Chapter 14
Content-Based Multimedia Retrieval

Flora Amato, Luca Greco, Fabio Persia, Silvestro Roberto Poccia, and Aniello De Santo

14.1 Introduction

During the last years, rapid diffusion of inexpensive tools (mostly embedded in smartphones, tablets, and other common mobile devices) for image, video, and audio capturing, together with improvements in storage technologies, have favored the creation of massive collections of digital multimedia content. However, the difficulty of finding relevant items increases with the growth of the amount of available data. To face this problem, two main approaches are often employed, both involving the use of metadata: the first includes manually annotating multimedia content by means of textual information (e.g., titles, descriptive keywords from a limited vocabulary, and predetermined classification schemes), while the second amounts to using automated feature extraction and object recognition to classify content.

The latter is naturally related to *content-based multimedia retrieval* (CBMR) systems and often represents the only viable solution since text annotations are mostly nonexistent or incomplete; moreover, CBMR has proven to dramatically reduce time and effort to obtain multimedia information, whereas frequent annotation additions and updates to massive data stores often require entering manually all the attributes that might be needed for the queries.

While early algorithms proposed in this field mainly focused on feature-based similarity search (over images, video, and audio), the interest of researchers

F. Amato (✉) • F. Persia • S.R. Poccia • A. De Santo
DIETI, University of Naples Federico II, Naples, Italy
e-mail: flora.amato@unina.it; fabio.persia@unina.it; silvestroroberto.poccia@unina.it; aniello.desanto@gmail.com

L. Greco
DIEM, University of Salerno, Fisciano, Italy
e-mail: lgreco@unisa.it

© Springer International Publishing Switzerland 2015

F. Colace et al. (eds.), *Data Management in Pervasive Systems*, Data-Centric Systems and Applications, DOI 10.1007/978-3-319-20062-0_14

has recently turned to the problem of understanding the semantics of a query rather than concentrating only on the optimization of the underlying computation (*semantic gap*). In this direction, interesting research topics include various kinds of interaction between humans and computers (experiential and affective computing), the introduction of new features to improve the detection/recognition process, the analysis of new media types (3D models, virtual reality, etc.), and the identification of test sets for benchmarking.

In this chapter, we describe the most representative work in the field of content-based multimedia retrieval. In particular, Sect. 14.2 presents a quick survey on text retrieval models and methods. In Sect. 14.3, we report the main trends and approaches for images and video retrieval. In Sect. 14.4, main researches and applications in content-based audio retrieval are covered. Finally, in Sect. 14.5 some cultural heritage applications and further domains for content-based multimedia retrieval are briefly discussed.

14.2 Content-Based Text Retrieval

Already in 1961, Swanson [53] had defined the process of indexing, cataloguing, and classifying documents as an attempt to represent information content within articles in an extremely compact form. He established the two fundamental objectives for information retrieval (IR) systems:

- *Effectiveness*, which concerns how well a system works in terms of percentage of relevant material retrieved and amount of irrelevant material excluded.
- *Economy*, which refers, on the one hand, to the cost of the overall working system, including indexing, storage, and searching, and, on the other hand, to the cost for the user to read irrelevant material.

Current search engines can recognize strings of characters and figures, but, unfortunately, they do not have the key to enter the semantic dimension of words: they cannot understand the richness and the imperfections of natural language. Therefore, even though users try to avoid the muddle of links and knots of the Web, resorting to search engines for help, they still cannot take for granted that the results of their searches will be satisfactory. Often, users searching the Web by means of a Web search engine meet the following obstacles:

- *False positives*: some retrieved Web pages do not match the actual user's intention.
- *False negatives*: some pages matching the user needs are not suggested by the system.

- *Hidden Web*: some interesting pages are in the Web but are not "visible" to the crawler[1] and thus have not been indexed.

A Web search engine is actually the most common example of an IR system. One such system can assist and even influence the activity of many users: knowledge workers, intellectual workers, top managers, and, in general, users who face the problem of information overload on a daily basis [27] creating value from the chaos [38]. In detail, IR involves a specific automatic text analysis activity that focuses on large collections of documents and aims at performing searches through specific queries.

Such queries consist mostly of words, phrases, and sentences written in natural language; moreover, complex Boolean expressions, regular expressions, or finite-state automata can be used [45].

14.2.1 Basic Crawling and Indexing Strategies

Documents can be crawled using several strategies:

- *Orthographic*, where words are treated just as strings of characters
- *Semantic*, where words are connected with the concepts they express
- *Statistical*, where the term frequency is systematically compared with a frequency lexicon [11]

In general, a document can be represented as a set of keywords (or key phrases) that contribute to the description of its content. During the indexing phase, when the collection and the storage of the data are performed, texts are usually preprocessed in order to remove *stop words*[2] and to perform *stemming*.[3] Additionally, the obtained words and stems can be reconnected to their synonyms in order to create relations between words and concept classes [32].

IR systems try to retrieve all the documents that are relevant to a user query while minimizing the number of nonrelevant documents retrieved.

---

[1]A *Web crawler* is the component of a Web search engine that systematically browses the World Wide Web (WWW), typically for the purpose of Web indexing. A Web crawler may also be called a Web *spider* or *automatic indexer*.

[2]Stop words are words that are filtered out before processing natural language data (text), such as articles or prepositions.

[3]The process of reducing inflected or derived words to their stem or morphological root, i.e., the part of the word that is common to all its inflected variants.

14.2.2 IR Models and Weighting Schemes

Information retrieval models are classified into three different categories [58]:

- *Algebraic methods* [80], where documents are represented as vectors, matrices, or tuples that are transformed into a one-dimensional similarity measure using algebra. Examples include the vector space model [81] that is the most used model in IR.
- *Probabilistic approaches* [78], where the relevance of a document depends on the weight of the contained words.
- *Boolean strategies* [12], where documents are represented by sets of terms; the similarities are derived using set-theoretic operations, and the model simply checks the presence/absence of queried terms in the documents. Examples of these models are the standard Boolean model, the extended Boolean model, and the fuzzy model.

