of the video shot playing the query concept sound. The main challenge facing this work is how to get the fundamental structure of a video despite the doubt caused by the temporal dissimilarity. Based on the sound extracted from a given database, the process presented here looks for a distinctive acoustic segment. The significance of these audio segments corresponds to the concepts in the semantic space. Therefore, the resulting similarity is a proximity measure between the sound query signal and the video shots.

4 Experiments and results

We evaluate our system in terms of the Average Precision (AP) for the distinction of the 26 environmental sounds across the 2000 videos. Our testing strategy is based on two comple-

mentary levels; the environmental sounds classification level and the video shots research level.

4.1 Environmental sounds classification experiments

We assess our approaches using threefold cross-validation: SVM, GMM and HMM. Every fold is trained for 40 % of the signal, and, then, tested on the remaining 60 %. Figure 5 shows the results of the SVM with the three different distance measures; namely, the cosine similarity, the Euclidian distance, and the Kullback–Leibler divergence.

SVM+ Kullback–Leibler divergence gives a better performance for the audio concepts classification such as "Ringtones", "Explosion" and "Police Alarm"; by distinction, the concepts such as "Car braking" and "Rain" are the best with the SVM+cosine similarity. The Concept "Laugh" is well detected by SVM+Euclidian distance, probably, for the reason that the human audio source takes an important part in discriminating other concepts. On average, SVM+ Kullback–Leibler divergence is the greatest among the three distance measures.

Figure 6 shows the results of the HMM with the same distance measures. The most favorable results of HMM is powerfully reliant on the total length of the positive examples of the concept; the HMM+Kullback–Leibler divergence gives and takes the greatest Average Precision and is capable to detain the feature across all the concepts.

The performance of the GMM+Kullback–Leibler divergence is shown in Fig. 7. To build the surrounding substance for this classification set, we experience the remaining distance measures. In contrast, in the three curves of the GMM histograms, we see that GMM+Kullback–Leibler complete are considerably superior to GMM+cosine similarity and GMM+Euclidian distance.

• *Discussion*

Figure 8 compares the most excellent results for each of the three modeling approaches (SVM+ Kullback–Leibler, HMM+Kullback–Leibler, GMM+Kullback–Leibler).

The figure compares the presentations in terms of Average Precision and accuracy rate. The Average Precision is computed one by one for each correct clip. The accuracy rate is the amount of the true clips. Noticeably, accuracies can keep the concepts high values with the little former probabilities purely by regarding all the clips as unconstructive. Equally, Average Precision is not deemed as a momentous part of this partiality. To attain a rigid classification from our

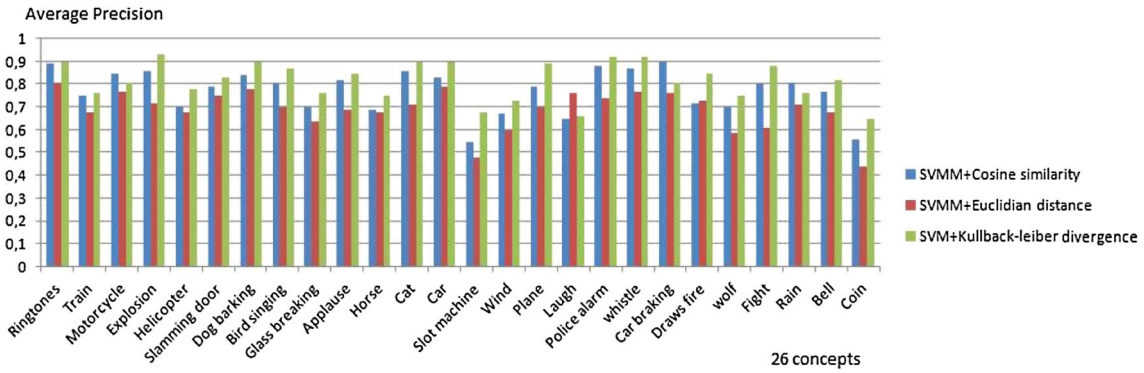

Fig. 5 AP across all 26 concepts for the SVM, using each of the three distance measures, cosine similarity, Euclidian distance, and Kullback–Leibler divergence

