Multimed Tools Appl (2011) 51:247–277
DOI 10.1007/s11042-010-0621-0

Personalization in multimedia retrieval: A survey

Yijuan Lu · Nicu Sebe · Ross Hytnen · Qi Tian

Published online: 3 November 2010

Abstract With the explosive broadcast of multimedia (text documents, image, video etc.) in our life, how to annotate, search, index, browse and relate various forms of information efficiently becomes more and more important. Combining these challenges by relating them to user preference and customization only complicates the matter further. The goal of this survey is to give an overview of the current situation in the branches of research that are involved in annotation, relation and presentation to a user by preference. This paper will present some current models and techniques being researched to model ontology, preference, context, and presentation and bring them together in a chain of ideas that leads from raw uninformed data to an actual usable user interface that adapts with user preference and customization.

Keywords Personalization · Information Access · Multimedia

Y. Lu · R. Hytnen
Department of Computer Science, Texas State University, San Marcos, TX 78666, USA

Y. Lu
e-mail: yl12@txstate.edu

R. Hytnen
e-mail: r.hytnen@gmail.com

N. Sebe
Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 14-38100 Povo, Trento, Italy
e-mail: sebe@disi.unitn.it

Q. Tian (✉)
Computer Science Department, University of Texas at San Antonio, San Antonio, TX 78249, USA
e-mail: qitian@cs.utsa.edu

1 Introduction

We are living in an era with explosion of multimedia content in databases, broadcasts, streaming media, internet, *etc.* The huge amount of multimedia data has generated new requirements for more effective access and more efficient search on these global information repositories especially based on users' preference. How to incorporate this preference in extracting content, indexing and retrieving multimedia data is becoming one of the most challenging and fastest-growing research areas. A consequence of the growing consumer demand for personalized multimedia information is that sophisticated technology is needed for representing, modeling, indexing, and retrieving multimedia data based on user's current need. In particular, we need robust techniques to index/retrieve and compress personalized multimedia information, new scalable browsing algorithms allowing access to very large multimedia databases, and interactive semantic visual interfaces integrating user's interest into unified multimedia browsing and retrieval systems.

The 1970's were dominated by the use of large mainframes, in the 80's computing power went to the user's desktop, and with the PC revolution, from the mid 90's the new frontier became the creation of a completely connected world. According to Merrill Lynch, from 2010 (peaking around 2030) we will enter the "Content Centric" era. In this vision, broadband networks will be pervasive and user personal content will be acquired, stored, and processed directly on the network [82]. Consider the following futuristic scenario:[1]

> John Citizen lives in Brussels, holds a degree in economics, and works for a multinational company dealing with oil imports. He enjoys traveling with emphasis on warm Mediterranean sites with good swimming and fishing. When watching TV his primary interest is international politics, particularly European. During a recent armed conflict he wanted to understand different perspectives of the war, including both relevant historical material as well as future projections from commentators. When he returns home from work, a personalized interactive multimedia program is ready for him, created automatically from various multimedia segments taken from diverse sources including multimedia news feeds, digital libraries, and collected analyst commentaries. The program includes different perspectives on the events, discussions, and analysis appropriate for a university graduate. The video program is selecting and summarizing video segments related to the war. Sections of the program allow him to interactively explore analyses of particular relevance to him, namely the impact of war on oil prices in various countries (his business interest), and its potential effect on tourism and accommodation prices across the Mediterranean next summer. Some presentations may be synchronized with a map display which may be accessed interactively. The audio program can select appropriate background music based on John's affective state. John's behavior and dialogue with the display are logged along with a record of the information presented to allow the system to better accumulate his state of knowledge and discern his interests in order to better serve him in the future. When John is away from home for business or leisure, he may receive the same personalized information on his mobile device as well, emphasizing information reflecting the neighborhood of his current Mediterranean location.

This "vision" reflects the strong need of personalized multimedia information retrieval. Facing a huge amount of multimedia information every day, users hope to read the news

[1] Courtesy to the FACS Consortium.

they are interested, watch the videos that are automatically selected and segmented for their need and listen to the music they like. This "vision" has many consequences from both societal and economic perspectives. Societal consequences impact personal information, continuous education, e-government, tourism, leisure, etc. The economic infrastructure of information provision and dissemination will also change, as new actors will enter the market (mainly providers of the new services to be offered), and existing actors will radically change their mode of operation (think about the new paradigm for the production of the TV programs of the future) and their business model.

In recent years, some multimedia systems have been built up to generate news, TV program, image, video *etc.* based on users' preference. Here, we briefly introduce some systems by their media types first. Then we discuss the common research topics shared by these systems in more details.

1.1 Webpage-based systems

WebMate [14] is a personal agent which provides personal browsing and searching service and using keywords to record user's preference. During users' browsing, the system uses the Trigger Pair Model to automatically extract keywords for refining document search. Hence, it learns the user interests incrementally and with continuous update and providing automatically documents (e.g. a personalized newspaper) that match the user interests [14].

SiteIF [59] is another personal agent for document recommendation. Different with WebMate, SiteIF uses sense-based document representation rather than word-based representation to build a model of the user's interests. By using WordNet and Word Domain Disambiguation technique, when the user browses the documents, SiteIF builds the user model as a semantic network whose nodes represent senses of the documents requested by the user. Then the filtering phase takes advantage of the word senses to retrieve new documents with high semantic relevance with respect to the user model [59]. Compared with word-based representation system, sense-based models are more accurate and are language independent. As blog becomes more and more popular, an online personalized blog reading system is proposed in [54]. This system first collects posts from the user's favorite blogs. Then similar posts related to the same topic will be clustered together. Finally, a personalized ranking algorithm will rank all the posts based on learned user's personal reading preferences and display to the user.

1.2 TV-based systems

PTVPlus [89] is an established online recommender system developed in the television listings domain. PTVPlus uses its recommendation engine to generate a set of TV program recommendations for a target user, based on their profiled interests, and it presents these recommendations in the form of a personalized program guide [89]. PTVPlus provides users with an interactive interface, from which users can browse, search, play and give feedback on recorded programs, for example, love, like, dislike etc. By collecting these user profiles including users' viewing behavior, playback history, implicit feedbacks, PTVPlus uses data mining techniques to discover similarity knowledge from collaborative filtering of user profiles and uses content-based reasoning (CBR) related methods to exploit this new knowledge during the recommendation process. PTVPlus attempts to develop a personalized TV recommendation system, which responds better to the personal preferences of individual users.

1.3 Image-based systems

Traditional content-based image retrieval (CBIR) systems typically use relevance feedback [111] to take consideration of human users' factor. After the retrieval results are displayed to the users, they can choose the images they like most or indicate relevance explicitly using a binary or graded relevance system. Then the search engine will perform a new search based on user's preference.

1.4 Video-based systems

Personalized video summarization is to find a condensed version of a full length video by identifying the most interesting and useful segments to the users. There are a number of different ways to generate personalized video summaries [6]; Tseng et al. [91] propose a system to filter video content before it is presented to the user via user profiles. Aizawa et al. [4, 34, 35] present a multimedia interactive browsing and summarizing system based on a user's life log. They use different sensors to detect and collect users' life log information, for example, a GPS receiver to identify user's location, physiological sensors to recognize users' brain wave and acceleration sensors to detect users' motion. The combined users' life log information is very helpful to summarize key frame images and generate video segment summaries that users are interested. Yu et al. [107] summarize video based on the users' individual browsing behavior and access patterns within an interactive customized browsing interface. With this interface, the viewer can get an overview of the video based on the key frames, select a shot to play, and jump to another shot at any time [107]. These users' behaviors are simulated with an ***Interest-guided Walk*** model (similar to Random Walk model for PageRank), and the probability of a shot being visited is taken as an indication of the interestingness and importance of that shot. By representing both Low-level features and user logs with virtual links among video objects and between users and video objects, a unified link analysis is applied to rank shots.

1.5 Other systems

MAGIC ("Metadata Automated Generation for Instructional Content") is developed by IBM to assist learning content authors and course developers in generating metadata for learning objects and information assets to enable wider reuse of these objects across departments and organizations [20]. MAGIC incorporates several software tools to analyze thousands of documents from internet, instructional videos, and other learning assets to provide a web-based user interface enabling authors to review and edit metadata at the time when the content is created [95].

