Chapter 8: Multimedia and Multimodal Information Retrieval

Alessandro Bozzon and Piero Fraternali

Dipartimento di Elettronica e Informazione, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
{alessandro.bozzon, piero.fraternali}@polimi.it

Abstract. The Web is progressively becoming a multimedia content delivery platform. This trend poses severe challenges to the information retrieval theories, techniques and tools. This chapter defines the problem of multimedia information retrieval with its challenges and application areas, overviews its major technical issues, proposes a reference architecture unifying the aspects of content processing and querying, exemplifies a next-generation platform for multimedia search, and concludes by showing the close ties between multi-domain search investigated in Search Computing and multimodal/multimedia search.

Keywords: multimedia information retrieval, digital signal processing, video search engines, multi-modal query interfaces.

1 Introduction

The growth of digital content has reached impressive rates in the last decade, fuelled by the advent of the so-called "Web 2.0" and the emergence of user-generated content. At the same time, the convergence of the fixed-network Web, mobile access, and digital television has boosted the production and consumption of audio-visual materials, making the Web a truly multimedia platform.

This trend challenges search as we know it today, due to the more complex nature of multimedia with respect to text, in all the phases of the search process: from the expression of the user's information need to the indexing of content and the processing of queries by search engines.

This Chapter gives a concise overview of Multimedia Information Retrieval (MIR), the long-standing discipline at the base of audio-visual search engines, and connects the research challenges in this area to the objectives and research goals of Search Computing.

MIR amplifies many of the research problems at the base of search over textual data. The grand challenge of MIR is bridging the gap between queries and content: the former are either expressed by keywords, like in text search engines, or, by extension, with non-textual samples (e.g., an image or a piece of music). Unlike in text search engines, where the query has the same format of content and can be matched almost directly to it, query processing in MIR must fill an enormous gap. To understand if an image, video or piece of music is relevant to some keyword, it is necessary to extract the hidden knowledge buried inside the aural and visual

S. Ceri and M. Brambilla (Eds.): Search Computing, LNCS 5950, pp. 135–155, 2010.

resources, a multi-sensorial recognition problem that in nature living organisms took quite a long time to solve.

Not surprisingly, MIR research revolves around the problem of extracting, organizing and making available for querying the knowledge present inside media assets. This problem is far from being solved in general, but many effective techniques have been devised for special cases, typically for the extraction of specific "features" from specific non-textual resources. Applications like music mood classification and similarity matching, face recognition, video optical character recognition are examples of these techniques, already deployed in commercial multimedia search solutions.

Since giving the full account of MIR research goes beyond the limits of this Chapter, we have organized the illustration so as to give a flavor of the essential themes. After exemplifying the numerous applications that motivate the growing interest in MIR (Section 1.1), Section 2 overviews the principal research topics in the development of a MIR solution: from the acquisition of content (Section 2.1), to its normalization for the purpose of processing (Section 2.2), to the extraction of the features useful for searching and their organization by means of suitable indexes (Section 2.3), to the languages and algorithms for processing queries (Section 2.4), to the problem of presenting search results (Section 2.5).

The variety of MIR solutions available can be abstracted by a common architecture, which is the subject of Section 3; a MIR system can be seen as an infrastructure for governing two main processes: the *content process*, treated in Section 3.1, comprises all the steps necessary to extract indexable features (called *metadata*) from multimedia elements; the query process, overviewed in Section 3.2, includes all the steps for executing a user's query.

The link between the content process and the query process is represented by *metadata*, which encode the knowledge that the MIR system is able to extract from the media assets, index and use for answering queries. Given that no single universal standard still exists for MIR metadata, Section 4 overviews some of the most popular formats that have been proposed in different application domains. How to extract such metadata from audiovisual data is the subject of Section 5, which presents a bird's eyes view of some feature extraction approaches for audio, image, and video content. This is the area where current research is most active, because the problem of understanding the content of non-textual data is far from being solved in a general way. In Section 6, we also provide an overview of the different query languages used in MIR, which go from simple keyword queries to structured languages.

To make the Chapter more concrete, Section 7 mentions a number of research and commercial MIR systems, where the architecture and techniques described in the preceding Sections have been put to work.

We conclude the Chapter with an outlook (in Section 8) of what lessons can be mutually learnt by researchers in MIR and Search Computing. As in MIR, Search Computing relies on a well balanced mix of offline content preparation (the wrapping and registration of heterogeneous data sources) and smart query processing; moreover, the presentation of multimedia search results requires smart solutions for easing the interpretation of complex results sets, which exhibit sophisticated internal structure, and a spatial as well as temporal distribution. MIR systems, pioneers of a search technology that goes beyond textual Web pages, may be an interesting source of inspiration for the multi-domain content integration, query processing, and result presentation challenges that Search Computing is facing.

1.1 Motivations, Requirements and Applications of Multimedia Search

“Finding the title and author of a song recorded with one’s mobile in a crowded disco”; “Locating news clips containing interviews to President Obama and accessing the exact point where the Health Insurance Reform is discussed”; “Finding a song matching in mood the images to be placed in a slideshow”. These are only a few examples of what multimedia information retrieval is about: satisfying a user’s information need that spans across multiple media, which can itself be expressed using more than one medium.

The requirements of a MIR application bring to the extreme or go beyond the problems faced in classical text information retrieval [37]:

- **Opacity of Content:** whereas in text IR the query and the content use the same medium, MIR content is opaque, in the sense that the knowledge necessary to verify if an item is relevant to a user’s query is deeply embedded in it and must be extracted by means of a complex pre-processing (e.g., extracting speech transcriptions from a video).
- **Query Formulation Paradigm:** as for traditional search engines, keywords may not be the only way of seeking for information: for instance, queries can be expressed by analogy, submitting a sample of content “similar” to what the user is searching for. In MIR, content samples used as queries can be as complex as an image, a piece of music, or even a video fragment.
- **Relevance Computation:** in text search, relevance of documents to the user’s query is computed as the similarity degree between the vectors of *words* appearing in the document and in the query (modulo lexical transformations). In MIR, the comparison must be done on a much wider variety of features, characteristic not only of the specific medium in which the content and the query are expressed, but even of the application domain (e.g., two audio files can be deemed similar in a music similarity search context, but dissimilar in a topic-based search application).

MIR applications requirements have been extensively addressed in the last three decades, both in the industrial and academic fields. As a consequence, MIR is now a consolidated discipline, adopted into a wide variety of domains [41], including:

- Architecture, real estate, and interior design (e.g., searching for ideas).
- Broadcast media selection (e.g., radio channel [58], TV channel).
- Cultural services (history museums [11], art galleries, etc.).
- Digital libraries (e.g., image catalogue [69], musical dictionary, bio-medical imaging catalogues [4], film, video and radio archives [52]).
- E-Commerce (e.g., personalized advertising, on-line catalogues [53]).
- Education (e.g., repositories of multimedia courses, multimedia search for support material).
- Home Entertainment (e.g., systems for the management of personal multimedia collections [27], including manipulation of content, e.g. home video editing [2], searching a game, karaoke).