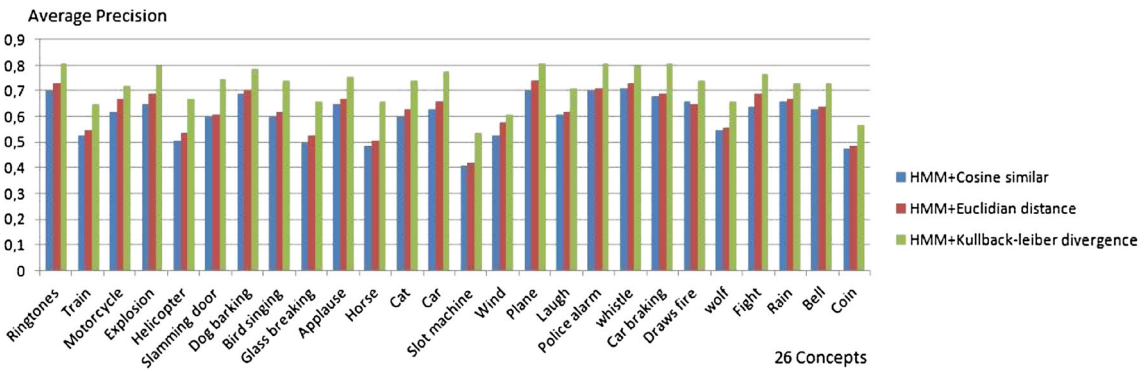

Fig. 6 AP across all 26 concepts for the HMM, using each of the three distance measures, cosine similarity, Euclidian distance, and Kullback–Leibler divergence

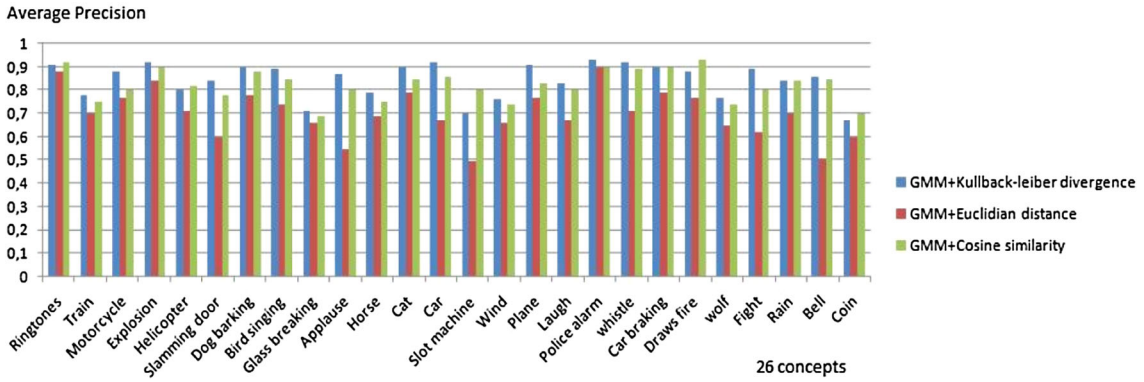

Fig. 7 AP across all 26 concepts for the GMM, using each of the three distance measures, cosine similarity, Euclidian distance, and Kullback–Leibler divergence