The standard Boolean, extended Boolean, vector space, and probabilistic techniques are also known as *statistical methodologies* [22].

Generally speaking, the statistical approaches represent documents and queries using features such as *terms* (words that occur in a given query or collection of documents) or *n-grams* (n-adjacent words) or other aspect obtained through the use of natural language processing methods [2]. Numerical weights can be assigned to each feature.

Among the weighting functions proposed in the literature [80], the following must be recalled:

- The mere frequency of occurrences of a term within a document
- The term specificity [73]
- The inverse document frequency, that is, a word is considered as less relevant if it is too frequent within a corpus of documents [80]
- The term frequency-inverse document frequency (TF-IDF) [73], whose value increases proportionally to the number of times a word appears in the document but is counterbalanced by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general

Other methods that are worth mentioning are the ones based on document-to-document similarity functions, where one can understand how close the content is of two different documents [44, 80]. These are procedures for document classification and clustering and may be *supervised*, if the classes are defined at the start of the process [21], or *unsupervised*, if a preliminary text analysis is required in order to acquire relevant homogeneous conceptual areas and classes [11].

Note that, in many applications, the results produced by information retrieval systems can be misleading: this happens because the processing of natural language is not always accurate and the semantic and syntactic aspects of language are not actually taken into account. Ontologies represent a first solution to this problem by allowing the hierarchical organization of the semantic aspects and the relations involved in the encyclopedic human knowledge. Ontology-oriented approaches to IR have been proposed by several authors [20, 35, 49, 70].

While ontologies are mostly based on nouns, a proposal that comes from the linguistics research area grounds its results on the connection between the syntactic structures and the semantics of the lexicon-grammar semantic predicates (verbs, their nominalization, and their adjectivalizations) [75]. In this field, other contributions that need to be mentioned are [23, 29, 30]. The use of syntactic relationships could increase the effectiveness by reducing the amount of irrelevant material with no loss of what is relevant. This issue has already been studied in the 1970s by Z.S. Harris and the University of Pennsylvania [43]. The use of semantics, instead, helps information retrieval to reduce irrelevant material by applying semantic techniques to disambiguate terms within documents: for instance, to distinguish between documents about *Apple* the computer company from documents about the fruit.

In general, the approaches that make use of semantic information can belong to different classes [39, 47]:

- *Neurolinguistic programming (NLP)-based methods* that ground the semantic analysis of free texts on the syntactic, lexical, and morphological ones [92]
- *Latent semantic indexing methods* that overperformed the traditional vector space techniques [52]
- *Neural nets methods*, based on the spreading activation search model (network nodes correspond to concepts contained into a thesaurus) [83]

In order to improve the performance of IR techniques, several extensions have been proposed; as an example, by embedding search concept or context into the query, or using query expansion and ranking techniques:

- *Conceptual IR* [28] that developed a concept-based retrieval method, relaying on Wikipedia-based explicit semantic analysis [37] and using hand-built thesauri with relations between words and concepts.
- *Contextual IR* [9, 87] aims at improving IR performances by integrating contextual information into the retrieval process and by adapting the results on users' needs. Users' profile and preferences can be profitably exploited to represent context and improve IR effectiveness [59, 67].
- *Automatic query expansion (AQE)*: AQE is an extremely promising technique to improve the retrieval effectiveness of document ranking, outdoing the reluctance and the difficulty of users in providing a more precise description of their information needs. In [14] and [19], the authors illustrate how AQE improves recall and precision of an IR system, describing the main computational aspects of this framework, comprising data acquisition and preprocessing, candidate feature generation and ranking, feature selection, and query reformulation.
- *Ranking and relevance feedback*: When the user gives a query, it is expected that she/he gets the documents most relevant to the query. For this reason, the obtained

documents are then ranked according to their degree of relevance, importance, etc. In addition, relevance feedback (RF) is one of the classical ways of refining search engine rankings: search engine first generates an initial set of rankings; then users select the relevant documents within this ranking, and based on the information in these documents, a more appropriate ranking is presented (e.g., the query may be expanded using the terms contained in the first set of relevant documents). In [77], the problem of ranking linked data is discussed, giving a comprehensive overview of existing ranking methods for the Web of data.

14.3 Content-Based Video and Image Retrieval

With the increasing proliferation of digital video contents, efficient techniques for analysis, indexing, and retrieval of videos according to their contents have become more and more important. A common first step for most content-based video analysis techniques available is segmenting a video into elementary shots, which can be composed to form a video sequence during video sorting or editing with either cut transitions or gradual transitions of visual effects, such as fades, dissolves, and wipes.

Shot boundaries are typically found by computing an image-based distance between adjacent frames and noting when the distance exceeds a certain threshold. The distance between adjacent frames can be based on statistical properties of pixels [48], compression algorithm [7], or edge differences [94]. The most widely used method is based on histogram differences.

In recent years, research has focused on the use of internal features of images and videos computed in an automated or semiautomated way [31]. The common strategy for automatic indexing had been based on using syntactic features alone. However, due to its complexity of operation, there has been a paradigm shift in the research of identifying semantic features [34]. User-friendly content-based retrieval (CBR) systems operating at semantic level would identify motion features as the key besides other features like color, objects, etc., as motion (either of camera motion or shot editing) adds to the meaning of the content.

Moreover, with the availability of image-capturing devices such as digital cameras and image scanners, the size of digital image collection is significantly increasing as well. Efficient image searching, browsing, and retrieval tools are required by users from many different domains. To this end, many general-purpose image retrieval systems have been developed. They can be classified in two different categories: *text based* and *content based*. In the text-based approach, the images are manually annotated by text descriptors that are then used by a DBMS to perform image retrieval. There are two main disadvantages with this approach. The first is that a considerable level of human effort is required for manual annotation. The second is the annotation inaccuracy due to the subjectivity of human perception [26, 84]. To overcome these disadvantages in text-based retrieval systems, the content-based image retrieval (CBIR) has been introduced. In CBIR, images are indexed by their visual content, such as color, texture, and shapes. A pioneering work was published by Chang in 1984 [15]. In the last years, several commercial products and experimental prototype systems have been developed, such as QBIC [33], Photobook [68], Virage [41], VisualSEEK [86], Netra [57], and SIMPLIcity [89].