- Investigation (e.g., human characteristics recognition [22], forensics [40]).
- Journalism (e.g. searching speeches of a certain politician [25] using his name, his voice or his face [23]).
- Multimedia directory services (e.g. yellow pages, Tourist information, Geographical information systems).
- Multimedia editing (e.g., electronic news service [16], media authoring).
- Remote sensing (e.g., cartography, ecology [81], natural resources management).
- Social (e.g. dating services, podcast [54] [56]).
- Surveillance (e.g., traffic control, surface transportation, non-destructive testing in hostile environments).

2 Challenges of Multimedia Information Retrieval

Multimedia search engines and their applications operate on a very heterogeneous spectrum of content, ranging from home-made content created by users to high value premium productions, like feature film video. The quality of content largely determines the kind of processing that is possible for extracting information and the kind of queries that can be answered. This Section overviews the main challenges in the design of a MIR solution, by following the lifecycle of multimedia content, from its entrance into the system (acquisition), to its preparation for analysis (normalization), to the extraction of metadata necessary for building the search engine indexes (indexing), to the processing of a user's query (querying) and, finally, to the presentation of results (browsing).

2.1 Challenge 1: Content Acquisition

In text search engines, content comes either from a closed collection (as, e.g., in a digital library) or is crawled from the open Web. In MIR, multimedia content can be acquired in a way similar to document acquisition:

- By crawling the Web or local media repositories.
- By user's contribution or syndicated contribution from content aggregators.

Additionally, multimedia content can also come directly from production devices directly connected to the system, such as scanners, digital cameras, smartphones, or broadcast capture devices (e.g., from air/cable/satellite broadcast, IPTV, Internet TV multicast, etc.).

Besides the heterogeneity of acquisition sources and protocols, also the size of media files make the content ingestion task more complicated, e.g., because the probability of download failures increases, the cost of storing duplicates or near duplicates becomes less affordable, and the presence of DRM issues on the downloaded content is more frequent.

As for textual data, but even more critical in the case of audiovisual content, is the capability of the content ingestion subsystem to preserve or event enhance the intrinsic quality of the downloaded digital assets, e.g., by acquiring them at the best resolution possible, given the bandwidth limitations, and preserving all the available *metadata* associated with them.

Metadata are textual descriptions that accompany a content element; they can range in quantity and quality, from no description (e.g., Webcam content) to multilingual data (e.g., closed captions and production metadata of motion pictures). Metadata can be found:

- Embedded within content (e.g., video close captions or Exchangeable image file format (EXIF) data embedded in images).
- In surrounding Web pages or links (e.g., HTML content, link anchors, etc).
- In domain-specific databases (e.g., IMDB [72] for feature films).
- In ontologies (e.g., like those listed in the DAML Ontology Library [71]).

The challenge here is building scalable and intelligent content acquisition systems, which could ingest content exploiting different communication protocols and acquisition devices, decide the optimal resolution in case alternative representations are available, detect and discard duplicates as early as possible, respect DRM issues, and enrich the raw media asset with the maximum amount of metadata that could be found inside or around it.

2.2 Challenge 2: Content Normalization

In textual search engines, context is subjected to a pipeline of operations for preparing it to be indexed [3]; such pre-processing includes parsing, tokenization, lemmatization, and stemming. With text, the elements of the index are of the same nature of the constitutive elements of content: *words*. Multimedia content needs a more sophisticated pre-processing phase, because the elements to be indexed (called "features" or "annotations") are numerical and textual metadata that need to be extracted from raw content by means of complex algorithms.

The processing pipeline for multimedia data is therefore longer than in text search engines, and can be roughly divided in two macro steps: content normalization (treated in this Section) and content analysis (treated in the next Section).

Due to the variety of multimedia encoding formats, prior to processing content for metadata extraction, it is necessary to submit it to a normalization step, with a twofold purpose: 1) translating the source media items represented in different native formats into a common representation format (e.g., MPEG4 [49] for video files), for easing the development and execution of the metadata extraction algorithms; 2) producing alternative variants of native content items, e.g., to provide freebies (free sample copies) of copyrighted elements or low resolution copies for distribution on mobile or low-bandwidth delivery channels (e.g., making a 3GP version [70] of video files for mobile phone fruition). The challenge here is to devise the best encoding format for addressing the needs of analysis algorithm and easing the delivery of content at variable quality, without exploding the number of versions of the same item to be stored in the search engine.

2.3 Challenge 3: Content Analysis and Indexing

After the normalization step, a multimedia collection has to be processed in order to make the knowledge embedded in it available for querying, which requires building the internal indexes of the search engine. Indexes are a concise representation of the content of an object collection, constructed out of the features extracted from it; the features used to build the indexes must be both sufficiently representative of the content and compact to optimize storage and retrieval.

Features are traditionally grouped into two categories:

- *Low level features*: concisely describe physical or perceptual properties of a media element (e.g., the colour or edge histogram of an image).
- *High level features*: domain concepts characterizing the content (e.g., extracted objects and their properties, geographical references, etc.).

As in text, where the retrieved keywords can be highlighted in the source document, also in MIR there is the need of locating the occurrences of matches between the user's query and the content. Such requirement implies that features must be extracted from a time continuous medium, and that the coordinates in space and time of their occurrence must be extracted as well (e.g., the time stamp at which a word occurs in a speech audio file, the bounding-box where an object is located in an image, or both pieces of information to denote the occurrence of an object in a video).

Feature detection may even require a change of medium with respect to the original file, e.g., the speech-to-text transcription.

Content analysis and indexing are the prominent research problem of MIR, as the quality of the search engine depends on the precision at which the extracted metadata describe the content of a media asset: after introducing the global scheme of the content analysis process in Section 3.1, we devote Section 4 to the various ways in which features (also called metadata) can be represented and Section 5 to the algorithms for computing them.

2.4 Challenge 4: Content Querying

Text IR starts from a user's query, formulated as a set of keywords, possibly connected by logical operators (AND, OR, NOT). The semantics of query processing is text similarity: both the text files and the query are represented into a common logical model (e.g., the word vector model [64]), which supports some form of similarity measure (e.g., cosine similarity between word vectors).

In MIR, the expression of the user's information need allows for alternative query representation formats and matching semantics. Examples of queries can be:

- *Textual*: one or more keywords, to be matched against textual metadata extracted from multimedia content.
- *Mono-media*: a content sample in a single media (e.g., an image, a piece of audio) to be matched against an item of the same kind (e.g., query by music or image similarity, query by humming) or of a different medium (e.g., finding the movies whose soundtrack is similar to an input audio file).
- *Multi-media*: a content sample in a composite medium, e.g., a video file to be matched using audio similarity, image similarity, or a combination of both.