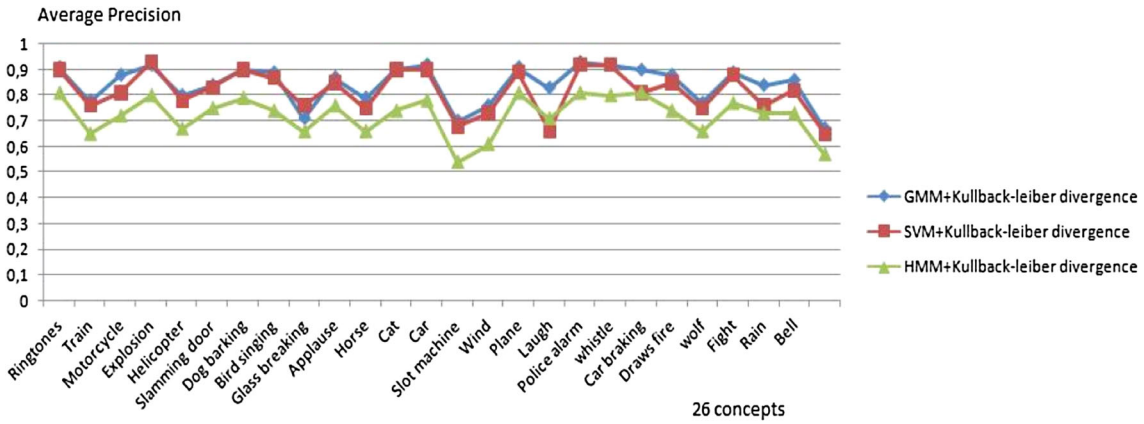

Fig. 8 Best results from Figs. 5, 6, and 7 illustrating the relative performance of each presentation

GMM-based rankings, we need to decide an entry for the distance-to-boundary values. Accordingly, we put this entry in competition with every class; in other words, the number of positive classifications corresponds to the previous class.

In addition, we notice that there is little variation by the use of this concept. It can be projected provided that the diverse labels will be obvious in the soundtrack and maintained by an extensive quantity of data training. Like the former possibilities, the major determining factor of the performance of these classifiers signifies that the larger the number of data training is, the harder the classifier will be. This is, still, bewildering. In some cases, these aspects may be illustrated; for instance, the concept "explosion" has the same Average Precision as that of the other concept "helicopter".

A number of concepts consist of a distinctive minority; an audio representative may be modeled by SVM rather than by GMM. We notice that the Average Precision for the "car" concept is visibly better with the SVM than with GMM. This also proposes that the presentation could be enhanced by separating some classes into additional subclasses (e.g., "car" possibly will be refined to "truck" and "motor").

In addition, we have discerned that several concepts such as "explosion" "helicopter" and "draws fire" are largely contained in other concepts such as "wolf" and "coin". It is destructive to make use of such highly overlapped labels for SVM training since it is not possible to divide pure positive and negative segments at the level of the entire clips. The GMM model is unable to deal with this difficulty since it is talented to represent the clip as mixture of two concepts. This may make it clear why its performance, averaged over all the classes, shows more than the other approaches.

The GMM+kullback–Leibler approach constantly gives the best results. For instance, GMM+kullback–Leibler attain higher Average Precision than the SVM or HMM for 23 out of the 26 concepts. Nevertheless, the edge of enhancement is relative. The SVM or HMM achieve relative glowing in contrast and are simpler to construct and to assess. Accordingly, depending on the nature of the database and the importance of the main precision, these may be suitable approaches.

4.2 Video shots research experiments

Since a set of labeled audio signal training is used in the classifiers, we are actually attracted by the way the classifier