The section is organized as in the following: Section 14.3.1 describes the possible categories of *video indexing and retrieval technique*, starting from the *shot segmentation* process, which is a preliminary phase essential for content-based video retrieval. Eventually, Sect. 14.3.2 shows the most important techniques for *content-based image retrieval*, first introducing the very important concept of *semantic gap* that exists between low-level and high-level features and then describing both the high-level semantic-based image retrieval and the low-level image features.

14.3.1 Content-Based Video Retrieval

As mentioned before, the shot segmentation is an essential phase for content-based video retrieval. A shot is defined as the consecutive frames from the start to the end of recording in a camera. It shows a continuous action in an image sequence [42]. There are two different types of transitions that can occur between shots: abrupt (discontinuous) also referred to as cut and gradual (continuous) such as fades, dissolves, and wipes.

There are several different approaches for detecting the shot in a video. The most commonly used are listed below:

- Shot boundary detection scheme based on rough-fuzzy set
- Shot boundary detection in low-pass filtered histogram space
- The hidden Markov model technique
- Shot change detection based on sliding window method
- Histogram-based detection
- Shot segmentation by graph partitioning
- Key frame extraction
- Key frame selection using adaptive temporal sampling
- Feature extraction
- Clustering
- Clustering algorithm based on K-L divergency
- Hierarchical clustering method

After performing one of the possible shot segmentation approaches, some *video indexing techniques* can be applied. There are two main possible categories of indexing techniques:

- Syntactic indexing
- Semantic indexing

As regards *syntactic indexing*, some of the prominent content-based retrieval (CBR) systems are IBM's QBIC [36], ViBE [17], VisualSEEK [86] and VideoQ [16], Photobook [68] and FourEyes [69] at MIT., Chabot [64], MARS [61], Virage [8], and Jacob [6]. These systems use syntactic features as the basis for matching and employ either *query by example* or *query-through-dialog box* to interface with the user. Thus, they operate at a lower level of abstraction, and therefore, the user needs to be highly versed in the details of the CBR system to take advantage of them.

On the contrary, in *semantic indexing*, a number of psychological studies and experiments emphasize the need for extracting the semantic information from images and video data. The two important researches in this direction are:

- Demonstrating that higher similarity ratings are produced by perceptually relevant semantic features as opposed to features derived from color histograms on the images [79].
- The performance and the efficiency of searching are generally greatly improved by using semantic cues [66] as compared to when low-level features are employed.

In summary, there is a great need to extract semantic indices for making the CBR system serviceable to the user. Though extracting all such indices might not be possible, there is great scope for furnishing the semantic indices with a certain well-established structure.

Video contains multiple types of audio and visual information, which are difficult to extract, combine, or trade off in general video information retrieval. It is based on the following concepts:

- Similarity measure
- Video retrieval using visual information
- Textual query for video retrieval
- Refinement and relevance feedback
- Pseudo-relevance feedback
- Negative pseudo-relevance feedback

14.3.2 Content-Based Image Retrieval

The main difference between *content-based* and *text-based* retrieval systems is that human interaction is an indispensable part of the latter system. Humans tend to use high-level features (concepts) to interpret images and measure their similarity. In general, there is no direct link between the high-level concepts and the low-level features [84]. Though many complex algorithms have been designed to describe color, shape, and texture features, these algorithms cannot adequately model image semantics and have a lot of limitations while dealing with broad content image databases [63].

In [26], Eakins mentioned three levels of queries in CBIR:

- *Level 1*: Retrieval by primitive features such as color, texture, shape, or the spatial location of image elements
- *Level 2*: Retrieval of objects of given type identified by derived features, with some degree of logical inference
- *Level 3*: Retrieval by abstract attributes, involving a big amount of high-level reasoning about the aim of the objects or scenes depicted

A CBIR system should provide full support in bridging the *semantic gap* between numerical image features and the richness of human semantics [85, 95] in order to support query by high-level concepts.

The state-of-the-art techniques in reducing the *semantic gap* include mainly five categories:

- Using object ontology to define high-level concepts
- Using machine learning tools to associate low-level features with query concepts
- Introducing relevance feedback (RF) into retrieval loop for continuous learning of users' intention
- Generating semantic template (ST) to support high-level image retrieval
- Making use of both the visual content of images and the related textual information (e.g. tags, keywords, descriptions, etc.) from the Web

Retrieval at Level 3 is difficult and less common. Possible Level 3 retrieval can be found in domain-specific areas such as art museums or newspaper libraries. Current systems mostly perform retrieval at Level 2. There are three fundamental components in these systems:

1. Low-level image feature extraction
2. Similarity measure
3. *Semantic gap* reduction

Low-level image feature extraction is the basis of CBIR systems. To perform CBIR, image features can be either extracted from the entire image or from regions (region-based image retrieval, RBIR).

To perform RBIR, the first step is to implement image segmentation. Then, low-level features such as color, texture, shape, or spatial location can be extracted from the segmented regions. Similarity between two images is defined based on region features.