Accepting in input queries expressed by means of non-textual samples requires real-time content analysis capability, which poses severe scalability requirements on MIR architectures. Another implication of non-textual queries is the need for the MIR architecture to coordinate query processing across multiple dedicated search engines: for example, an image similarity query may be responded by coordinating an image similarity search engine specialized in low-level features matching and a text search engine, matching high-level concepts extracted from the query (e.g., object names, music gender, etc).

The grand challenge of MIR query processing is in part the same as for textual IR: retrieving the media objects more relevant to the user's query with high precision and recall. MIR adds the specific problem of content-based queries, which demand suitable architectures for analysing a query content sample on the fly and matching its features to those stored in the indexes. We devote Section 3.2 to a brief overview of the query process.

2.5 Challenge 5: Content Browsing

Unlike data retrieval queries (such as SQL or XPATH queries), IR queries are approximate and thus results are presented in order of relevance, and often in a number that exceeds the user's possibility of selection. Typically, a text search engine summarizes and pages the ranked results, so that the user can quickly understand the most relevant items.

In MIR applications, understanding if a content element is relevant poses additional challenges. On one side, content summarization is still an open problem [5]: for example, a video may be summarized in several alternative ways: by means of textual metadata, with a selection of key frames, with a preview (e.g., the first 10

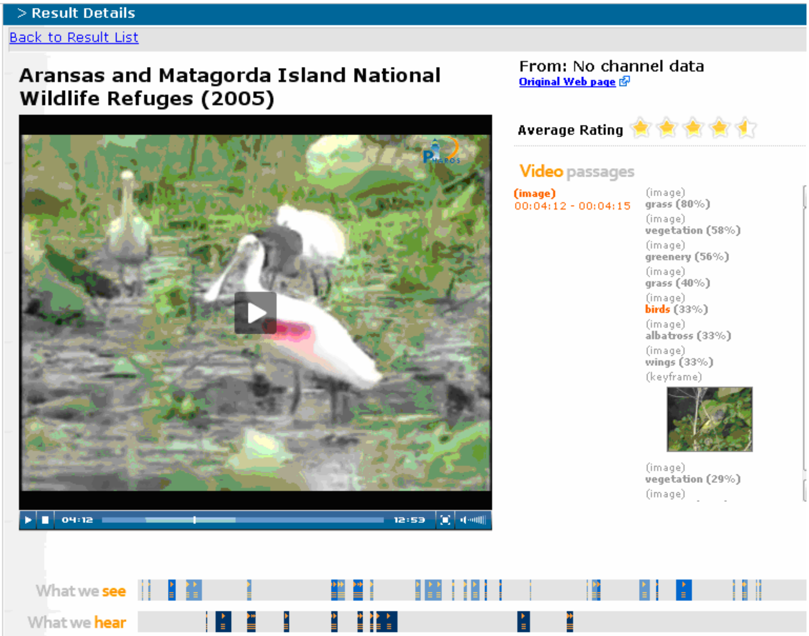

Fig. 1. Visual and aural time bars in the interface of the PHAROS search platform [10]

seconds), or even by means of another correlated item (e.g., the free trailer of a copyrighted feature film). The interface must also permit users to quickly inspect continuous media and locate the exact point where a match has occurred. This can be done in many ways, e.g., by means of annotated time bars that permit one to jump into a video where a match occurs, with VCR-like commands, and so on.

Figure 1 shows a portion of the user interface of the PHAROS multimedia search platform [10] for accessing video results of a query: two time bars (labelled "what we hear", "what we see") allow one to locate the instant where the matches for a query occur in the video frames and in the audio, inspect the metadata that support the match, and jump directly to the point of interest.

The challenge of MIR interfaces is devising effective renditions (visual, but also aural) that could convey both the global characteristics of the result set (e.g., the similarity distribution across a result collection) and the local features of an individual result item that justify the query match.

3 The MIR Architecture

The architecture of a MIR system [9] can be described as a platform for composing, verifying, and executing search processes, defined as complex workflows made of atomic blocks, called search services, as illustrated in Figure 2. At the core of the architecture there is a *Process Execution Engine* which is a runtime environment, optimized for the scalable enactment of data-intensive and computation-intensive workflows made of search services. A *search service* is a wrapper for any software component that embodies functionality relevant to a MIR solution.

The most important categories of MIR workflows are *Content Processe*s, which have the objective of acquiring multimedia content from external sources (e.g. from the user or a from video portal) and extracting features from it; and the *Query Processes*, which have the objective of acquiring a user's information need and computing the best possible answer to it. Accordingly, the most important categories of search services are *content services*, which embody functionality relevant to content acquisition, analysis, enrichment, and adaptation; and *query services*, which implements all the steps for answering a query and computing the ranked list of results.

Examples of content services can be: algorithms for extracting knowledge from media elements, transducers for modifying the encoding format of media files; examples of query services, instead, are: query disambiguation services for inferring the meaning of ambiguous information needs, or social network analysis services for inferring the preferences of a user and personalizing the results of a user's query.[1]

[1] In Figure 2 metadata are given in output to the content owner. They enrich the processed content and thus increase its value, and thus can be used by the content owner for publishing purposes or for building a separate query processing solution.

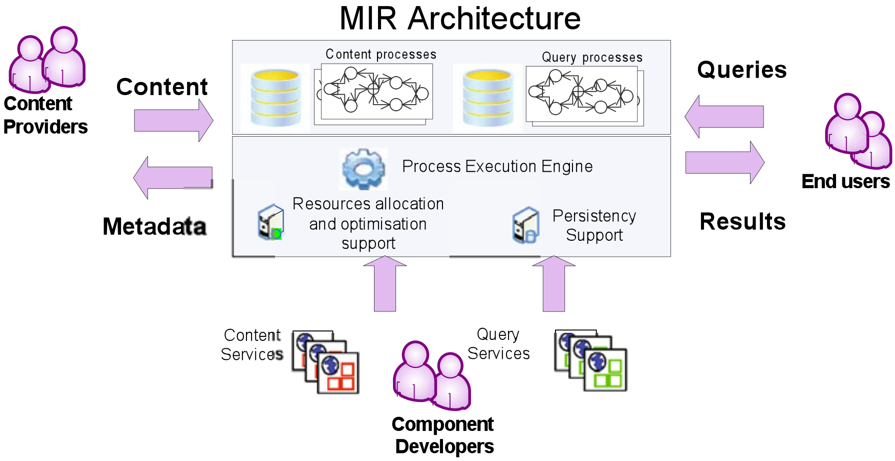

Fig. 2. Reference architecture of a MIR system

3.1 The Content Process

A content process (as the one schematized in Figure 3) aims at gathering multimedia content and at elaborating it to make it ready for information retrieval. A MIR platform may host multiple content processes, as required for elaborating content of different nature, in different domains, for different access devices, for different business goals, etc.