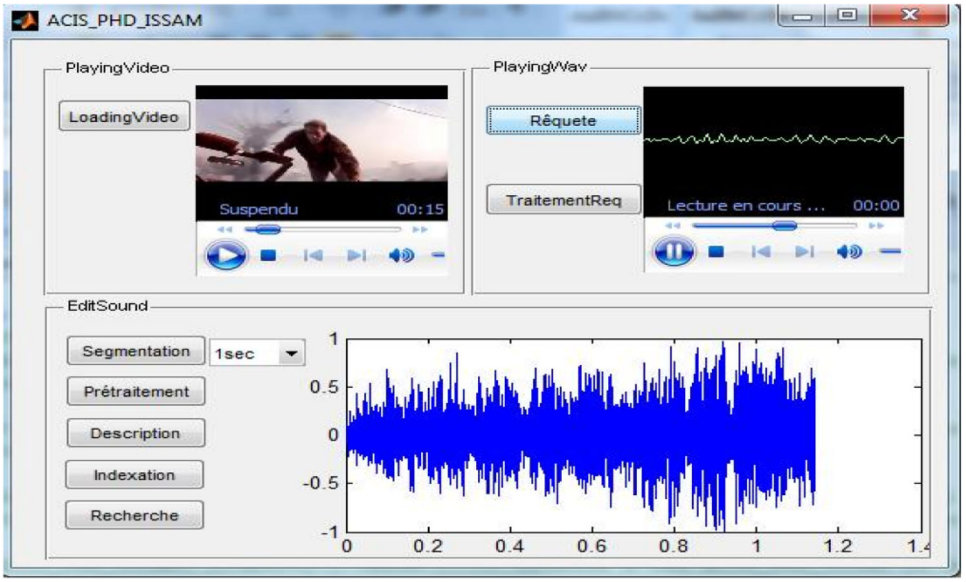

Fig. 9 Graphical interface for user query integration

behaves when it faces data that it has never run before. We can deduce that if the target is to take generalization performance, there is no self-determining circumstance to support one learning more than another. For this reason, to evaluate our video retrieval system, an experiment approach is pursued [26]. This approach maintains two complementary significant cases; the subjective and objective research. The subjective research is principally a qualitative approach to the study of the human behavior and the reasons that govern such a behavior. Researchers have the propensity to grow to be subjectively engrossed by the subject substance in this kind of research method. The objective research is principally quantitative; the researchers have a tendency to stay objectively separated from the subject substance. The objective research is logically regarded as a quantitative approach because it looks for the exact measurements and analysis of the aim of the concepts that respond to its query.

4.2.1 Subjective evaluation

To achieve a subjective experiment process, we used 2000 video clips. The details of the data clips are described in Sect. 3.1. A diversity of 80 sound signal queries and the release of the best corresponding shots are presented to them one by one using a specific interface. For each sound query, they are drilled to pick the retrieved video shots that sound similar to it. Otherwise, they decide 'not found' if they end up with the result that none of the recovered videos seem comparable to the query.

Ten users estimate 80 sound queries by looking at the retrieved video shot clips for each query. This is taken as $10 \times 80 = 800$ samples for evaluating the performance. As an alternative to the conventional precision and to recall the rates of retrieval, the performance is calculated in stipulations of probability of the relevant video shots in the corresponding clips. This experiment demonstrates the probability relative to the relevant video shots. It can be distinguished that the retrieved video shots are higher than the relevant shots by 40 %. In addition, the probability of the relevant retrieved video shots is ≈ 0.8. This is basically the worst-case measure of the retrieval sound system.

4.2.2 Objective evaluation

Figure 9 shows a graphical interface for the user's query integration. The scale of this experiment is restricted here by sound concepts mentioned in Sect. 3.1. Table 2 details six sound queries by examples and the corresponding retrieved video shot clips. We perceive that the retrieved video shots are perfectly related to the query. It is practical that the response means the desired sounds by request but the visual scene does not reflect the source of the sound. For example, in the sound query 3 we find trees but no helicopter, and in the sound query 2 we find an image of dead soldiers rather than an explosion flame.

Table 2 Sound query examples and corresponding video shots retrieved

Sound query	Video shots retrieved description
1/Dog barking	2 barking dogs in a home garden
2/Explosion	Soldiers died in a scene of war
3/Helicopter	Trees whose leaves are trembling
4/Police alarm	The passage of several police cars in full speed
5/Airplane	Plane landing in an airport runway
6/Car	Scene of cars race