P-Karaoke ("Personalized Karaoke") system is proposed by Microsoft Research Asia [37]. P-Karaoke is a real multimedia system, which consists of music, personal video and personal photo collection. Given a song, P-Karaoke automatically selects personal videos and photos according to their content, user's preferences or music and utilizes them as the background videos of the Karaoke [37]. PicToon is a personalized cartoon system, which can generate funny, lively, and artistic cartoon from personalized input images [15]. Microsoft Research Asia is also developing personalized eHealth system. It aims to build a user-centric & data-driven personalized eHealth software and services platform for improving quality of life and efficiency of treatment [21].

The above systems try to generate personalized news, TV, image, video, music, cartoon, and even health treatment. From the above vision and systems, several basic and important research topics may be derived:

- How to represent multimedia content descriptions?
- How to model user preferences and represent context?
- How to automatically annotate content?
- How to generate automated presentation authoring?
- How to perform distributed retrieval and filtering of content descriptions and user preferences?

We discuss in the following each of these research directions and we focus on a unified concept that seemingly integrates the user, his/her context, and his/her interests.

2 Personalized access

Considering the scenario introduced previously, what is needed is the development of a new modality for access to information of interest through automatic creation of interactive, mixed-media, personalized presentations on demand or prescribed. These presentations may be represented in the form of Intelligent multimedia Objects (or I-Objects)[2] that embed selected (e.g. novel and pertinent to the user's information need) video clips from multimedia documents and provide the tools that adapt the presentations automatically to the user requirements (explicitly or implicitly expressed) that range from content selection to content presentation and rendering.

Automatically synthesized interactive multimedia presentations should go beyond the simple collection of sequences of potentially relevant multimedia segments: I-Objects can generate presentations comparable to professionally edited and composed interactive multimedia programs. The creation of a personalized I-Object, as shown in Fig. 1 can be obtained through the following steps:

1. **Automated multimedia document annotation**. Multimedia documents are automatically annotated with metadata expressing the (i) semantic and (ii) affective/emotional/ rhetoric content of each video clip [32]. Semantic annotations identify the conceptual content of the document while affective annotations denote their perspective, e.g. documents may convey an attitude or a bias in their content by virtue of language, movement, juxtaposition, color, rhythm, etc. Affective annotations give information like "video sequence with dramatic presentation", "expresses negative opinion vehemently", "evocative music sequence", "visually stimulating picture", etc.
2. **Modeling of user preferences and acquisition of user profiles** Understanding the semantic value of a user preference is an ongoing challenge. There are some useful technologies for modeling this kind of data like MPEG-7/21 but they are currently insufficient to the task. Current technologies use some basic concepts such as historical usage and keywords to understand user preference. The profile and context of users are created and updated by observing their behavior when interacting with a presentation through a device, by logging a record of what they have seen, and by allowing them to explicitly express their preferences or modify their profiles.

[2] The idea of I-Objects and part of the description come from discussions and documents created during the FACS consortium interactions.

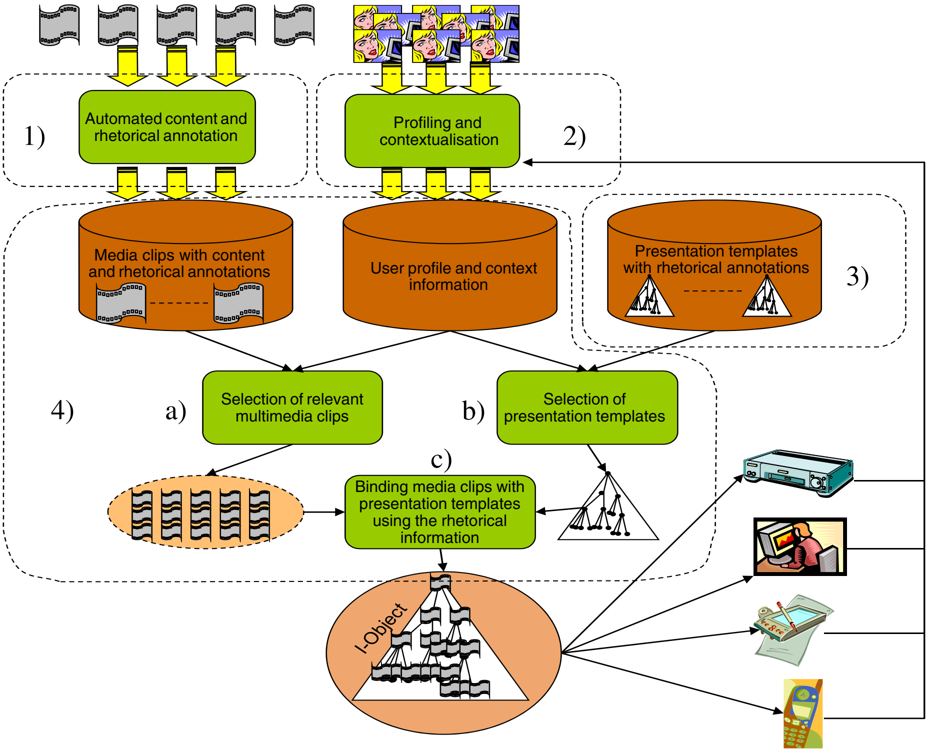
1)
Automated content and rhetorical annotation
Profiling and contextualisation
2)
Media clips with content and rhetorical annotations
User profile and context information
Presentation templates with rhetorical annotations
3)
4)
a)
Selection of relevant multimedia clips
b)
Selection of presentation templates
c)
Binding media clips with presentation templates using the rhetorical information
I-Object

Fig. 1 Automated generation of I-Objects

3. **Multi-modal input and output**. The best results in understanding user needs and presenting data in a way a user finds appealing and can benefit the most from have come from system in which multiple types of input are to be understood and presented in a multi-sensory way to the user. There is extensive research in cognitive science that is being applied to computer science in this field.
4. **Creation of presentation templates**. Experts of multimedia presentation authoring need to create a set of presentation templates enriched with affective information. A presentation template describes the presentation structure as well as the active/intelligent behavior of the I-Objects that are generated, but does not include the media clips to be presented. All components of a presentation template are annotated with specific affective descriptions that indicate what type of clips should be used for that component, from an affective perspective.
5. **Personalized generation of I-Objects**. A personalized I-Object can be an active/intelligent presentation with style and content pertinent to the user interest and prior knowledge. A personalized I-Object can be generated by a) selecting clips of documents having semantic content relevant to the user’s requirements, b) selecting the most appropriate presentation template on the base of the user profile, c) binding elements of the selected template to the selected clips, according to their rhetorical annotations. A personalized I-Object should be capable of further refining the final generated presentation by adapting itself to the user’s environment, access device and preferences.