The input to the process is twofold:

- Multimedia Content (image, audio, video).
- Information about the content, which may include *publication metadata* (HTML, podcast [55], RSS [62], MediaRSS [45], MPEG7, etc), *quality information* (encoding, user's rating, owner's ratings, classification data), *access rights* (DRM data, licensing, user's subscriptions), and *network information* (type and capacity of the link between the MIR platform and the content source site - e.g., access can be local disk-based, remote though a LAN/SAN, a fixed WAN, a wireless WAN, etc).

The output of the process is the textual representation of the metadata that capture the knowledge automatically extracted from the multimedia content via content processing operations. The calculated metadata are integrated with the metadata gathered by the content acquisition system (shown as an input in Figure 3), which are typically added to the content manually by the owner or by the Web users (e.g., as tags, comments, closed captions, and so on). Section 4 and Section 5 respectively provide a discussion on the state of the art of metadata vocabularies and analysis techniques for extracting metadata from multimedia assets.

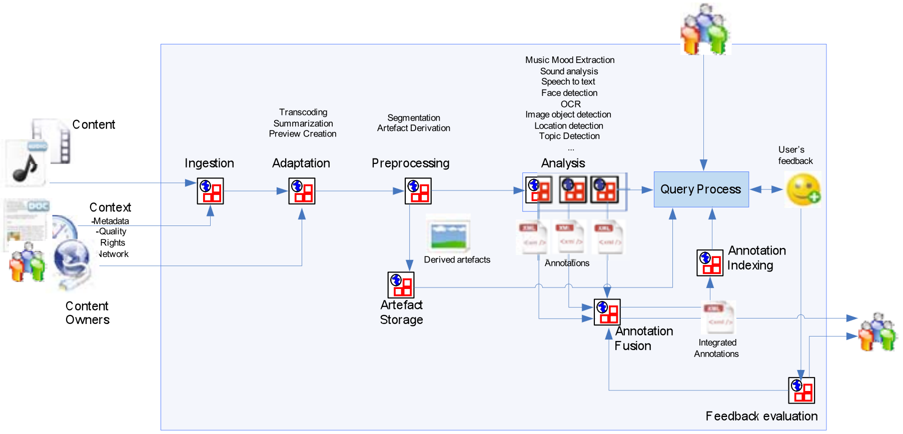

Fig. 3. Example of a MIR Content Process

A content process can be designed so as to dynamically adapt to the external context, e.g., as follows:

- By analyzing the content metadata (e.g., manual annotations) to dynamically decide the specific analysis operators to apply to a media element (e.g., if the collection denotes indoor content, a heuristic rule may decide to skip the execution of outdoor object detection).
- By analyzing the access rights metadata to decide the derived artefacts to extract (e.g., if content has limited access, it may be summarized in a freebie version for preview)
- By analyzing the geographical region where the content comes from (e.g., inferring the location of the publisher may allow the process to apply better heuristic rules for detecting the language of the speech and call the proper speech-to-text transcription module).
- By understanding the content delivery modality (e.g., a real-time stream of a live event may be indexed with a faster, even if less precise, process for reducing the time-to-search delay interval).

3.2 The Query Process

A MIR Query process (like the one schematized in Figure 4) accepts in input information need and formulates the best possible answer from the content indexed in the MIR platform.

The input of the query process is an information need, which can be a keyword or a content sample. The output is a result set, which contains information on the objects (typically content elements) that match the input query. The description of the objects in the result set can be enriched with metadata coming from sources external to the MIR platform (e.g., additional metadata on a movie taken from IMDB, or a map showing the position of the object taken from a Geographical Information System).

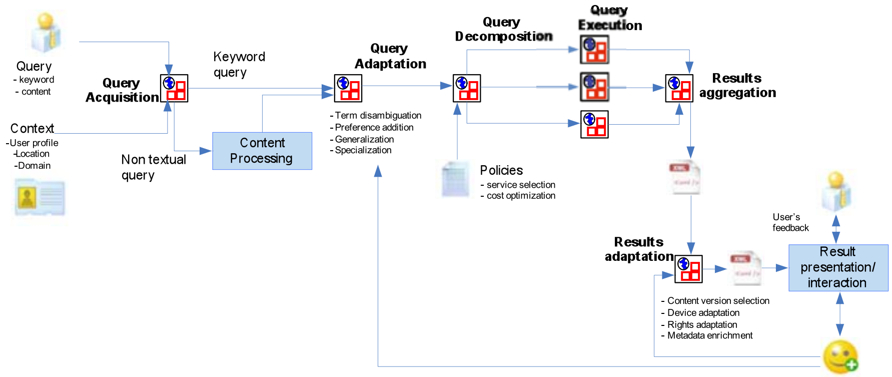

Fig. 4. Example of a MIR Query Process

A collateral source of input is the query context, which expresses additional circumstances about the information need, often implicit. Well-known examples of query context are: user preferences, past users' queries and their responses, access device, location, access rights, and so on. The query context is used to adapt the query process: for instance, it can be used to expand the original information need of the user with additional keywords reflecting her preferences, to disambiguate a query term based on the application domain where the query process is embedded, or to provide the best shape of results for the current user.

Queries are classified as **mono-modal**, if they are represented in a single medium (e.g., a text keyword, a music fragment, an image) or **multi-modal**, if they are represented in more than on medium (e.g., a keyword AND an image). As for Search Computing, also in MIR queries can be classified as **mono-domain**, if they are addressed to a single search engine (e.g., a general purpose image search engine like Google Images [26] or a special purpose search service as Empora [20] garments search), or **multi-domain**, if they target different, independent search services (e.g., a face search service like Facesaerch [23] and a video search service like Blinkx [7]). Table 1 exemplifies the domain and mode classification of queries.

Table 1. Examples of Mono/Multi modal, Mono/Multi domain queries in a MIR system

	Mono Domain	Multi Domain
Mono Modal	Find all the results that match a given keyword; Find all the images similar to a given image.	Find theatres playing movies acted by an actor having the voice similar to a given one.
Multi Modal	Find all videos that contain a given keyword and that contain a person with a face similar to a given one.	Find all CDs in Amazon with a cover similar to a given image.

4 Metadata

The content process produces a description of the knowledge extracted from the media assets, possibly integrated with information gathered during the content acquisition phase. This articulated knowledge must be represented by means of a suitable formalism: the current state of the practice in content management presents a number of metadata vocabularies dealing with the description of multimedia content [24]. Many vocabularies allow the description of high-level (e.g., title, description) or low-level features (e.g., colour histogram, file format), while some enable the representation of administrative information (e.g., copyright management, authors, date). In a MIR system, the adoption of a specific metadata vocabulary depends on its intended usage, especially for what concerns the type of content to describe. In the following we illustrate a few relevant and diverse examples that cover the main metadata format categories.