14.4 Content-Based Audio Retrieval

Nowadays, content-based audio retrieval systems can be crucial for several application domains: from music retrieval to speech recognition, including audio segmentation, environmental sound recognition, and acoustic surveillance. As seen for text, images, and video retrieval problems, the identification of the right content-based features is the most important challenge in the design of a good audio retrieval system. In this particular case, the process of extracting features from audio signals is aimed at obtaining a compact and machine-processable description of the meaningful information within those signals. The most common features, such as mel-frequency cepstral coefficients (MFCCs), which were used primarily for speech recognition, are often used for other domains so that the problem of feature identification is generally addressed independently from the specific domain. That said, we can identify the following main fields of research for content-based retrieval systems:

- *Segmentation*, which aims at distinguishing different types of sound (music, silence, speech, etc.) by identifying homogeneous parts in an audio stream; once types have been identified, each one will be processed by means of a particular technique.
- *Music information retrieval*, which deals with the retrieval of similar pieces of music, artists, and genres [25]. The challenging problem of *music transcription* is also addressed although it requires a much deeper analysis to extract pitch, attack, duration, and signal source of each sound in a music track [50].
- *Automatic speech recognition* refers to the recognition of spoken word, language [74], and the extraction of emotions.
- *Environmental sound retrieval* includes types of sound that are neither speech nor music. The problem of audio surveillance is also addressed.

The identification of similar audio content, as perceived by humans, is actually an *inverse* problem. In fact, it deals with the estimation of model parameters by processing observed data. A semantic gap is then introduced because of the mismatch between high-level concepts and low-level description: a classical music track is seen (at a low level) as a series of samples (numeric values) by computers, but it is actually a sequence of notes with specific pitch and durations. Humans can bridge the semantic gap thanks to prior knowledge, so they perceive motifs, themes, and also emotions. The great challenge for the research community today is to make machines able to complete the task and narrow the semantic gap.

14.4.1 A Framework for Audio Retrieval Systems

A typical framework for content-based audio retrieval consists of three main modules [62]:

- An *input module* which aims at extracting features from one or multiple tracks contained in an *audio database*. As seen for text, video, and image retrieval problems, a feature extraction process is required to reduce the amount of data to be handled by the system. In the particular case of audio, the raw waveform would be too big for direct processing and also inadequate for retrieval. Common feature extraction techniques for audio ensure a size reduction of several orders of magnitude. Extracted features are stored in a *feature database*.

- A *query module* that allows the users to interact with the system. A user can provide a query object that contains audio fragments of interest (examples) as well as hummed or whistled melodies. Features must also be extracted from the query object to make a similarity comparison between the query and feature database items possible.
- A *retrieval module* which is the core of the system. Here the similarity between different feature-based media descriptions is computed.

Also for audio retrieval, the most used approach for similarity estimation relies on the vector space model so that a distance measurement is actually performed. Unfortunately, mathematical metrics seem not to match well human perception of similarity. This often leads to low-quality retrieval results, especially when the query itself is imperfect (humming or whistling). In most cases, the user has also the possibility to give a relevance feedback [55], by specifying which retrieved object is relevant for her/his needs. The quality of extracted features plays a very important role in the retrieval task: it is necessary to capture audio properties that show high variation across the available audio samples so that selected features carry only meaningful information and allow to imitate human perception by filtering, for example, components of the original signal which are not perceivable by humans. Typically, different content-based audio features allow to capture different information that can be used for similarity purposes; for example, *pitch* is useful for determining the musical note from an audio signal, while *mel-frequency cepstral coefficients* (MFCCs) are more suitable for the analysis of timbral characteristics.

14.4.2 Properties of Audio Signals

The first distinction we can make about audio signals is between tones and noise [62]: *tones* are “capable of exciting an auditory sensation having pitch,”[4] while noise typically has no pitch. Moreover, we can distinguish between *pure tones* where “the instantaneous sound pressure is a simple sinusoidal function of time” and *complex tones* that contain “sinusoidal component of different frequencies.”[5] Complex tones are usually categorized into *harmonic* and *inharmonic* complex tones that contain partials with frequencies, respectively, at integer or non-integer multiples of the fundamental frequency. Noise can also be classified according to its temporal and spectral characteristics: for example, *stationary noise* is characterized by “negligibly small fluctuations of level within the period of observation,” *broad-*

[4]ANSI/ASA S3.20-1995 (R2008) Bioacoustical Terminology at http://webstore.ansi.org/RecordDetail.aspx?sku=ANSI%2FASA+S3.20-1995+%28R2008%29.

[5]Ibid.

band noise has no pitch, *white noise* equally contains all frequencies within a band, and *colored noise* has a spectral power function of frequency. Audio signals can also be described in terms of psychoacoustic attributes:

- *Duration* is the time between the start and the end of an audio signal and is typically divided into *attack, decay, sustain,* and *release* (these parameters describe the envelope of the sound and in some cases are not all present).
- *Loudness* is related to sound pressure level changes and is defined as "the attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud."
- *Pitch* is "the attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high." It is often used as a synonym of the fundamental frequency.
- *Timbre* is the most complex attribute. It is defined as "the attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are dissimilar." Different musical instruments playing the same musical note actually have a different timbre.

The auditory perception of these attributes is really complex since they cannot be considered as independent. For example, sound pressure and the waveform may alter pitch perception, while loudness is influenced by sound duration (longer sounds appear louder). In order to take into account these issues, proper features have been designed. In the next section, a brief survey on recent audio feature extraction works is presented.

14.4.3 Audio Feature Extraction and Classification Researches

Audio feature sets must be designed in order to properly represent each aspect or attribute of audio signals [60, 88]. Different research works have been proposed in this area. In [93], some fundamental features are discussed: loudness, tone,[6] pitch, and mel-frequency cepstral coefficients (MFCCs); these are typically extracted from audio frames having a duration between 25 and 40 ms. The work in [62] proposes an audio feature taxonomy for content-based audio retrieval and compares properties of state-of-the-art and traditional features. Early works focus on pitch detection. In particular, [88] discusses the use of pitch histograms as a method to represent the pitch content of music signals in both symbolic and audio forms; a multiple-pitch detection algorithm for polyphonic signals is used to calculate pitch histograms for audio signals. Such a method has proven to obtain good results in automatic musical genre classification. In [51], a novel set of tempo-related audio features for application in audio retrieval is discussed, and it relies on the definition of the

[6]Tone is typically related to the brightness and the bandwidth of sound.

cyclic beat spectrum. A stochastic method for the representation of sung melodic contour is proposed in [65]: such a representation is obtained by fitting probability distribution functions. Semantic music retrieval by the use of social tags and Web-mined documents is also investigated in [10]: here feature kernel combination is used. In [24], an approach for automatic extraction of high-level audio features through distributed computing is presented, and an audio analysis method to create high-level annotation is discussed in [72].