The key research issues that have to be addressed are the following:

1. **Representation of multimedia content descriptions**. It is necessary to define a model for describing multimedia content that enhances and integrates emerging standards, such as MPEG7/21, to represent all the video content metadata. MPEG7 can be used as a knowledge representation language capable of describing semantics in real world applications and can facilitate a description of domain ontologies.
2. **Representation of user preferences and context**. Advanced models for personalization and conceptualization of information are required. Models should include those dedicated to the context of the original source media as well as those reflecting the situation of the user. User models need also to include structures to describe personal demographic profiles and preferences, local setting (e.g., time, place, and organization), task assignment and goal, and recursively, other domain specific ontological structures, to enable contextual reasoning. User context and profile should be dynamically updated and refined to reflect external input including the user retrieval and interaction history. The context models can be described in knowledge representation structures that adhere to the relevant international standards.
3. **Discovery of semantic concepts in data**. For systems that want to discover the semantics of new incoming media, discovering semantic concepts is just as important as the ability to determine the relationships between them. It is necessary that a particular brand of feature extraction provide distinct enough concepts to be useful. This makes the integration of multiple data sources of key interest. Research on this topic begins with automatically dividing a video into its related parts for analysis and ends with low level annotation.
4. **Automated content annotation**. Starting with the extraction of low-level, machine recognizable features from the temporal image and audio files, one can identify higher-level objects, events, sounds, words, and phrases including names, times, and places. These can be combined to derive actions, activities, and concepts of the subject matter and finally can be classified based on their affective information. All such derived knowledge may be coded into a complex database that enables search, correlation, and recombination along any dimensions. The goal is to bridge the "semantic gap" between the extraction of the underlying raw feature data and the need for a semantic description of the content and its style [85].
5. **Multi-modal input and output**. Cognitive science shows that there are benefits to both presenting information in a multi-sensory fashion and to receiving it in multi-modal ways. For example, a combination of speech and pointing device may assist a system in developing a more contextually accurate query for information than just a pointing device or speech recognition alone. A corollary to this is that multimedia data may be interpreted as a multi-modal interface as well and can benefit to a degree from the techniques discussed. The cutting edge of research right now is focusing on how to accurately model what it means to receive data and fuse them into a consistent internal representation and how to break apart the resultant data into a presentation for the user.
6. **Automated presentation authoring**. New techniques to generate professional quality interactive multimedia presentations starting from annotated (portions) of multimedia documents are needed. It is important to investigate how media and content descriptions can be embedded in I-Objects, jointly with the intelligence to combine them and to autonomously produce end-user interactive multimedia presentations. Intelligent objects can include content metadata, state, behavior and they should be capable of interoperation with the environment (user interactions, user profiles, and

contexts). They should be able to modify the environment (user profiles and contexts) and should allow to be modified by it (object behavior and presentation). It is necessary to create and use prototypes of presentation layouts built according to semantic, affective, and presentation optimization strategies that allow the I-Object to generate the most appropriate presentation. The same I-Object might generate different presentations corresponding to differences in both the user profiles (e.g., age and education attributes) and the affective context in which it is delivered, and then be appropriately rendered for the user's output device.

7. **Distributed retrieval and filtering of content descriptions and user preferences**. Content, context, and user profile descriptions can be represented using standardized XML conventions in order to enable import, export, interpretation and real-time customization in a widely distributed, non-uniform, information and user network. As evidenced by recent activity (e.g. RSS and Podcast), we anticipate a plethora of virtually continuous user-generated multimedia content sources in the future.

3 Current research

Current research on presentation is actually a long chain of applications working together to provide data, context, preference and layout. The following subsections take a step by step approach to through the chain of applications to explore what the current research at each step looks like. The first step is automatic annotation. Though automatic discovery and annotation may seem like a distant field of research, they are actually very related because the models chosen to represent annotation and context affect how they can be indexed and compared. From annotation, the next step in the process is ontological modeling and then indexing and adapting queries into the data based on user preference.

3.1 Representation and discovery of semantic concepts and their relationships

A video typically contains a series of concepts in isolated chunks of video. A reasonable analogy would be the scenes of a movie. Each scene has independent meaning and semantics and it does not seem reasonable to analyze disparate scenes to find their joint context. Rather, each scene is independently analyzed and then a set of scenes compiled to represent a series of concepts. To this end, the usual first step in semantic analysis of a video is segmenting the video (which contains a lot of theory itself) into manageable parts usually called shots. In turn, shots are a series of frames that provide an appropriate boundary for analysis. In most approaches, only a few representative frames of the shot are chosen for analysis. Detecting shot boundary has matured quite a bit and there are several well known techniques currently in use.

The simplest technique is just to compare, pixel by pixel, two consecutive images. If by some distance measure, the difference between the two images is above a certain threshold, a shot boundary is assumed. This idea is built upon by considering the HSV histograms of two consecutive images and computing their distance. Again if the difference is above a particular threshold, then a boundary is assumed. Other techniques include dividing an image into chunks and computer gray scales or the mean and standard deviation. Once boundaries are discovered, concept detection can begin. At this stage the research for the automated generation of content annotation begins.

3.2 Automated generation of media content descriptions

MPEG-7 and MPEG-21 are the default standards for describing multimedia content. Recently it has been shown that these standards can be used as a knowledge representation language and have been used to describe the semantics of the content of sport videos utilizing complex sports ontologies (such as soccer ontologies) [94]. The resulting metadata descriptions are MPEG-7 compliant and can be used by any MPEG-7 application.

The fundamental obstacle in automatic annotation is the semantic gap between the digital data and their semantic interpretation [85]. Progress is currently being made in known object retrieval [25, 65], while promising results are reported in object category discrimination [23], all based on the invariance paradigm of computer vision. Most of object retrieval systems use color, shape, texture, motion-based retrieval method, which are still not good enough to understand video contents. There is a big distance between semantic retrieval technology and commerce applications. Significant solutions to access the content and knowledge contained in audio/video documents are offered by StreamSage [88] and Infomedia [38]. While the field of content-based retrieval is very active by itself, much is to be achieved by combination of multiple modalities: data from multiple sources and media (video, images, text) can be connected in meaningful ways to give us deeper insights into the nature of objects and processes.

So far multimodal data knowledge mining has mostly been carried out separately on each information channel. Today, however, knowledge sources that marry multiple descriptions are urgently needed to support the analysis and retrieval of mixed-media. The picture–text combination for example is widely considered to be the richest option for information access. In worldwide, task-based retrieval evaluations such as TRECVID, an integrated approach combining text and visual information is the essential ingredient in the most successful systems [86], in combination with the use of machine learning techniques.

Wold et al. [103] present a system which analyzes sounds based on their pitch, loudness, brightness, and bandwidth over time and tracked the mean, variance, and autocorrelation functions of these properties. Other approaches (e.g. [24]) are based on methods developed in the digital speech processing community using Mel Frequency Cepstral Coefficients (MFCCs) and motivated by perceptual and computational considerations. However, the MFCCs ignore some of the dynamic aspects of music. A different approach is taken by the SOM-enhanced Jukebox system [75], where characteristics of frequency spectra are extracted and transformed according to psychoacoustic models focusing on the rhythmic characteristics.

Naphade et al. [68–70] develop a multi-modal representation of a concept or feature called a multi-ject or multimedia object. A multi-ject uses a feature extraction algorithm and bands them together with semantic labels. It also attempts to attach significance to any non-verbal audio in the video. A trustworthy data set is developed with expert help (in this case it involves training a hidden Markov model) and future searches are compared against these features. In this way you have a series of high level contextual labels paired with image and audio features in a manner that conveys the probability that particular comparative scenes should be given the label of a matching well defined canonical object from the data set. The relationships between multi-jects is defined in an object called a multi-net which is essential a graph weighted by the probability of a relationship. The presence of particular multi-ject in a scene can then influence the probability that another multi-ject is in a scene base on its multi-net relationship.

Multi-jects (and most other attempts) suffer from having a static data set from which to compare new data for indexing. These data sets require supervised build up and maintenance. Additionally, the concept of a multi-ject in general is certainly not trivial. This can result in the indexing and retrieval process being slow Wang [97]. Wang tries to solve both of these problems by focusing instead of concepts that occur simultaneously. Like Naphade, concepts are represented by a few key frames from a shot. There is however, there is a difference in representation. Wang represents a concept as a HSV color histogram and an edge histogram. Given a video, concepts extracted from the video can be compared to the existing set of concepts to see how closely they match. A candidate set of features is built and a Bayesian network tests the probability that a feature is a new feature worth saving. In this way, the videos are not only annotated, by new annotations can be created dynamically.

Song et al. [87] have a concept representation similar to Wang but rely on clustering techniques to find image relationships. In this method, images are clustered based on visual similarity and distance in time. Using a variant of linear regression, concepts that least likely fit their cluster become candidates for supervisor review and annotation. The upside to this particular technique is that not all images need annotation. Well fit concepts can simply share their annotations with each other. Lavrenko et al. [50] have used a different method with which to search images for concepts. In this technique, an image is segmented into rectangular parts, either with some algorithm or just overlaying a regularly spaced grid. Each rectangle is then analyzed for features and the distribution of features and words is recorded in a model called a continuous relevance model. A nice feature that results from griding each image is that regions of the data can be labeled which provides a finer granularity over the annotation process. Research on this technique focuses heavily on improving the mathematics of determining word and feature distribution. Other examples include gm-LDA and correspondence-LDA [9].