MPEG-7 [42] is an XML vocabulary that represents the attempt from ISO to standardize a core set of audio-visual features and structures of descriptors and their (spatial/temporal) relationships. By trying to abstract from all the possible application domains, MPEG-7 results in an elaborate and complex standard that merges both high-level and low-level features, with multiple ways of structuring annotations. MPEG7 is also extensible, so to allow the definition of application-based or domain-based metadata.

Dublin Core [19] is a 15-element metadata vocabulary (created by domain experts in the field of digital libraries) intended to facilitate discovery of electronic resources, with no fundamental restriction on the resource type. Dublin Core holds just a small set of high-level metadata and relations (e.g. title, creator, language, etc…), but its simplicity made it a common annotation scheme across different domains. It can be encoded using different concrete syntaxes, e.g., in plain text, XML or RDF.

MXF (Material Exchange Format) [17] is an open file format that wraps video, audio, and other bit streams (called "essences"), aimed at the interchange of audio-visual material, along with associated data and metadata, in devices ranging from cameras and video recorders to computer systems for various applications used in the television production chain. MXF metadata address both high-level and administrative information, like the file structure, key words or titles, subtitles, editing notes, location, etc. Though it offers a complete vocabulary, MXF has been intended primarily as an exchange format for audio and video rather than a description format for metadata storage and retrieval.

Exchangeable Image File Format (EXIF) [31] is a vocabulary adopted by digital camera manufacturers to encode high-level metadata like date and time information, the image title and description, the camera settings (e.g., exposure time, flash), the image data structure (e.g., height, width, resolution), a preview thumbnail, etc. By being embedded in picture raw contents, EXIF metadata is now a de-facto standard for image management software; to support extensibility, EXIF enables the definition of custom, manufacturer-dependent additional terms.

ID3 [32] is a tagging system that enriches audio files by embedding metadata information. ID3 includes a big set of high-level (such as title, artist, album, genre) and administrative information (e.g. the license, ownership, recording dates), but a very small set of low-level information (e.g. BPM). ID3 is a worldwide standard for audio metadata, adopted in a wide set of applications and hardware devices. However, ID3 vocabulary is fixed, thus hindering its extensibility and usage as format for low-level features.

Other examples of multimedia-specific metadata formats are SMEF [67] for video, IPTC [34] for images, and MusicXML [61] for music. In addition, several communities created some domain-specific vocabularies like LSCOM [39] for visual concepts, IEEE LOM [36] for educational resources, and NewsML [34] for news objects.

5 Techniques for Content Processing

The *content process* described in Figure 3 aims at creating a representation of the multimedia collection suitable for indexing and retrieval purposes. The techniques applied for analysing content are application dependent, and relate both with the nature of the processed items and with the aim of the applications. The content process is not exclusive of MIR, but also applies classical text-based IR systems. The processing of text is a well-understood activity which embodies a standard sequence of operations (language detection, spell checking, correction and variant resolution, lemmatization, and stop-word removal), which convert documents into a canonical format for a more efficient indexation [3].

MIR systems deal with more complex media formats, like audio, video and images, and therefore require a more articulated analysis process to produce the metadata needed for indexing. In essence, a MIR content process can be seen as an acyclic graph of operators, in which each operator extracts different features from a media item, possibly with the help of the metadata previously extracted by other already executed operators. The various operators embody diverse algorithms for content analysis and feature extraction, which are the subjects of the research challenges briefly introduced in Section 2.3.

The operations that constitute the MIR content process can be roughly classified in three macro categories: *transformation, feature extraction,* and *classification*, based on the stage at which they occur in the analysis process and on the abstraction level of the information they extract from the raw content:

- **Transformation:** this kind of operation converts the format of media items, for making the subsequent analysis steps more efficient or effective. For instance, a video transformer can modify an MPEG2 movie file to a format more suitable for the adopted analysis technologies (e.g., MPEG); likewise, an audio converter can transform music tracks encoded in MP3 to WAV, to eliminate compression and make content analysis simpler and more accurate.
- **Feature Extraction:** calculates low-level representations of media contents, i.e. feature vectors, in order to derive a compact, yet descriptive, representation of a pattern of interest [14]. Such representation can be used to enable content based search, or as input for classification tasks. Examples of visual features for images are *colour*, *texture, shape,* etc. [28]; examples of aural features for music contents are *loudness*, *pitch*, *tone (brightness and bandwidth)*, *Mel-filtered Cepstral Coefficients*, etc. [76].
- **Classification:** assigns conceptual labels to content elements by analyzing their raw features; the techniques required to perform this operations are commonly known as machine learning. For instance, an image classifier can assign to image files annotations expressing the subject of the pictures (e.g., mountains, city, sky, sea, people, etc.), while an audio file can be analyzed in order to discriminate segments containing speech from the ones containing music.

Table 2. Content analysis techniques in MIR systems

Audio Analysis	Image Analysis	Video Analysis
Audio segmentation [44]: to split audio track according to the nature of its content. For instance, a file can be segment according to the presence of noise, music, speech, etc.	*Semantic Concept extraction* [38]: the process of associating high-level concepts (like *sky*, *ground*, *water*, *buildings*, etc.) to pictures.	*Scene detection* [59]: detection of scenes in a video clip; a scene is one of the subdivisions of a play in which the setting is fixed, or that presents continuous action in one place [60].
Audio event identification [57]: to identify the presence of events like gunshots and scream in an audio track.	*Optical character recognition* [6]: to translate *images* of handwritten, typewritten or printed text into an editable text.	*Video text detection and segmentation* [50]: to detect and segment text in videos in order to apply image OCR techniques.
Music genre (mood) identification [12] [46]: to identify the genre (e.g., rock, pop, jazz, etc.) or the mood of a song.	*Face recognition and identification* [75] [82]: to recognize the presence of a human face in an image, possibly identifying its owner.	*Video summarization* [5]: to create a shorter version of a video by picking important segments from the original.
Speech recognition [21]: to convert words spoken in an audio file into text. Speech recognition is often associated with Speaker identification [51], that is to assign an input speech signal to one person of a known *group*	*Object detection and identification* [80]: to detect and possibly identify the presence of a known object in the picture.	*Shot detection* [15]: detection of transitions between shots. Often shot detection is performed by means of Keyframe segmentation [13] algorithms that segment a video track according to the key frames produced by the compression algorithm.