Several works have also been proposed in the field of audio classification. The work of Liu [56], for example, represents a first attempt at building a multimedia search engine for the Internet: a fuzzy inference system for audio classification and retrieval. Support vector machines (SVMs) have also been employed with a multi-classification strategy [40] to address this problem. Experiments with SVM methods showed good results, an example of effective training for audio categorization with low error rate. In [60, 71], Gaussian Mixture Models (GMM) were used to classify audio by grouping N-dimensional feature vectors. These approaches involve the use of different kinds of standard features (MFCC, loudness, sharpness, etc.) as well as original features (entropy modulation, stationary segment duration, and so on). A hidden Markov model-based approach is presented in [76], where acoustic segment models are used to classify effectively music genres. Recently, some hybrid approaches have been proposed for audio classification. In [18], a weight factor based on support vector is applied to the Euclidean distance and k-NN rule to improve accuracy by 28 % with respect to the classic Euclidean-based k-NN classifier. The modified two-dimensional root cepstrum was introduced in [46] and tested to classify recordings of different classes of gunshots: an accuracy of 98.93 % was reported.

14.4.4 Applications and Tools for Content-Based Audio Retrieval

Different researches in the field of audio retrieval have focused on the development of applications and tools. Until now, the most successful application is Shazam [90]. The algorithm, introduced by Wang in 2003, relies on the fingerprinting technique and combinatorial hashing. It has proven to be robust against noise and distortions, and its computational efficiency allows great scalability; in fact, it is the most used audio search algorithm for mobile audio search engines. Experimental results on a database of about 20,000 tracks showed a search time of about 5–500 ms. Early tools for audio retrieval date back to 2000. Works such as Maryas, LibXtract, MIRtoolbox, and Mirage [13, 54, 82] relied on fast Fourier transform (FFT) algorithms to analyze audio content, extract features, and optimize audio similarity. MARSYAS[7] is a software framework for rapid prototyping and experimentation

[7] http://marsyas.info

with audio analysis and synthesis with specific emphasis to music signals and music information retrieval. The basic goal is to provide a general, extensible, and flexible architecture that allows easy experimentation with algorithms and provides fast performance that is useful in developing real-time audio analysis and synthesis tools. LIBXTRACT[8] is a simple, portable, lightweight library of audio feature extraction functions. The purpose of the library is to provide a relatively exhaustive set of feature extraction primitives that are designed to be "cascaded" to create extraction hierarchies. For example, "variance," "average deviation," "skewness," and "kurtosis" all require the "mean" of the input vector to be precomputed. MIRtoolbox[9] is a MATLAB toolbox dedicated to the extraction of musical features from audio files, including routines for statistical analysis, segmentation, and clustering. MIRtoolbox integrates a user-friendly syntax that enables to easily combine low- and high-level operators into complex flowcharts. MIRAGE[10] is an implementation of the research in automatic playlist generation and music similarity. Mirage analyzes music collection and computes acoustic similarity models for each song. It is also able to automatically generate playlists of similar music. It implements psychoacoustic modeling and Gaussian models as well as the FFT algorithm. RythMiXearch is a project presented by Kato that implements a query by example approach to music search. This method is able to accept two examples for each query and mix them; then, using latent Dirichlet allocation, it builds clusters and retrieves similar objects.

14.5 Further Application Domains for Content-Based Multimedia Retrieval

In the last section, we show possible applications of previously discussed methodologies, especially in the cultural heritage domain that represents in Italy a resource of inestimable value. The Internet of Things and Internet of Services are becoming the basic building blocks toward a unified ICT platform for a variety of applications. Since cellular phones and participatory sensor networks easily allow public and professional users to analyze and share local knowledge, there is an increasing interest in applying outcomes from information retrieval research to mobile applications. Several works have been proposed in this regard.

In [4], a novel data model for 3D objects and a tool for building, querying, and storing 3D objects into a relational database are presented. The preliminary environment is designed to be powerful enough to include functionalities and features of modern 3D description languages (X3D and Collada).

[8] http://libxtract.sourceforge.net/

[9] http://www.mathworks.com/matlabcentral/fileexchange/24583-mirtoolbox

[10] http://hop.at/mirage/

A multimedia semantic approach for recommending item in browser system, based on semantic contents and low-level features, is proposed in [1]. Here, a multimedia object recommender prototype for browsing the “Uffizi Gallery” digital picture collection is implemented to investigate the effectiveness of the proposed approach, based on users’ satisfaction. The recommender system aims at helping users browse digital reproduction of Uffizi Gallery by suggestions computed with a novel method for recommendations.

Another interesting application is introduced in [5] by the DATABENC district. The aim of the project is to customize the system for an indoor museum, providing a personalized visiting experience for tourists (e.g., by designing specific “talking objects”).

In the SNOPS project for smart cities, the focus is on the context-aware recommendation services. A recommendation strategy for planning browser activities exploiting object features, user behaviors, and context information is described in [3]. The contextual knowledge is modeled by using ontologies similar to CONON [91].

In this work, a system that provides personalized visiting paths to the tourists of Herculaneum ruins is presented. The effectiveness of this approach is evaluated by investigating the browsing effectiveness and users’ satisfaction.

In this chapter, we have described how different information retrieval approaches can be used and mixed in order to build complex pervasive information systems that enhance user experience. These techniques find wide application in several emerging contexts such as smart cities.