3.3 Affective computing

While classification into semantic categories is a comparatively mature field, having produced a range of approaches and results, annotation according to affective or emotional categories of video is a relatively young domain, gaining more and more importance [5, 32, 47, 53, 55, 58, 66, 96, 105]. If the affective content of a video is detected, it will be very easy to build an intelligent video recommendation system, which can recommend videos to users based on users' current emotion and interest. For example, when the user is sad, the system will automatically recommend happy movies to the user; when the user is tired, the system will suggest relaxing film.

All the current affective analysis systems try to solve the following problems [83]: 1) identification of valid affective features; 2) bridging the gap between affective features and affective states; 3) establishing an affective model to take user's personality into consideration; 4) representing affective state.

In general, there are three kinds of popular affective analysis methods. *Categorical affective content analysis methods* usually define a few basic affective groups and discrete emotions, for example, "happy", "sadness" and "fear". Then classify video/audio to these predefined groups. Moncrieff et al. [66] analyze changes in sound energy of the non-literal components of the audio tracks of films and detect four sound energy events commonly used in horror film: "surprise or alarm", "apprehension or emphasis of a significant event", "surprise followed by a sustained alarm", and "building apprehension up to a climax". They find that these four sound energy events convey well established meanings through their dynamics to portray and deliver certain affect, sentiment related to the horror film genre [66].

Lu et al. [58] propose a hierarchical framework to detect four mood categories: contentment, depression, exuberance and anxious/frantic from acoustic music data. They extract three feature sets from each music: intensity feature set (energy in subband), timbre feature set (spectral shape feature and spectral contrast feature), and rhythm feature set (rhythm strength, rhythm regularity and tempo). They build up a hierarchical framework, which uses intensity feature to classify a music clip into *Contentment* and *Depression* group and *Exuberance* and *Anxious/Frantic* group, then uses timbre and rhythm features to determine which exact mood the music clip is. In the framework, a Gaussian mixture model (GMM) with 16 mixtures is utilized to model each feature set regarding each mood cluster (group). Kang et al. [47] detect emotional events such as fear, sadness and joy from videos by computing intra-scene context (shots' coherences, shot's interactions, dominant features in color and motion information) and inter-scene context (scene's relationship with other scenes). Xu et al. [105] identify video/audio segments which make audience laugh in comedy and scary segments in horror films as affective contents. They use Hidden Markov Models (HMM) based audio classification method to detect audio emotional events (AEE) such as laughing, horror sounds etc. Then use the AEE as a clue to locate corresponding video segment.

The second type of affective analysis method is called *Dimensional affective content analysis* method, which commonly employs the Dimensional Affective Model to compute affective state. The psychological Arousal-Valence (A-V) Affective Model [39, 49, 78, 81] is one popular Dimensional Affective Model. Arousal stands for the intensity of affective experience and Valence characterizes the level of "pleasure". Hanjalic and Xu [30, 31] did research on affective state representation and modeling by using A-V Affective Model. According to A-V affective model, the affective video content can be represented as a set of points in the two-dimensional *(2-D) emotion space* that is characterized by the dimensions of *arousal* (intensity of affect) and *valence* (type of affect). By using the models that link the arousal and valence dimensions to low-level features extracted from video, the affective video content can be mapped onto the 2-D emotion space. Then *affect curve* (arousal and valence time curves) can be easily detected as reliable representations of expected transitions from one feeling to another along a video. Pleasure-Arousal- Dominance (P-A-D) model [64] is another popular affective model. Pleasure stands for the degree of pleasantness of the emotional experience, Arousal stands for the level of activation of the emotion, and Dominance describes the level of attention or rejection of the emotion. Based on P-A-D model, Arifin et al. [5] propose to use Dynamic Bayesian Networks (DBNs) to build up a P-A-D value estimator, which estimates the P-A-D values of the video shots of the input video. Then video can be segmented based on the estimated P-A-D content. Different from the Arousal and Valence modeling proposed by Hanjalic and Xu, this work takes the influences of former emotional events and larger emotional events into consideration.

The third type of affective analysis method is *Personalized affective content analysis method.* The representative work is reported in [96], which introduces more personalization factors into affective analysis for Music Video (MV) retrieval. First, they build a user interface and record the users' feedback in the user profile database. Each profile records MV's ID, user's descriptions about MV's Arousal and Valence (two scores describing their opinions about Arousal and Valence level). When users play MV, they also can use feedback to change their opinions on MV at any time. Based on users' profile, two Support Vector Regression (SVR) models (Arousal model and Valence model) are trained to fit the user's affective descriptions. Finally, the affective features extracted from MV are fed into the trained models to get the personalized affective states. They also provide a novel Affective Visualization interface for efficient and user-friendly MV retrieval. Through this interface, the user can easily log into the system, search MV based on their affective states (for example, anger, happy, sad/blue, or peaceful) and also provide his/her feedback on each MV.

3.4 Personalization and contextualization

Given a source of richly annotated media, disseminating it efficiently becomes a primary task if useful applications with easy to use interfaces are to be developed. Users find it increasingly difficult to navigate information as its volume and variety increase and so adaptable systems become highly desirable [44]. It is this need that motivates the current research on personalization and customization. Personalization and conceptualization actually can be subdivided in to several topics. The first is profile modeling or deciding in what way can we organize and access a user profile so that it is useful and efficient. What kinds of structure should the data have allows a content provider to meaningfully record user information? As well discuss, MPEG-7/21 is of primary interest here though it has limitations. Secondly, how do you actually determine the users profile and optimize the user experience? There are several preference based algorithms currently in use like clustering, usage history and semantic interest. Finally, in a less related topic, how do you effectively acquire and share these profiles. A website may store information about your visit in the form of a cookie on your personal machine and more detailed information about your preferences remotely in the website applications own database. What are intelligent and efficient ways to share this data between interested parties? The nature of adaptability is also in question [44]. Division of labor can be an important question for designing a system that is comfortable for a user. To little interaction and a user may not understand the implications of customization or personalization, too much and a user may not have enough control.

Jameson defines several types of adaption and interaction [45]. Database inference model is one example. Typically, this entails gathering usage about usage but can sometimes, less frequently be accumulated data from several users. In this instance, you try to make general inference about what users typically want. Collaborative filtering is an example of this type of inference. Theory inference models create some notion of a user model ahead of time which can then be customized and adapted. A Bayesian model can often be used to create stereotypes or preconfigured templates of users' preference. Decision making models make inferences about the user but also try to figure out what adaptations are best. Input variables may be weighted and selected from for optimal use. The following section discusses some ways in which users are modeled in various systems.

It's not so easy to transparently accumulate accurate data on users [22, 74] to determine proper contexts. Experimentally, the best results for personalization tended over time, to be based on some accumulation of explicit feedback from the user and implicit observation of usage and history. Eynard [22] makes the distinction between customization and personalization. Customization involves direct user manipulation of the users own profile. Personalization relies on perceived interest garnered through metrics like time spent on a topic and the topics previously searched. Research has shown that the most effective profiles are created when users can give active feedback on the personalization process. That is to say that if a user can help trim unhelpful optimizations, then the remaining options are more likely to be useful. Pandora radio is an example of this kind of customized personalization. In their system, they try to find music based on your perceived preferences, but they give you the option to tell them they made a mistake. Pandora tries to fine tune their selections based on the user input and over time the user is happier with the presentation.

The idea led to several pieces of software designed to help users understand their own usage through visualization and back linking to referencing data. Research in profile representation and context discovery have driven the development of a wide range of algorithms that are useful for finding better content for a user.