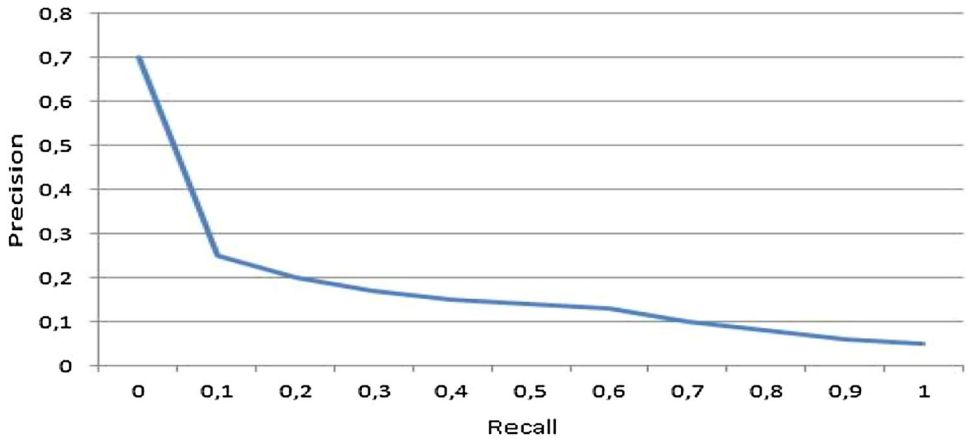

Fig. 10 Precision v/s recall for concepts

For the 26 sound concepts mentioned previously, the precision and recall rates recorded are illustrated in Fig. 10. The retrieval performance is improved using about 80 sound signals as a test sample from each concept.

Recall is spotted from 0 to 1.0 by making an allowance for video shots from the list of the retrieved shots. The precision is obtained by averaging the precision values at every occurrence of a correctly retrieved video shot.

The objective evaluation that uses high level shows that our process can perform as good as the other methods [27]. The advantage of this process is its easy computation. Moreover, the subjective evaluation results indicate that the system is capable of retrieving the video shots containing sounds that are relevant to the user's sound query. To estimate the system performance, it should be acknowledged that video with text labeling-based retrieval is basically different from the example-based retrieval. Text descriptions speak about the concepts that reside in intense volumes. This means that it is easier to solve particular features of descriptions of video by counting the suitable keywords in the query. This is particularly true for complex acoustic events as queries because they require a diversity of perceptual qualities. Based on this analysis, the subjective experiments evaluations, in general, are strict and conformist estimates of the system performance.

5 Conclusion

In this paper, we have illustrated a number of variants of a video–soundtrack retrieval system based on some environmental sounds. Indeed, we have integrated a new audio analysis process in an attempt to separate the audio sources. In particular, we have tried different models for audio classification; namely, the Support Vector Machines, the Gaussian Mixture Model and the Hidden Markov Model using Euclidian, cosine and Kullback–Leibler distance measure. We have produced a support for GMM+Kullback–Leibler. Based on this audio concepts classification, we have constructed a video retrieval system. We show that the integrations of our approaches are efficient to enhance the retrieval effectiveness. Subsequently, video retrieval is relatively a current research field and there are assortments of attractive ways for the future works. Mainly, we are planning to combine the recognition of the same sounds or with other sounds in only one video sequence. Though there have been several studies about the mixture of sounds, there is a little interest in environmental sound extraction. We are currently starting to investigate how to combine together the running of many appropriate distinguishing processes of different sound concepts.

Acknowledgments The authors would like to acknowledge the financial support of this work by grants from the General Direction of Scientific Research and Technological Renovation (DGRSRT), Tunisia, under the ARUB program 01/UR/11/02.

References

1. Saunders J, Lockheed Martin Co (1996) Real-time discrimination of broadcast speech/music. In: IEEE International Conference on Acoustic, Speech, Signal Process, Atlanta, pp 993–996
2. Williams G, Ellis, Daniel PW (1999) Speech/music discrimination based on posterior probability features. In: 6th European Conference on Speech Communication and Technology. Budapest
3. Scheirer E, Slaney M (1997) Construction and evaluation of a robust multifeature speech/music discriminator. In: IEEE International Conferences on Acoust, Speech, Signal Process, Munich, pp 1331–1334
4. Ajmera J, McCowan I, Bourlard H (2003) Speech/music segmentation using entropy and dynamism features in a HMM classification framework. Elsevier Speech Commun 40(3):351–363
5. Zhang T, Kuo C-CJ (2001) Audio content analysis for online audiovisual data segmentation and classification. IEEE Trans Speech Audio Process 9(4):441–457 Fall
6. Guo G, Li SZ (2003) Content-based audio classification and retrieval by support vector machines. IEEE Trans Neural Netw 14(1):209–215
7. Wold E, Blum T, Wheaton J (1996) Content-based classification, search and retrieval of audio. IEEE Trans Multimed 3(3):27–36