References

1. Albanese, M., d’Acierno, A., Moscato, V., Persia, F., Picariello, A.: A multimedia semantic recommender system for cultural heritage applications. In: 2011 Fifth IEEE International Conference on Semantic Computing (ICSC), pp. 403–410. IEEE, Palo Alto (2011)
2. Amato, F., Mazzeo, A., Penta, A., Picariello, A.: Building rdf ontologies from semi-structured legal documents. In: CISIS, pp. 997–1002 (2008)
3. Amato, F., Chianese, A., Moscato, V., Picariello, A., Sperli, G.: Snops: a smart environment for cultural heritage applications. In: Proceedings of the Twelfth International Workshop on Web Information and Data Management, pp. 49–56. ACM, Maui (2012)
4. Amato, F., Mazzeo, A., Moscato, V., Picariello, A.: Building and retrieval of 3d objects in cultural heritage domain. In: 2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), pp. 816–821. IEEE, Palermo (2012)
5. Amato, F., Chianese, A., Mazzeo, A., Moscato, V., Picariello, A., Piccialli, F.: The talking museum project. Proc. Comput. Sci. **21**, 114–121 (2013)
6. Ardizzone, E., La Cascia, M.: Automatic video database indexing and retrieval. In: Zhang, H., Aigrain, P., Petkovic, D. (eds.) Representation and Retrieval of Video Data in Multimedia Systems, pp. 29–56. Springer, New York (1997). doi:10.1007/978-0-585-31786-13. http://dx.doi.org/10.1007/978-0-585-31786-13
7. Arman, F., Hsu, A., Chiu, M.Y.: Image processing on encoded video sequences. Multimedia Syst. **1**(5), 211–219 (1994). doi:10.1007/BF01268945. http://dx.doi.org/10.1007/BF01268945

8. Bach, J.R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R.C., Shu, C.F.: Virage image search engine: an open framework for image management. In: Storage and Retrieval for Image and Video Databases (1996). doi:10.1117/12.234785. http://dx.doi.org/10.1117/12.234785
9. Bahrami, A., Jun, Y.: Methods and systems for context based query formulation and information retrieval. U.S. Patent No. 7,970,786. U.S. Patent and Trademark Office, Washington, DC, June 2011
10. Barrington, L., Yazdani, M., Turnbull, D., Lanckriet, G.R.: Combining feature kernels for semantic music retrieval. In: ISMIR, pp. 614–619 (2008)
11. Bolasco, S.: Statistica testuale e text mining: alcuni paradigmi applicativi. Quaderni di Statistica **7**, 17–53 (2005)
12. Bordogna, G., Pasi, G.: A fuzzy linguistic approach generalizing boolean information retrieval: a model and its evaluation. J. Am. Soc. Inf. Sci. **44**(2), 70–82 (1993)
13. Bullock, J.: libxtract: a lightweight library for audio feature extraction. In: Proceedings of the 2007 International Computer Music Conference, vol. 2, pp. 25–28. ICMA, Copenhagen (2007)
14. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. (CSUR) **44**(1), 1 (2012)
15. Chang, S.K., Liu, S.H.: Picture indexing and abstraction techniques for pictorial databases. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**(4), 475–484 (1984). doi:10.1109/TPAMI.1984.4767552
16. Chang, S., Chen, W., Meng, H.J., Sundaram, H., Zhong, D.: Videoq: an automated content based video search system using visual cues. In: Proceedings of ACM Multimedia, pp. 313–324 (1997)
17. Chen, J.Y., Taskiran, C., Delp, E., Bouman, C.: Vibe: a new paradigm for video database browsing and search. In: Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, 1998, pp. 96–100 (1998). doi:10.1109/IVL.1998.694510
18. Chung, Y.Y., Choi, E.H.C., Liu, L., Shukran, M.A.M., Shi, D.Y., Chen, F.: A new hybrid audio classification algorithm based on svm weight factor and euclidean distance. In: Proceedings of the 2007 Annual Conference on International Conference on Computer Engineering and Applications, CEA'07, pp. 152–157. World Scientific and Engineering Academy and Society, Stevens Point (2007). http://dl.acm.org/citation.cfm?id=1348258.1348286
19. Clarizia, F., Colace, F., De Santo, M., Greco, L., Napoletano, P.: Mixed graph of terms for query expansion. In: 11th International Conference on Intelligent Systems Design and Applications (ISDA), 2011, pp. 581–586 (2011). doi:10.1109/ISDA.2011.6121718
20. Colace, F., Santo, M.D., Greco, L.: An adaptive product configurator based on slow intelligence approach. Int. J. Metadata Semant. Ontol. **9**(2), 128–137 (2014). doi:10.1504/IJMSO.2014.060340. http://dx.doi.org/10.1504/IJMSO.2014.060340
21. Colace, F., Santo, M.D., Greco, L., Napoletano, P.: Text classification using a few labeled examples. Comput. Hum. Behav. (2014). doi:10.1016/j.chb.2013.07.043. http://www.sciencedirect.com/science/article/pii/S0747563213002823
22. Cummins, R.: A standard document score for information retrieval. In: Proceedings of the 2013 Conference on the Theory of Information Retrieval, p. 24. ACM, New York (2013)
23. D'Agostino, E., Elia, A., Vietri, S.: Lexicon-grammar, electronic dictionaries and local grammars of Italian. In: Leclère, Ch., Laporte, É., Piot, M., Silberztein, M. (eds.) Lexique, Syntaxe et Lexique-Grammaire. Papers in Honour of Maurice Gross, Lingvisticae Investigationes Supplementa, vol. 24. IGML, Amsterdam (2004)
24. Deliege, F., Chua, B.Y., Pedersen, T.B.: High-level audio features: distributed extraction and similarity search. In: Ninth International Conference on Music Information Retrieval, Philadelphia, pp. 565–570, September 2008
25. Downie, J.S.: Music information retrieval. Annu. Rev. Inf. Sci. Technol. **37**(1), 295–340 (2003). doi:10.1002/aris.1440370108. http://dx.doi.org/10.1002/aris.1440370108
26. Eakins, J., Graham, M.: Content-based image retrieval, Technical Report (1999)
27. Edmunds, A., Morris, A.: The problem of information overload in business organisations: a review of the literature. Int. J. Inf. Manag. **20**(1), 17–28 (2000)