McCarthy [63] introduced contexts as formal objects. Context is very useful for localizing the reasoning to a subset of facts known to an agent about a specific problem. Contexts are seen as local, domain-specific, goal-driven theories of the world, and are building blocks of what an agent knows. Contexts have been used extensively in AI to formalize agents which have a representation of the changing beliefs, intentions and goals of other agents involved in dialogue or negotiation [26, 27, 51, 73]. Context is also used in linguistics where it refers to the words in an utterance that are near in the focus of attention and further specify it [16]. In information systems, context has been described in terms of context types [8] that relate to application types. Context types proposed in [48] include organizational, domain-based, personal, and physical context. Each is subdivided into more specific types like work flow and structure, domain ontology, knowledge profiles, usage history, interest profiles, etc. In ubiquitous environments context types including location and time are often used for proximate selection and for context-driven actions [80].

Rigo and Oliviera [76] attempt to monitor usage histories of users and make adjustments to content presented to a user based on the actual semantics of the content browsed. The content needs to pre-annotated so this isn’t a system that actually derives the meaning of some media. The idea is simply to monitor the content browsed by a user and create a log. The log is analyzed and matched to usage patterns called classes. A class is a stereotype of interests groups shared by a group of people whose usage indicates they belong in the class. Relationships between content can be paired and provided to the user a-priory over similar content that might be otherwise similar but different in context.

Jameson [1] has a relatively novel approach at determining user preference. His group recognizes the difference in need a user has when searching for the same data for different reasons. A user may need for example papers published specifically by an author, papers related to a particular paper, or papers constrained to a certain time frame. The user may make nearly identical queries in each situation. The general approach to this problem is to create temporally relevant preference models. The long term model is the general trend of user over an extended period. The medium term model addresses the current need of the user and ontological relevance is more heavily weighted. The short term model is very quickly adapting and more inaccurate. It focuses more on the relationship between objects rather than their ontology. Any number of referencing paradigms can be built off the cumulative information in these three profiles.

Jameson uses three main techniques to help build data relationships based on user usage history. Usage mining tries to find regularities in rarely used and often used data. This can be useful in ranking competing results for relevance. Structural mining tries to find regularities in the linkage between data sources and content mining strives to find regularities in the context. Context mining is useful in determining if some context is underrepresented by the query results thus indicating the need to find more resource around that context.

Research relating to actual storage and retrieval typically develops query frameworks. Kadlik and Jalenic [46] propose a series of ontological agents that support each other though it's not clear in what way they do this. User profiles are stored as a distributed hash on an ontology server. The server organizes various ontological agents to find content. It also handles communication and organization. They chose the distributed hash as a model because of the dynamic, distributed environment they envision, but they note the inherent weakness resulting from the lack of clustering relating concepts meaningfully and the trouble of developing unique identifiers for concepts.

In summary, research in user preference is divided between preference modeling and preference development. What factors go into developing good semantics and what models are most expressive of those semantics are intricately linked—not unlike research in video annotation. For now, MPEG-7/21 standard is undergoing heavy research and it seems likely this model is will be the default standard for multimedia preference annotation.

3.5 Multimodal analysis

Multimodal presentation is a huge topic. To understand the problem, one needs to understand why and when it may be useful, how to determine which modes of input are most useful, how to interpret the data into a internal format independent of modality, data retrieval, how to exploit the presentation of this data to the advantage of the user and how a system like this might even be modeled. While each sub-problem has some relation to presentation generation, the aspects of the problem that are most relevant are probably fission, the process of divesting concepts into disparate pieces for presentation, and fusion, the modeling of various data types into a contextual language that be useful for queries.

Cognitive science has shown that benefits of multi-modal interfaces center heavily on concurrent processing in the brain of the user. There is an apparent parallelization going on when a user visualizes and listens to data relating to the same topic [11]. What this means from a system development standpoint is that it is useful to know how to present data to most affect this advantage. While studies show that learning capabilities may rise only 10% [11], error correction and accuracy go up. Multiple sources of data input to a system better relate the problem than a single data source. The trick is in which way do you select modalities that are useful to each other. Speech and lip movement might be paired redundantly while pointer and speech would be complimentary. To this end, there are several categories of paired modalities of which to be aware. The following list was compiled by Oviatt et al. [11]

- Data fusion is a situation in which data sources are very similar. Oviatt gives the example of two web cams viewing the same scene from different angles. You get high redundancy and no information loss since you have the raw data. The trade off is that you must now deal with a higher noise level.
- Feature fusion are complimentary models (like lip movement and speech) that are time dependent.
- Decision fusion is useful for managing modalities with minimal relationship to each other but which have a unified semantic. The classic example is point and talk combinations.

Decision level fusion is the most common and can be further subdivided. Again, the demarcation here is provided by Oviatt et al. [11] and loosely conveys the types of structures most used in decision level fusion. This is precisely the reason decision level fusion is of even higher interest than feature and data level fusion in presentation generation. It tends to represent data in a semantic way through key-value pairs and statistical analysis. Recall that in Section 3.4 this kind of structure and methodology is also prevalent and we therefore have a kind of symmetry between user input and input from data queries in general. Developing contextual awareness from user input might then give a very precise semantic pairing with a database of context annotated content.

Another aspect of multi-modal modeling that is of key interest in presentation generation is fission. Fission is the process of taking information, or context to be presented to a user and deciding on a channel and form for output. If you can represent the same data via photo, audio commentary or video you have to select one or more formats to present. Once the format(s) are chosen, the actual presentation has to be constructed. I leave the presentation construction to Section 3.6. While there is a lot of research in temporal presentation, the most interesting work may be in the field of machine learning. In this case, monitoring features of the user may give clues as to what is of greatest interest so emphasis can be provided on those channels.

This field of research is deeply tied to user preference on both sides of the interaction process. Users want to use multi-modal systems [72]. They way they want to use them is highly dependent on the type of data and request being made and therefore so is the resulting output. Oviatt used a series of simulations and task analysis with users to experimentally determine how a user prefers to use a multi-modal interface. The most significant conclusion is that preference is highly tied to requests involving spatial regions and the use of several modes of input are usually intermittent and discontinuous. When several modes are used, their data tend to be complimentary rather than redundant. In other words, users simultaneous perform related tasks with different inputs rather than the same task. This seems a natural conclusion as Oviatt also notes that this pattern occurs in general communication with language and gesturing (Figure 2).

By way of a real word example, Zhou et al. [112] provides an overview of their multi-modal presentation system dubbed RIA. The following diagram is taken from their paper as it is useful in understanding the method in which they modeled the data.

In this model, step 1, the input interpreter is the beginning of the fusion process described above. The conversation facilitator essentially develops a contextual framework that allows the presentation broker to do create a sort of template. Part b of the diagram is an example conversation. Based on the template, step 4 fills the template with media specific data. It is here the abstraction of the fission process lives.

There is heavy research being done on developing fusion engines and presenting data. The method surveyed in this paper for presentation generation relies upon an abstraction called an intelligent object and is the subject of the next section.

3.6 Automated generation of presentations from intelligent objects

There are at least two problem sets to solve for presentation from intelligent objects. The first challenge is a reasonable container to hold information about the object and its annotations. This is a problem with similarities to personalization, but which has its own specific container solutions such as MPEG4. The second half of the problem set is in representing the presentation and finding a descriptive way to process relationships. This seems generally to be represented as a constraint satisfaction problem where much of the work is in developing a suitable language to express the constraint relationships.