Arbitrary combinations of transformation, feature extraction, and classification operations can result in several analysis algorithms. Table 2 presents a list of 14 typical audio, image, and analysis techniques; the list is not intended to be complete, but rather to give a glimpse on the analysis capabilities currently available for MIR systems. To provide the reader with a hook to the recent advancements in the respective fields, each analysis technique is referenced with a recent survey on the topic. The techniques shown in Table 2 can be used in isolation, to extract different features from an item. Since the corresponding algorithms are probabilistic, each extracted feature is associated with a confidence value that denotes the probability that item X contain feature Y. To increase the confidence in the detection, different analysis techniques can be used jointly to reinforce each other. Using the example of the movie file, the fact that in a single scene both the face and the voice of a person are identified as belonging to an actor "X" can be considered as a correlated event, so to describe the scene as "scene where actor X appears" with a high confidence. The cross reinforcement of analysis techniques is called *annotation fusion*: multiple features extracted from media are fused together to yield more robust classification detection [43]. For instance, multiple content segmentation techniques (e.g., shot detection and speaker's turn segmentation) can be combined in order to achieve better video splitting; voice identification and face identification techniques can be fused in order to obtain better person identification. Typically, the use of multiple techniques simultaneously can be computationally expensive, thus limiting this solution to such domains where accuracy in the content descriptions is more important than indexing speed.

6 Examples of MIR Query Languages

In MIR, a user's query is matched against the representation of content provided by one (or more) of the metadata formats described in Section 4. Given such a variety of data representations, there is not a standardized query language for MIR systems, as every retrieval framework provides its own proprietary solution. For such a reason, several proposals for a unified MIR query language have emerged in the last years, and this Section will provide an overview.

Given that multimedia objects are usually described textually, a natural choice for the query language is exploiting mature text retrieval techniques: for instance, free-text or keyword-based search, context queries, Boolean queries, pattern queries [3], or faceted queries [63].

Even if based on a conventional IR query languages, though, a query language for MIR must comply with additional requirements typical of aural and visual media types or of specific application domains [29]:

- *Schema independence*: given the multitude of metadata representation formats, a query language should not rely on a specific schema.
- *Arbitrary search scope granularity*: the query language must allow search of information both in the whole media object and in chunks thereof.
- *Media objects as query conditions*: a MIR query language should support content-based queries, in one of two ways: 1) providing a media object to use as a query condition and the information about the algorithm to use for its on-the-fly analysis; 2) providing a set of previously calculated low-level features to use as a query condition.
- *Arbitrary similarity measure*: the query language should enable the flexible representation of arbitrary ranking functions, so to suite application-specific needs.

The last two decades have witnessed to a lot of efforts in the definition of more expressive and structured query languages, designed specifically for multimedia retrieval. Among the most recent efforts, POQL^{MM} [30], is a general purpose query language for object oriented multimedia databases exposing arbitrary data schema. MuSQL [78] is a music structured query language, composed of a schema definition sub-language and a data manipulation sub-language. In [35], authors propose a query language for video retrieval enabling queries at both image and semantic levels, for retrieving videos with both exact matching and similarity matching.

One of the latest attempts in providing a unified language for MIR is represented by the MPEG Query Format [1]. MPQF is part of the MPEG-7 standard, and provides a standardized interface for MIR systems based on the XML metadata representation formats. MP7QF derives from the well-known set of

XML-based query languages (e.g., *XPath* and *XQuery*), from which it inherits both the syntax and the semantics. MP7QF provides a rich set of multimedia query types (e.g., *QueryByMedia*, *QueryByDescription*, *QueryByFreeText*, *SpatialQuery*, *TemporalQuery*, *QueryByXQuery*, etc.), and specifies a set of precise output parameters describing the response of a multimedia query request by allowing the definition of the content as well as structure of the result set. MP7QF can also be extended with novel query operators. For instance, in [79] authors introduce the *SpatioTemporalOperator*.

7 Examples of Research and Commercial MIR Solutions

In this Section we overview a number of research projects that have prototyped the architecture and techniques of a MIR solution, as well as a sample of commercial systems that enable querying multimedia content.

7.1 European and Regional Research Projects

PHAROS [10] is an Integrated Project of the Sixth Framework Program (FP6) of the European Community. PHAROS has developed an extensible platform for MIR, based on the automatic annotation of multimedia content of different nature: audio, images and video. PHAROS content annotation process has a plug-in architecture: the content process can be defined (with a proprietary tool) and deployed in a distributed manner, possibly incorporating external components, invoked as web services. On top of the PHAROS platform two showcase applications, one for fixed Internet and one for mobile networks, have been prototyped.

VITALAS [18] is an FP6 Research Project which has implemented a prototype system for the intelligent access to multimedia professional archives. VITALAS was conceived as a B2B tool to develop and validate technologies applicable to large consumer-facing MIR search engines. The main objective is to enable scalable cross-media indexing and retrieval, as well as methods for content aggregation through the automatic extraction of metadata. VITALAS has produced a prototype implementation of automatic annotation algorithms, visual interfaces for searching in large audio-visual archives, and search personalization techniques.

THESEUS [73] is an ongoing German research program aimed at developing a new Internet-based infrastructure to better exploit the knowledge available on the Internet. To this end, application-oriented basic technologies and technical standards are being developed and tested. For instance, the THESEUS project created and supports the Open Source project SMILA [66] (Semantic Information Logistics Architecture), a reliable, standardized industrial strength enterprise framework for building searches solutions to various kinds of information (i.e. accessing unstructured information). Since June 2008, SMILA is an official project of the Eclipse Foundation.

Quaero [74] is a French collaborative research and development program that aims at developing multimedia and multilingual indexing, processing, and management tools to build general public search applications on large collections of multimedia information (multilingual audio, video, text, etc.). The challenge of Quaero is to integrate search and indexing components with audio/images/video processing techniques, semantic annotation methodologies and automatic machine translation technologies, with a specific focus on improving the quality and relevance of these later technologies and techniques.

7.2 Examples of Commercial MIR Systems

Midomi [47] is an example of audio processing technologies applied to music search engine. The interface allows users to upload voice recordings of public songs, and then to query such music files by humming or whistling. Another similar application is **Shazam** [65]. Shazam is a commercial music search engine that enables users to identify tunes using their mobile phone. The principle consists in using its mobile phone to record a sample of few seconds of a song from any source (even with bad sound quality) and the system returns the identified song with the necessary details: artist, title, album, etc. Similar systems for music search are also provided by BMat [8]. **Voxalead**™ [77] is an audio search technology demonstrator implemented by Exalead to search in TV news, radio news, and VOD programs by content. The system uses a third-party speech-to-text transcription module transcribe political speeches in several languages.

The field of image search technologies also appears to be mature. **Google Images** [26] and **Microsoft Bing** [48], for instance, now offer a *”show similar images”* functionality, thus proving the scalability of content-based image search on the Web. Other notable examples of image MIR engine are **Tiltomo** [68], which also performs search according to the image theme, and **SAPIR** [33], a search engine developed within the homonymous EU-founded project which also provide geographic and video search. **Blinkx** [7] is another example of search engine on videos and audios streams. Blinkx, like Voxalead™, uses speech recognition to match the text query to the video or audio speech content. Blinkx represents an example of mature video MIR solution as they claim to have over 30 million hours of video indexed.