28. Egozi, O., Markovitch, S., Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. (TOIS) **29**(2), 8 (2011)
29. Elia, A., Vietri, S., Postiglione, A., Monteleone, M., Marano, F.: Data mining modular software system. In: SWWS, pp. 127–133 (2010)
30. Elia, A., Guglielmo, D., Maisto, A., Pelosi, S.: A linguistic-based method for automatically extracting spatial relations from large non-structured data. In: Algorithms and Architectures for Parallel Processing, pp. 193–200. Springer, Heidelberg (2013)
31. Fablet, R., Bouthemy, P., Pérez, P.: Statistical motion-based video indexing and retrieval. In: International Conference on Content-Based Multimedia Information Access, pp. 602–619 (2000)
32. Faloutsos, C., Oard, D.W.: A survey of information retrieval and filtering methods. Technical Report, CS-TR-3514, University of Maryland (1995)
33. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and effective querying by image content. J. Intell. Inf. Syst. **3**(3-4), 231–262 (1994). doi:10.1007/BF00962238. http://dx.doi.org/10.1007/BF00962238
34. Fan, J., Elmagarmid, A.K., Zhu, X., Aref, W.G., Wu, L.: Classview: hierarchical video shot classification, indexing, and accessing. Trans. Multimedia **6**(1), 70–86 (2004). doi:10.1109/TMM.2003.819583. http://dx.doi.org/10.1109/TMM.2003.819583
35. Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., Motta, E.: Semantically enhanced information retrieval: an ontology-based approach. Web Semant. Sci. Serv. Agents World Wide Web **9**(4), 434–452 (2011)
36. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: the qbic system. Computer **28**(9), 23–32 (1995). doi:10.1109/2.410146
37. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI, vol. 7, pp. 1606–1611 (2007)
38. Gantz, J., Reinsel, D.: Extracting value from chaos. IDC iview, pp. 1–12 (2011)
39. Grossman, D.A.: Information Retrieval: Algorithms and Heuristics, vol. 15. Springer, New York (2004)
40. Guo, G., Li, S.Z.: Content-based audio classification and retrieval by support vector machines. IEEE Trans. Neural Netw. **14**(1), 209–215 (2003)
41. Gupta, A., Jain, R.: Visual information retrieval. Commun. ACM **40**(5), 70–79 (1997). doi:10.1145/253769.253798. http://doi.acm.org/10.1145/253769.253798
42. Han, S., Yoon, K., Kweon, I.: A new technique for shot detection and key frames selection in histogram space. In: 12th Workshop on Image Processing and Image Understanding, pp. 475–479 (2000)
43. Harris, Z.S.: Linguistic transformations for information retrieval. In: Papers in Structural and Transformational Linguistics, pp. 458–471. Springer, Dordrecht (1970)
44. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G., Milios, E.: Information retrieval by semantic similarity. Int. J. Semant. Web Inf. Syst. **2**(3), 55–73 (2006)
45. Hopcroft, J.E.: Introduction to Automata Theory, Languages, and Computation. Pearson Education India, New Delhi (1979)
46. Hossein, M.: Automatic audio classification using modified two dimensional root cepstral features. In: The International Conference on Electrical Engineering (2008)
47. Ishikawa, Y., Subramanya, R., Faloutsos, C.: Mindreader: querying databases through multiple examples. Computer Science Department, p. 551 (1998)
48. Jain, R.: Dynamic vision. In: 9th International Conference on Pattern Recognition, 1988, vol. 1, pp. 226–235 (1988). doi:10.1109/ICPR.1988.28212
49. Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N.K., Alpaslan, F.N.: An ontology-based retrieval system using semantic indexing. Inf. Syst. **37**(4), 294–305 (2012)
50. Klapuri, A., Davy, M. (eds.): Signal Processing Methods for Music Transcription. Springer, New York (2006)