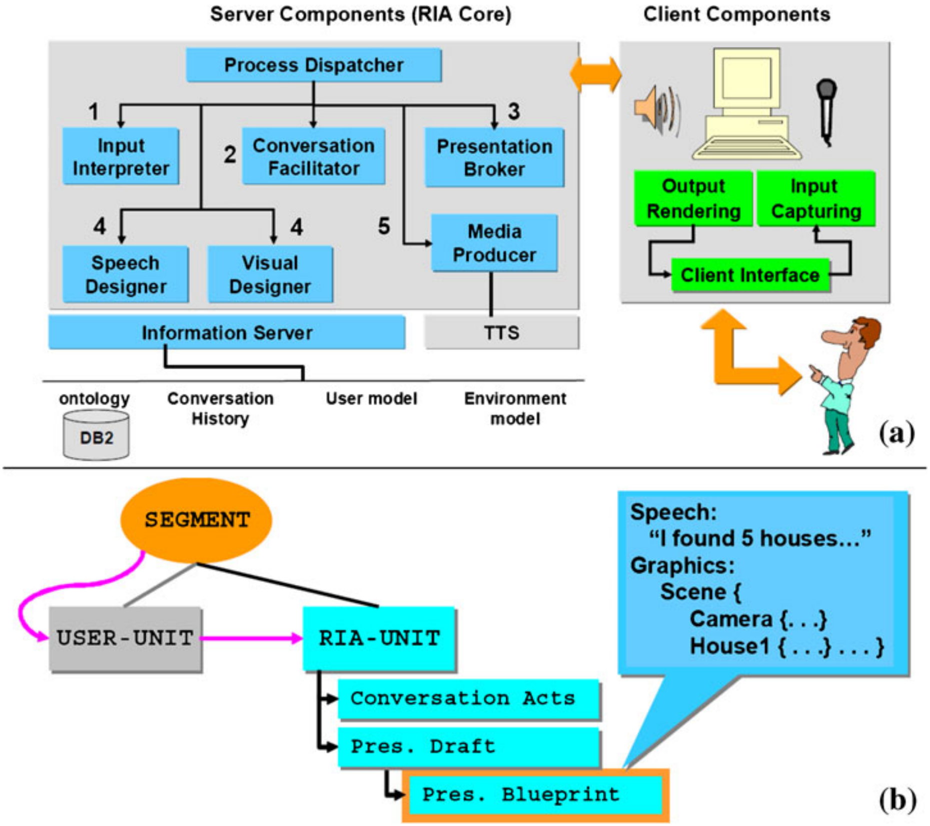

Fig. 2 RIA Multi-modal presentation architecture [112]

There are several existing technologies available that can be used to define storage containers for objects. Formats such as MPEG4 [67] provide an encapsulated object model, in which one or more instances of a media object can be stored along with relevant annotation information. MPEG4 supports a multi-layered model in which one or more low-level media encoding are packaged into a composite object as a collection of MPEG streams. While complex stream collections are possible, containing multiple encodings for multiple target environments, such an approach typically leads to an explosion of encodings within a single package, as the final presentation target environment is defined. Practical use of multiple MPEG4 streams is also limited by minimal support for complex stream processing within most MPEG4 codes. XML languages such as SMIL [12] provide an indirect reference model for defining the multiple media encodings associated with a single abstract media item. SMIL contains an extensible content selection mechanism that could be used to define a semantic and syntactic hierarchy to serve as the basis for final form object selection. Neither MPEG4 nor SMIL provide complete solutions to the object encoding problem, since multiple layers of semantic markup will need to be supported, but both technologies provide a valid starting point for investigating presentation component persistent storage.

The existing approaches for encoding presentation style information are less complete than media object storage. The GRiNS editor [13, 36] supported the use of a language to specify generic presentation templates for defining abstract layouts from which to generate run-time presentation instances. However, this language supported only physical attributes (e.g., screen size, rendering capabilities, available codes) when associating media objects with layout elements. Such policies will need to be expanded to define more generic mapping of content to affective, semantic, and physical device properties.

Given a set of presentation style descriptions and candidate media objects, final objects need to be generated subject to knowledge of the target environment and the assets available for presentation. Many text-based systems can make use of style sheet transformations (using XSLT and CSS), but these languages are not rich enough to support a wide range of media and affective content characterizations. Other approaches to solving portions of this generation problem include those of Weitzman et al. [101], who used a method based on relational grammars, and Zhou [110], who used a visual planning model. More recently, several implementations of presentation generation have been defined that are based on the use of rhetorical structure theory (RST) [60], but all of these approaches are tailored to generate text descriptions based on a structured markup at the text encoding level.

To look at a more standardized query frameworks, Anthony Jameson [41] provides detailed analysis. Jameson discusses the use of an ontological language called SWQL (pronounced SWEQUEL) which standards for Semantic Web Query Language. SWQL is based in part on XQUERY but uses the graph structure from OWL (Web Ontology Language) to organize its internal type system. In this model, the data is organized as a series of items which can be considered a node or property. Nodes can be either literals or objects. Nodes are connected by properties. Various functions are defined on each item including NodeTest and PropertyTest. NodeTest can be used to group and compare types—for instance "publications" while PropertyTest can be used to navigate the graph structure.. Filtering structures like predicates are available for paring down collections of nodes by some condition.

Another technique to model the user query is MPEG-7/21. However, the personalization of MPEG-7/21 is limited and cannot account for the domain knowledge described in specific ontology. and thus it cannot exploit combined MPEG-7/domain-specific metadata descriptions. MPEG-7 does provide a rich set of API to document semantics however. This means it is generally the case that the problem is because the MPEG-7/21 context model is primarily limited by the fact that it is keyword based and it is clear that keywords are not enough to adequately express complex semantic relationships. There is, therefore, some research based on extending the MPEG-7/21 protocol to allow it to exploit the rich context provided in MPEG-7.

Tsinaraki and Christodoulakis present an attempt at an extension [92, 93]. The traditional MPEG-7/21 environment contains four categories of properties. User characteristics, terminal capabilities, environment, and network sections are represented. Of these the most important for personalization is obviously user characteristics which contain sections on general information, preferences, history, presentation and accessibility. What is missing from these particular categories is a way in which to express the semantic preferences about content viewed. Their canonical example is trying to specify an interest in goals for the home team in a soccer game. It is not sufficient to just specify "goal" or "goal" and "home team" because there is inherent ambiguity here. There is no inherent meaning to the keywords that allow one to record deep, semantic preference information. This information does exist in the MPEG-7 MDS and so the general concept is a new definition that is a general container for semantic knowledge that is sub-classed and weighted for preference. By way of example [93] they provide the following:

```
<UserPreferences id="UP1">
  <FilteringAndSearchPreferences>
    <CreationPreferences xsi:type="SCreationPreferencesType">
      <ContentPreferences>
        <SemanticBag preferenceValue=90
          xsi:type="UWeighedSemanticBagType">
          <SemanticBase xsi:type="EventType" id="ChGoal">
            <Relation type="agent" target="#Charisteas"/>
            <Relation type="exemplifies" target="#Goal"/>
            <Relation type="patient" target="#Ricardo"/>
          </SemanticBase>
       </SemanticBag>
      </ContentPreferences>
   </CreationPreferences>
 </FilteringAndSearchPreferences
</UserPreferences>
```

In this example the relationships become apparent: Charisteas goals versus Ricardo. In the description, you can see the value placed on the preference, the kind of event, the entities involved and a notion of the relationship between them (exemplifies). Compared to the set of keywords, Charisteas, goal, and Ricardo, it becomes obvious that the deeper meaning inherent in the user preference is only adequately expressed with their model and not the keywords.

Agius and Angelides provide a somewhat more complicated model they dub the hanging basket model [2]. In their model they define three planes: the event plane, the object plane and the property plane. The event plane is a set of events in time in a particular media stream. It's represented as a simple digraph. The object plane is the space in which each event is expressed in space and hierarchy between objects. Finally the property plane consists of a set of properties that define the objects of events. The description in MPEG-7 event plane would record the series of event time points and the relationships to the events. The object and property plane would be annotated similarly and they try to show that this is a sufficiently descriptive model for semantic retrieval.

Having a semantic model, the idea then is to build an internal data warehouse that reflects the ontology of the model. Jameson refers this as data integration and considers that should be able to index by type. The OWL structure provides a means of typing and type testing and this is the means by which you may want to look up relevant data. The important difference in indexes by type rather than by indexing in more traditional databases is that there may exist equivalence statements between types and properties on top of the hierarchal structure that is already present.

Regardless of the ontological model, there are quiet problems that need to be addressed. One example is that of duplicate detection [41]. Multiple references may source the same object but the object may represent slightly differently. An authors name abbreviated in different ways may confuse a system into thinking these are two separate pieces of information when in fact they are identical. Jameson tackles this particular problem of bibliographies but the general problem is still a difficult question and research on this topic continues.

There are also attempts to create a formatting model out of currenttext formatting systems like CSS and XSLT [29]. Guerts et al. propose an extension to CSS and SMIL that accounts for media properties like visual and temporal layout, schemes and transformation. The paper documents several useful case studies to make the argument that their extensions are viable.