8 Conclusion and Perspectives

In this chapter we presented the problem of Multimedia Information Retrieval, highlighting its challenges, major technical issues, and application areas, and we proposed a reference architecture unifying the aspects of content processing and querying. Then, we provided a survey on the existing research and commercial solution for multimedia search, showing the existence of several mature MIR technologies, products, and services.

In a context where the production of content has become massive thanks to the availability of cheap and high-quality recording devices, MIR solutions represent a fundamental tool for the access to content collections.

MIR systems could benefit from the Search Computing approach in various ways:

- At the architecture level, a MIR system is often implemented on top of a set of distributed Web services, each specialized in a different content analysis and/or query processing technique. In such a distributed scenario, MIR query processing resembles the computation of a multi-domain query: the query “find news clips where President Obama discusses the Health Insurance

Reform could be resolved by joining the results, ranked by confidence, of two metadata sources, one capable of locating the face of President Obama in a video and one able to process the text transcriptions to identify the topic of a discussion.

- At the user interface level, the Liquid Query paradigm could be used to enable the exploration of large multimedia collection. The user could start with a focused query (e.g., a keyword query on the text transcripts of news clips) and then expand the query by joining other metadata sources, exposed as service interfaces: e.g., looking for other videos featuring the same speaker, or produced by the same media agency.

On the other hand, MIR can also extend the capabilities of Search Computing systems by enabling new ways of matching and ranking in multi-domain queries. For instance:

- **MIR Systems as Domain-Specific Search Engines:** MIR systems can be adopted in Search Computing as a special category of ranked search services: for instance, in a multi-domain query for a market analysis application, a text transcription search engine could be wrapped as a service interface for selecting and ranking news clips according to the probability that they deal with a given company, provided in input as a keyword.
- **MIR Operations as Query Operators:** including non-textual content items as query conditions can enrich the expressive power of Search Computing systems. For instance, users can express their information need as images or audio files, leaving to the analysis and annotation operations the task of extracting, in a textual form, the concept to use as a join condition (e.g., the name of a person given an image of its face). Even more interestingly, a multi-domain query could directly exploit content similarity to compute joins: in a trip planning multi-domain query, the destinations could be joined to the result of the query based on their similarity to the user's favourite beach, supplied in input as an image.

References

[1] Adistambha, K., Döller, M., Tous, R., Gruhne, M., Sano, M., Tsinaraki, C., Christodoulakis, S., Yoon, K., Ritz, C., Burnett, I.: The MPEG-7 Query Format: A New Standard in Progress for Multimedia Query by Content. In: Proceedings of the 7th International IEEE Symposium on Communications and Information Technologies (ISCIT 2007), pp. 479–484 (2007)

[2] Adobe Premiere (2009), `http://www.adobe.com/products/premiere/`

[3] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval, 1st edn. Addison Wesley, Reading (1999)

[4] Baldi, A., Murace, R., Dragonetti, E., Manganaro, M., Guerra, O., Bizzi, S., Galli, L.: Definition of an automated Content-Based Image Retrieval (CBIR) system for the comparison of dermoscopic images of pigmented skin lesions. Biomed. Eng. Online (2009)

[5] Barbieri, M., Agnihotri, L., Dimitrova, N.: Internet Multimedia Management Systems IV. In: Proceedings of the SPIE, vol. 5242, pp. 1–13 (2003)

[6] Beitzel, S.M., Jensen, E.C., Grossman, D.A.: Retrieving OCR Text: A Survey of Current Approaches. In: Symposium on Document Image Understanding Technologies, SDUIT (2003)

[7] Blinkx – Video Search Engine (2009), http://www.blinkx.com/

[8] BMat - 2009 (2009), http://www.bmat.com/

[9] Bozzon, A., Brambilla, M., Fraternali, P.: Model-Driven Design of Audiovisual Indexing Processes for Search-Based Applications. In: 7th IEEE International Workshop on Content-Based Multimedia Indexing, pp. 120–125. IEEE Press, New York (2009)

[10] Bozzon, A., Brambilla, M., Fraternali, P., Nucci, F., Debald, S., Moore, E., Neidl, W., Plu, M., Aichroth, P., Pihlajamaa, O., Laurier, C., Zagorac, S., Backfried, G., Weinland, D., Croce, V.: Pharos: an audiovisual search platform. In: Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in information Retrieval, SIGIR 2009, Boston, MA, USA, July 19 - 23, p. 841. ACM, New York (2009)

[11] Buchenwald demonstrator, University of Twente (2009), http://vuurvink.ewi.utwente.nl:8080/Buchenwald/

[12] Caringella, N., Zoia, G., Mlynek, D.: Automatic genre classification of music content: a survey. IEEE Signal Processing Magazine 23(2), 133–141 (2006)

[13] Carrato, K.S.: Temporal video segmentation: a survey. Signal Processing: Image Communication 16, 477–500 (2001)

[14] Cees, G.M.: Concept-Based Video Retrieval. Foundations and Trends in Information Retrieval 4(2), 215–322 (2009)

[15] Cotsaces, C., Nikolaidis, N., Pitas, I.: Video Shot Boundary Detection and Condensed Representation: A Review. IEEE Signal Processing Magazine (2006)

[16] Delve Networks - Online Video Platform and Content Management (2009), http://www.delvenetworks.com/

[17] Devlin, B., Wilkinson, J.: The Material Exchange Format. In: Gilmer, B. (Hrsg.) File Interchange Handbook, pp. 123–176. Elsevier Inc., Focal Press (2004)

[18] Diou, C., Papachristou, C., Panagiotopoulos, P., Stephanopoulos, G., Dimitriou, N., Delopoulos, A., Rode, H., Aly, R., de Vries, A.P., Tsikrika, T.: VITALAS at TRECVID 2008. In: Proceedings of the 6th TREC Video Retrieval Evaluation Workshop, Gaithersburg, USA, November 17-18 (2008)

[19] Dublin Core Metadata Initiative (2009), http://dublincore.org/

[20] Empora Online Shop (2009), http://www.empora.com

[21] Eu, H., Hedge, A.: Survey of continuous speech recognition software usability. Cornell University, Ithaca, NY (1999), http://ergo.human.cornell.edu/AHProjects/Hsin99/Voice%20Recognition%Paper.pdf (retrieved April 5, 2004)

[22] Eyealike platform for facial similarity (2009), http://www.eyealike.com/home

[23] Facesaerch (2009), http://www.facesaerch.com/

[24] Geurts, J., van Ossenbruggen, J., Hardman, L.: Requirements for practical multimedia annotation. In: Workshop on Multimedia and the Semantic Web Heraklion, Crete, pp. 4–11 (2005)

[25] Google Election Video Search (2009), http://googleblog.blogspot.com/2008/07/in-their-own-words-political-videos.html