8. Malkin R, Waibel A (2005) Classifying user environments for mobile applications using linear autoencoding of ambient audio. Proc IEEE Int Conf Acoustic Speech Signal Process 5:509–512
9. Milner BL, Smith D (2006) Acoustic environment classification. ACM Trans Speech Lang Process 3(2):1–22
10. Chu S, Narayanan S, Kuo C-CJ (2006) Content analysis for acoustic environment classification in mobile robots. In: International Conference on Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic System, Arlington, pp 16–21
11. Su F, Yang L, Lu T, Wang G (2011) Environmental sound classification for scene recognition using local discriminant bases and hmm. In: 19th ACM international conference on Multimedia, Newyork, pp 1389–1392
12. Okuyucu C, Sert M, Yazici A (2013) Audio feature and classifier analysis for efficient recognition of environmental sounds. IEEE International Symposium on Multimedia. Anaheim, pp 125–132
13. Xia-qing X, Quan-wei B, Lei H, Xu W (2013) Study and application of semantic-based image retrieval. J China Univ Posts Telecommun 20(2):136–142
14. Andre-Obrecht R (1988) A new statistical approach for automatic segmentation of continuous speech signals. IEEE Trans Acoustic Speech Signal Process 36(1):29–40
15. Thornburg H (2005) Detection and modeling of transient audio signals with prior information. Ph.D. dissertation, Stanford Univ., Stanford
16. Ellis DPP, Lee K (2004) Minimal-impact audio-based personal archives. 1st ACM Workshop Continuous Archiving and Recording of Personal Experiences CARPE-04, New York
17. Lie Lu, Hanjalic A (2006) Audio elements based auditory scene segmentation. In: IEEE International Conference on Acoustic, Speech, Signal Process, Toulouse, France
18. Wichern G, Thornburg H, Mechtley B, Fink A, Tu K, Spanias A (2007) Robust multi-feature segmentation and indexing for natural sound environments. In: IEEE/EURASIP International Workshop Content- Based Multimedia Indexing, Bordeaux, France, pp 69–76
19. Jafer E, Mahdi AE (2003) Wavelet based voiced/unvoiced classification algorithm. EURASIP Conference focused on video/ image processing and multimedia communications, pp 667–672
20. Feki I, Ben Ammar A, Alimi AM (2012) New process to identify audio concepts based on binary classifiers encapsulation. Int J Comp Elect Eng 4(4):515–518
21. Feki I, Ben Ammar A, Alimi AM (2014) Query sound-by-example video retrieval framework. In: IEEE proceedings of International Conference on Hybrid Intelligent Systems, Kuwait, pp 297–302
22. Vasconcelos N (2004) On the efficient evaluation of probabilistic similarity functions for image retrieval. IEEE Trans Inform Theory 50(7):1482–1496
23. Helén M, Virtanen T (2007) Audio query by example of audio signals using Euclidean distance between Gaussian mixture models. IEEE International Conference on Audio, Speech and Signal Processing, Honolulu, USA, pp 225–228
24. Zhao J, Zhang Z, Han S, Qu C, Yuan Z, Zhang D (2011) SVM based forest fire detection using static and dynamic features. Comp Sci Inform Syst 8(3):821–841
25. Rabiner L, Juang B (1993) Fundamentals of speech recognition. Prentice Hall, New Jersey
26. Weitao W, Yuehui J, Tan Y, Yidong C (2012) A video quality assessment method using subjective and objective mapping stategy. In: IEEE International Conference on Cloud Computing and Intelligent Systems, vol 2, Hangzhou, pp 514–518
27. Jadhav SM, Patil VS (2012) Review of significant researches on multimedia information retrieval. In: IEEE International Conference on Communication, Information and Computing Technology, Mumbai, pp 1–6