There are a few attempts to develop a presentation structure independent of media type, which is focused more on the media relationships. One such example is an idea based on Weitzman's work. Guerts et al. [28] tries to develop a language to express the document as a constraint satisfaction problem. They start with the position that simple constraints with numerical domains and outputs are insufficient and try to abstract some special relationships like transitivity and association for objects. This is followed up with boolean level relationships that begin to full express a system of order in objects. It becomes clearer why an object has some proximity to another object now. Once rule types are defined, they illustrate the process of automatically generating the specific presentations rules based on their relationships. The presentation structure is then graphed as a grid rather than a tree so that relationships exist vertically and horizontally. This approach tries to provide a structure that shows objects in proximity based on relations defined in more logical terms than just keyword analysis. This allows an individual device to render the content appropriately for its technology since the underlying meaning of the content is no longer dependent upon its actual display.

4 Challenges

Despite the considerable progress of academic research in multimedia information retrieval [52], there has been relatively little impact of multimedia information retrieval (MIR) research on commercial applications with some niche exceptions such as video segmentation and image retrieval. One example of an attempt to merge academic and commercial interests is Riya (www.riya.com). Riya's goal is to have a commercial product that uses the academic research in face detection and recognition and allows the users to search through their own photo collection or through the Internet for particular people. Another example is the MagicVideo Browser (www.magicbot.com) which transfers MIR research in video summarization to household desktop computers and has a plug-in architecture intended for easily adding new promising summarization methods as they appear in the research community. Like (www.like.com) provides an interactive search interface for powerful visual shopping. Users can draw a box around a specific detail of a product they like. Then the visual shopping will return items with matching details. Through Like.com, users can easily search products by the color, shape and pattern they like. An interesting long-term initiative is the launching of Yahoo! Research Berkeley (research.yahoo.com/Berkeley), a research partnership between Yahoo! Inc. and UC Berkeley whose declared scope is to explore and invent social media and mobile media technology and applications that will enable people to create, describe, find, share, and remix media on the Web. Nevenvision (www.nevenvision.com) is developing technology for mobile phones that utilizes visual recognition algorithms for bringing in ambient finding technology. However, these efforts are just in their infancy, and it is important to avoid a future where the MIR community is isolated from real-world interests. We believe that the MIR community has a golden opportunity in the growth of the multimedia search field that is commonly considered the next major frontier of search [7].

An issue in the collaboration between academic researchers and industry is the opaqueness of private industry. Frequently it is difficult to assess if commercial projects are using methods from the field of content-based MIR. In the current atmosphere of intellectual property lawsuits, many companies are reluctant to publish the details of their systems in open academic circles for fear of being served with a lawsuit. Nondisclosure can be a protective shield, but it does impede open scientific progress. This is a small hurdle if the techniques developed by researchers have significant direct application to practical systems.

To assess research effectively in multimedia retrieval, task-related standardized databases on which different groups can apply their algorithms are needed. In text retrieval, it has been relatively straightforward to obtain large collections of old newspaper texts because the copyright owners do not see the raw text as having much value. However image, video, and speech libraries do see great value in their collections and consequently are much more cautious in releasing their content. More and more academic and research labs try to collect huge amount of images from internet (i.e. google, flickr, facebook) for large scale image retrieval and object recognition. MIT collects 80 millions images from internet for object and scene recognition [90]. Stanford opens ImageNet database (~10 millions images). These images are collected by WordNet Nouns keywords and organized based on WordNet hierarchy [17]. While it is not a research challenge, obtaining large multimedia collections for widespread evaluation benchmarking is a practical and important step that needs to be addressed. One possible solution is to see that task-related image and video databases with appropriate relevance judgments are included and made available to groups for research purposes as was done with TRECVID. Useful video collections could include news video (in multiple languages), collections of personal videos, and possibly movie collections. Image collections would include image databases (maybe on specific topics) along with annotated text (the use of library image collections should also be explored). One critical point here is that sometimes the artificial collections like Corel might do more harm than good to the field by misleading people into believing that their techniques work, while they do not necessarily work with more general image collections. Therefore, cooperation between private industry and academia is strongly encouraged. The key point here is to focus on efforts which mutually benefit both industry and academia. As was noted earlier, it is of clear importance to keep in mind the needs of the users in retrieval system design, and it is logical that industry can contribute substantially to our understanding of the end-user and also aid in the realistic evaluation of research algorithms. Furthermore, by having closer communication with private industry, we can potentially find out what parts of their systems need additional improvements to increase user satisfaction. In the example of Riya, they clearly need to perform object detection (faces) on complex backgrounds and then object recognition (who the face is). In the context of consumer digital photograph collections, the MIR community might attempt to create a solid test set which could be used to assess the efficacy of different algorithms in both detection and recognition in real-world media.

There is active research that attempts to solve the annotation collection problem by easing the process by which such collections can be created. So far results are focused largely on reducing the supervised input of initial annotations by correlating concept co-occurrence and building relationship graphs between them and by choosing images which are most likely to need better annotation versus images that are likely to be sufficiently annotated based on similar imagery. To this end there are several methods for representing a concept and computing likeness. The general point seems to be heading towards teaching a machine to independently isolate a concept and solving its relationship with other known concepts. A concept abstracted this way has no semantic labeling, but does provide the ability to rank candidate images more precisely.

The potential landscape of personalized multimedia information retrieval is quite wide and diverse. Following are some potential areas for additional MIR research challenges.

4.1 Multimedia input analysis and output generation

Many research challenges remain in areas such as inter-media segmentation, partial input parsing and interpretation, and partial multimedia reference resolution [62]. New interactive devices (e.g., force, olfactory, and facial expression detectors) need to be developed and tested to provide new possibilities, such as human emotional state detection and tracking. Techniques for media integration and aggregation should be further refined to ensure synergistic coupling among multiple media, managing input that is impartial, asynchronous, or varies in level of abstraction. Algorithms developed for multimedia input analysis have proven beneficial for multimedia information access [18]. More and more physical sensors information is used as personal profile for multimedia systems. Aizawa et al. [4, 34, 35] use physiological sensors to detect user's brainwave activity while capturing the video content. These brainwave activities are used as a measure of the user's interest to generate video summaries, which consist of a series of key frames representing the most pertinent video segments, based on a given inclusion threshold value related to the significance of brainwave activity associated with the video content. In the affective computing, facial expressions, head movements, and body gestures detectors are used to detect and understand human affective behaviors for personalized selection [108].

Important questions remain regarding methods for effective content selection, media allocation (e.g., choosing among language, non-speech audio, or gesture to direct attention), and modality selection (e.g., realizing language as visual text or aural speech). In addition, further investigation remains to be done in media realization (i.e., choosing how to say items in a particular media), media coordination (cross modal references, synchronicity), and media layout (size and position of information) [19].

4.2 Human centered methods

We should focus as much as possible on the user who may want to explore instead of search for media [40–42]. It has been noted that decision makers need to explore an area to acquire valuable insight, thus experiential systems which stress the exploration aspect are strongly encouraged. Studies on the needs of the user are also highly encouraged to give us a full understanding of their patterns and desires.

Whether we talk about the pervasive, ubiquitous, mobile, grid, or even the social computing revolution, we can be sure that computing is impacting the way we interact with each other, the way we design and build our homes and cities, the way we learn, the way we communicate, the way we play, the way we work. Simply put, computing technologies are increasingly affecting and trans-forming almost every aspect of our daily lives. Unfortunately, the changes are not always positive, and much of the technology we use is clunky, unfriendly, unnatural, culturally biased, and difficult to use. As a result, several aspects of daily life are becoming increasingly complex and demanding. We have access to huge amounts of information, much of which is irrelevant to our own local socio-cultural context and needs or is inaccessible because it is not available in our native language, we cannot fully utilize the existing tools to find it, or such tools are inadequate or nonexistent. Thanks to computing technologies, our options for communicating with others have increased, but that does not necessarily mean that our communications have become more efficient. Furthermore, our interactions with computers remain far from ideal, and too often only literate, educated individuals who invest significant amounts of time in using computers can take direct advantage of what computing technologies have to offer.