[26] Google Images (2009), http://images.google.com

[27] Google Picasa (2009), http://picasa.google.com/

[28] Hanbury, A.: A survey of methods for image annotation. Journal of Visual Languages and Computing 19(5), 617–627 (2008)

[29] Henrich, A., Robbert, G.: Combining multimedia retrieval and text retrieval to search structured documents in digital libraries. In: Proc. 1st DELOS Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland, vol. 01/W001 (2000)

[30] Henrich, A., Robbert, G.: POQLMM: A Query Language for Structured Multimedia Documents. In: Proceedings of the First International Workshop on Multimedia Data and Document Engineering, Lyon, France, pp. 17–26 (2001)

[31] Japan Electronics and Information Technology Industries Association: Exchangeable image file format for digital still cameras: EXIF. Version 2.2 (2002)

[32] ID3 (2009), http://www.id3.org/

[33] IST SAPIR Large Scale Multimedia Search and P2P (2009), http://sapir.isti.cnr.it/index

[34] International Press Telecommunications Council (2009), http://www.iptc.org/IPTC4XMP/

[35] Le, T.H., Thonnat, M., Boucher, A., Bremond, F.: A Query Language Combining Object Features and Semantic Events for Surveillance Video Retrieval. In: Proceedings of Advances in Multimedia Modeling, 14th MMM Conference, Kyoto, Japan, pp. 307–317 (2008)

[36] Learning Object Metadata (2009), http://ltsc.ieee.org/wg12/

[37] Lew, M., et al.: Content-Based Multimedia Information Retrieval: State of the Art and Challenges. ACM Transactions on Multimedia Computing, Communications, and Applications 2(1) (2006)

[38] Liu, Y., Zhang, D., Lu, G., Ma, W.: A survey of content-based image retrieval with high-level semantics. Pattern Recogn. 40(1), 262–282 (2007)

[39] LSCOM Lexicon Definitions and Annotations (2009), http://www.ee.columbia.edu/ln/dvmm/lscom/

[40] LTU technologies (2009), http://www.ltutech.com/

[41] Martínez, J.M.: MPEG-7 Overview (version 10), ISO/IEC JTC1/SC29/WG11N6828, Palma de Mallorca (2004)

[42] Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7: Multimedia Content Description Interface, 396 p. Wiley, Chichester (2002)

[43] Maragos, P., Potamianos, A., Gros, P.: Multimodal Processing and Interaction, Audio, Video, Text. Multimedia Systems and Applications, vol. 33. Springer, Heidelberg (2008)

[44] Marsden, A., Mackenzie, A., Lindsay, A.: Tools for Searching, Annotation and Analysis of Speech, Music, Film and Video; A Survey. Literary and Linguistic Computing 22(4), 469–488 (2007)

[45] Media RSS (2009), http://en.wikipedia.org/wiki/Media_RSS

[46] Meyers, O.C.: A Mood-Based Music Classification and. Exploration System, MS Thesis, Massachusetts Institute of. Technology (MIT), USA (2007)

[47] MiDoMi (2009), http://www.midomi.com/

[48] Microsoft Bing (2009), http://www.bing.com/images

[49] MPEG Industry Forum (2009), http://www.m4if.org/

[50] Ngo, C., Chan, C.: Video text detection and segmentation for optical character recognition. Multimedia Systems 10(3), 261–272 (2004)

[51] Petrovska-Delacrétaz, D., El Hannani, A., Chollet, G.: Text-Independent Speaker Verification: State of the Art and Challenges. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) COST 277. LNCS, vol. 4391, pp. 135–169. Springer, Heidelberg (2007)

[52] Pictron Solutions (2009), http://www.pictron.com/

[53] Pixta, Visual search technologies (2009), http://www.pixsta.com/

[54] Pluggd Podcast Search Engine (2009), http://www.pluggd.tv/

[55] Podcast (2009), http://en.wikipedia.org/wiki/Podcasting

[56] Podscope Podcast Search Engine (2009), http://www.podscope.com/

[57] Potamitis, I., Ganchev, T.: Generalized recognition of sound events: Approaches and applications. Studies in Computational Intelligence, vol. 120, pp. 41–79. Springer, Heidelberg (2008)

[58] Radio Oranje speech search, Univeristy of Twente (2009), http://hmi.ewi.utwente.nl/choral/radiooranje.html

[59] Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: a systematic survey. IEEE Transactions on Image Processing 14(3), 294 (2005)

[60] Rasheed, Z., Shah, M.: Scene detection in Hollywood movies and TV shows. In: Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (2003)

[61] Recordare (2009), http://www.recordare.com/xml.html

[62] RSS (2009), http://en.wikipedia.org/wiki/RSS_file_format

[63] Sacco, S.M., Tzitzikas, Y.: Dynamic Taxonomies and Faceted Search, Theory, Practice, and Experience. The Information Retrieval Series, vol. 25, p. 340. Springer, Heidelberg (2009)

[64] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (November1975) (2003)

[65] SHAZAM (2009), http://www.shazam.com/

[66] SMILA (2009), http://www.eclipse.org/smila/

[67] SMEF (2009), http://www.bbc.co.uk/guidelines/smef/

[68] TILTOMO (2009), http://www.tiltomo.com/

[69] Tineye, Image Search Engine (2009), http://tineye.com/

[70] The 3GP video standard (2009), http://www.3gp.com/

[71] The DAML Ontology Library (2009), http://www.daml.org/ontologies

[72] The Internet Movie Database (2009), http://www.imdb.com

[73] The Theseus programme (2009), http://theseus-programm.de

[74] The Quaero Program (2009), http://www.quaero.org

[75] Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology 18(11), 1473–1488 (2008)

[76] Typke, R., Wiering, F., Veltkamp, R.C.: A survey of music information retrieval systems. In: ISMIR 2005, pp. 153–160 (2005)

[77] Voxalead$^{\mathrm{TM}}$ (2009), http://voxalead.labs.exalead.com

[78] Wang, C.C., Wang, J., Li, J., Sun, J.G., Shi, S.: MuSQL: A Music Structured Query Language. In: Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., Chia, L.-T. (eds.) MMM 2007. LNCS, vol. 4352, pp. 216–225. Springer, Heidelberg (2006)

[79] Wattamwar, S.S., Ghosh, H.: Spatio-temporal query for multimedia databases. In: Proceeding of the 2nd ACM Workshop on Multimedia Semantics (MS 2008), pp. 48–55. ACM, New York (2008)

[80] Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Survey 38(4) (2006)

[81] Yu, G., Chen, Y., Shih, K.: A Content-Based Image Retrieval System for Outdoor Ecology Learning: A Firefly Watching System. In: International Conference on Advanced Information Networking and Applications, vol. 2, p. 112 (2004)

[82] Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Comput. Surv. 35(4), 399–458 (2003)