51. Kurth, F., Gehrmann, T., Müller, M.: The cyclic beat spectrum: tempo-related audio features for time-scale invariant audio identification. In: Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR), Victoria, pp. 35–40, 8–12 October 2006
52. Landauer, T.K., Dumais, S.T.: A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. **104**(2), 211 (1997)
53. Landauer, R., Swanson, J.: Frequency factors in the thermally activated process. Phys. Rev. **121**(6), 1668 (1961)
54. Lartillot, O., Toiviainen, P.: A matlab toolbox for musical feature extraction from audio. In: International Conference on Digital Audio Effects, pp. 237–244 (2007)
55. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. **2**(1), 1–19 (2006). doi:10.1145/1126004.1126005. http://doi.acm.org/10.1145/1126004.1126005
56. Liu, M., Wan, C., Wang, L.: Content-based audio classification and retrieval using a fuzzy logic system: towards multimedia search engines. Soft Comput. **6**(5), 357–364 (2002). doi:10.1007/s00500-002-0189-3. http://dx.doi.org/10.1007/s00500-002-0189-3
57. Ma, W., Manjunath, B.: Netra: a toolbox for navigating large image databases. In: Proceedings of International Conference on Image Processing, 1997, vol. 1, pp. 568–571 (1997). doi:10.1109/ICIP.1997.647976
58. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
59. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts. In: Advances in Information Retrieval, pp. 362–367. Springer, Heidelberg (2011)
60. Mckinney, M., Breebaart, J.: Features for audio and music classification. In: Proceedings of the International Symposium on Music Information Retrieval, pp. 151–158 (2003)
61. Mehrotra, S., Rui, Y., Ortega-Binderberger, M., Huang, T.: Supporting content-based queries over images in mars. In: Proceedings of IEEE International Conference on Multimedia Computing and Systems '97, pp. 632–633 (1997). doi:10.1109/MMCS.1997.609791
62. Mitrovic, D., Zeppelzauer, M., Breiteneder, C.: Features for content-based audio retrieval. Adv. Comput. **78**, 71–150 (2010). http://dblp.uni-trier.de/db/journals/ac/ac78.htmlMitrovicZB10
63. Mojsilovic, A., Rogowitz, B.: Capturing image semantics with low-level descriptors. In: Proceedings of 2001 International Conference on Image Processing, vol. 1, pp. 18–21 (2001). doi:10.1109/ICIP.2001.958942
64. Ogle, V.E., Stonebraker, M.: Chabot: retrieval from a relational database of images. Computer **28**(9), 40–48 (1995)
65. Ohishi, Y., Goto, M., Itou, K., Takeda, K.: A stochastic representation of the dynamics of sung melody. In: Dixon, S., Bainbridge, D., Typke, R. (eds.) ISMIR, pp. 371–372. Austrian Computer Society, Vienna (2007). http://dblp.uni-trier.de/db/conf/ismir/ismir2007.htmlOhishiGIT07
66. Papathomas, T., Conway, T., Cox, I., Ghosn, J., Miller, M., Minka, T., Yianilos, P.: Psychophysical studies of the performance of an image database retrieval system. In: IS&T/SPIE Symposium on Electronic Imaging: Science and Technology, Conference on Human Vision and Electronic Imaging III, pp. 591–602 (1998)
67. Patel, A.J.: Systems and methods for highlighting search results. U.S. Patent No. 6,839,702, 4 Jan 2005
68. Pentland, A., Picard, R., Sclaroff, S.: Photobook: content-based manipulation of image databases. Int. J. Comput. Vis. **18**(3), 233–254 (1996). doi:10.1007/BF00123143. http://dx.doi.org/10.1007/BF00123143
69. Picard, R., Minka, T.: Vision texture for annotation. Multimedia Syst. **3**(1), 3–14 (1995). doi:10.1007/BF01236575. http://dx.doi.org/10.1007/BF01236575
70. Pickard, A.J.: Research Methods in Information. Facet, London (2013)
71. Pinquier, J., Arias, J., André-Obrecht, R.: Audio classification by search of primary components. In: International Workshop on Image, Video and Audio Retrieval and Mining, Sherbrooke (2004)

72. Pohle, T., Knees, P., Seyerlehner, K., Widmer, G.: A high-level audio feature for music retrieval and sorting. In: DAFx-10 (2010)
73. Quinlan, J.R.: Combining instance-based and model-based learning. In: ICML, pp. 236–243 (1993)
74. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Upper Saddle River (1993)
75. Ranchhod, E.: Lexique-grammaire du portugais: prédicats nominaux supportés par estar. Lingvisticae Investigationes **13**(2), 351–367 (1989)
76. Reed, J.: A study on music genre classification based on universal acoustic models. In: ISMIR 2006, pp. 89–94 (2006)
77. Roa-Valverde, A.J., Sicilia, M.A.: A survey of approaches for ranking on the web of data. Inf. Retr. **17**, 295–325 (2014)
78. Robertson, S.E., Jones, K.S.: Relevance weighting of search terms. J. Am. Soc. Inf. Sci. **27**(3), 129–146 (1976)
79. Rogowitz, B.E., Frese, T., Smith, J., Bouman, C.A., Kalin, E.: Perceptual image similarity experiments. In: Proceedings of the SPIE Human Vision and Electronic Imaging III, vol. 3299 (1998)
80. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
81. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
82. Schnitzer, D.: High-performance music similarity computation and automatic playlist generation. Technical Report, University of Technology, Vienna (2007)
83. Scholtes, J.C.: Unsupervised learning and the information retrieval problem. In: 1991 IEEE International Joint Conference on Neural Networks, pp. 95–100. IEEE, Seattle (1991)
84. Sethi, I.K., Coman, I.L., Stan, D.: Mining association rules between low-level image features and high-level concepts (2001). doi:10.1117/12.421083. http://dx.doi.org/10.1117/12.421083
85. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
86. Smith, J.R., Chang, S.F.: Visualseek: a fully automated content-based image query system. In: Proceedings of the Fourth ACM International Conference on Multimedia, MULTIMEDIA '96, pp. 87–98. ACM, New York (1996). doi: 10.1145/244130.244151. http://doi.acm.org/10.1145/244130.244151
87. Tamine-Lechani, L., Boughanem, M., Daoud, M.: Evaluation of contextual information retrieval effectiveness: overview of issues and research. Knowl. Inf. Syst. **24**(1), 1–34 (2010)
88. Tzanetakis, G., Ermolinskyi, A., Cook, P.: Pitch histograms in audio and symbolic music information retrieval. In: Proceedings of the Third International Conference on Music Information Retrieval: ISMIR, pp. 31–38 (2002)
89. Wang, J., Li, J., Wiederhold, G.: Simplicity: semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001). doi:10.1109/34.955109
90. Wang, A., et al.: An industrial strength audio search algorithm. In: ISMIR, pp. 7–13 (2003)
91. Wang, X.H., Zhang, D.Q., Gu, T., Pung, H.K.: Ontology based context modeling and reasoning using owl. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, pp. 18–22. IEEE, Orlando (2004)
92. Wendlandt, E.B., Driscoll, J.R.: Incorporating a semantic analysis into a document retrieval strategy. In: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 270–279. ACM, New York (1991)
93. Wold, E., Blum, T., Keislar, D., Wheaten, J.: Content-based classification, search, and retrieval of audio. IEEE MultiMedia **3**(3), 27–36 (1996). doi:10.1109/93.556537

94. Zabih, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classifying scene breaks. In: Proceedings of the Third ACM International Conference on Multimedia, MULTIMEDIA '95, pp. 189–200. ACM, New York (1995). doi:10.1145/217279.215266. http://doi.acm.org/10.1145/217279.215266
95. Zhou, X.S., Huang, T.S.: Cbir: from low-level features to high-level semantics. In: Proceedings of SPIE Image and Video Communication and Processing (2000). doi:10.1117/12.382975. http://dx.doi.org/10.1117/12.382975