Clearly, a Human-Centered Computing research agenda should include a broad understanding and a multidisciplinary approach, as Brewer, et al., [10] propose in the specific context of developing regions.

4.3 Multimedia collaboration

Discovering more effective means of human-human computer-mediated interaction is increasingly important as our world becomes more wired. In a multimodal collaboration environment many questions remain: How do people find one another? How does an individual discover meetings/collaborations? What are the most effective multimedia interfaces in these environments for different purposes, individuals, and groups? Multimodal processing has many potential roles ranging from transcribing and summarizing meetings to correlating voices, names, and faces, to tracking individual (or group) attention and intention across media. Careful and clever instrumentation and evaluation of collaboration environments [62] will be the key to learning more about just how people collaborate.

Very important here is the query model which should benefit from the collaboration environment. One solution would be to use an event-based query approach [56] that can provide the users a more feasible way to access the related media content with the domain knowledge provided by the environment model. This approach could be extremely important when dealing with live multimedia where the multimedia information is captured in a real-life setting by different sensors and streamed to a central processor.

4.4 Interactive search and agent interfaces

Emergent semantics and its special case of relevance feedback methods are quite popular because they potentially allow the system to learn the goals of the user in an interactive way. Another perspective is that relevance feedback is serving as a special type of smart agent interface. Agents are present in learning environments, games, and customer service applications. They can mitigate complex tasks, bring expertise to the user, and provide more natural interaction. For example, they might be able to adapt sessions to a user, deal with dialog interruptions or follow-up questions, and help manage focus of attention. Agents raise important technical and social questions but equally provide opportunities for research in representing, reasoning about, and realizing agent belief and attitudes (including emotions). Creating natural behaviors and supporting speaking and gesturing agent displays [62, 100] are important user interface requirements. Research issues include what the agents can and should do, how and when they should do it (e.g., implicit versus explicit tasking, activity, and reporting), and by what means should they carry out communications (e.g., text, audio, video). Other important issues include how do we instruct agents to change their future behavior and who is responsible when things go wrong.

4.5 Neuroscience and new learning models

Observations of child learning and neuroscience suggest that exploiting information from multiple modalities (i.e., audio, imagery, haptic) reduces processing complexity. For example, researchers have begun to explore early word acquisition from natural acoustic descriptions and visual images (e.g., shape, color) of everyday objects in which mutual information appears to dramatically reduce computational complexity [77]. This work, which exploits results from speech processing, computer vision, and machine learning, is being validated by observing mothers in play with their pre-linguistic infants performing the same task.

Neuroscientists and cognitive psychologists are only beginning to discover and, in some cases, validate abstract functional architectures of the human mind. However, even the relatively abstract models available from today's measurement techniques (e.g., low fidelity measures of gross neuro-anatomy via indirect measurement of neural activity such as cortical blood flow) promise to provide us with new insight and inspire innovative processing architectures and machine learning strategies.

Caution should be used when such neuroscience-inspired models are considered. These models are good for inspiration and high-level ideas. However, they should not be carried too far because the computational machinery is very different. The neuroscience/cognition community tries to form the model of a human machine, and we are trying to develop tools that will be useful for humans. There is some overlap, but the goals are rather different.

Machine learning of algorithms using multimedia promises portability across users, domains, and environments. There remain many research opportunities in machine learning applied to multimedia such as on-line learning from one medium to benefit processing in another (e.g., learning new words that appear in newswires to enhance spoken language models for transcription of radio broadcasts). A central challenge will be the rapid learning of explainable and robust systems from noisy, partial, and small amounts of learning material. Community defined evaluations will be essential for progress; the key to this progress will be a shared infrastructure of benchmark tasks with training and test sets to support cross-site performance comparisons. Different users exhibit different patterns when they interact with computer systems. Machine learning algorithms are also needed to recognize such regularities and integrate them into the system, to personalize the system's interactions with its user [33], and to build up user models. User models may seek to describe (1) the cognitive processes that underlie the user's actions; (2) the differences between the user's skills and expert skills; (3) the user's behavioral patterns or preferences; (4) the user's characteristics [99]. In general, there is great potential in tapping into or collaborating with the artificial intelligence and learning research community for new paradigms and models of which neuro-based learning is only one candidate. Learning methods have great potential for synergistically combining multiple media at different levels of abstraction. Note that the current search engines (e.g., Yahoo!, Google, etc) use only text for indexing images and video. Therefore, approaches which demonstrate synergy of text with image and video features have significant potential. Note that learning must be applied at the right level as is done in some hierarchical approaches and also in the human brain. An arbitrary application of learning might result in techniques that are very fragile and are useless except for some niche cases. Bag-of-visual words model (BoW) is one popular method to represent images as visual documents composed of repeatable and descriptive visual words. The basic idea is to cluster a large number of local image descriptors extracted from images, then use the exemplar descriptor of each cluster as a visual word. All of visual words construct a comprehensive visual dictionary, which can describe all the images and with such a dictionary, a lot of text retrieval techniques can be directly applied on image and video retrieval. Although BoW model has been utilized for object recognition [57, 61, 79], image segmentation [102, 106], video event detection [98, 104, 113], and large-scale image retrieval [71, 84], how to generate a visual dictionary for images as descriptive as the text dictionary used for documents is still an open problem. Many works try to capture the special relationship between different visual words to generate more descriptive visual words [57, 61, 79, 106]. Zhang et al. [109] define Descriptive Visual Words (DVWs) and Descriptive Visual Phrases (DVPs) by identifying visual words and their combinations which are effective in representing certain visual objects or scenes. How to generate compact descriptive visual words, visual phrases, visual sentences, visual blocks and construct a comprehensive and efficient visual dictionary for images are still challenges.

Furthermore, services such as Blinkx and Riya currently utilize learning approaches to extract words in movies from complex, noisy audio tracks (Blinkx) or detecting and recognizing faces from photos with complex backgrounds (Riya). In both cases, only methods which are robust to the presence of real-world noise and complexity will be beneficial in improving the effectiveness of similar services.

4.6 Folksonomies

It is clear that the problem of automatically extracting content multimedia data is a difficult problem. Even in text, we could not do it completely. As a consequence, all the existing search engines are using simple keyword-based approaches or are developing approaches that have a significant manual component and address only specific areas. Another interesting finding is that, for an amorphous and large collection of information, a taxonomy-based approach could be too rigid for navigation. Since it is relatively easier to develop inverted file structures to search for keywords in large collections, people find the idea of tags attractive: by somehow assigning tags, we can organize relatively unstructured files and search [43]. About the same time, the idea of the wisdom of crowd became popular. So it is easy to argue that tags could be assigned by people and will result in wise tags (because they are assigned by the crowd) and this will be a better approach than the dictatorial taxonomy. The idea is appealing and made flickr.com and Del.icio.us useful and popular.

The main question arises: Is this approach really working—or can it be made to work? If everybody assigns several appropriate tags to a photo and then the crowd seeing that photo also assigns appropriate tags, then the wisdom of crowd may come into action. But if the uploader rarely assigns tags, and the viewers, if any, assign tags even more rarely, then there is no crowd, and there is no wisdom. There are other somewhat amusing attempts to crowd-source the annotation process for images. The most well known is probably Google's image-labeler developed by Luis van Ahn [114]. The idea is to pair people together and have them try to cooperatively label an image. If their labels match, they get a new picture, otherwise they have to guess again. The game has some variants for improved tagging such as pairing experts in a field together and giving them specific kinds of images. Rather than labeling a picture "car", two automotive experts were more likely to list a model or make. Robots would periodically play, displaying previously entered labels to new users as a form of validation. The images in this way can get assigned labels that are both sound and complete. Interesting game-like approaches (see, e.g., www.espgame.org, Peekaboom [3]) are being developed to assign these tags to images. Based on ad hoc analysis, it seems that very few tags are being assigned to photos on flickr.com by people who upload images and fewer are being assigned by the viewers. Moreover, it may happen that, without any guidance, people become confused about how to assign tags. It appears that the success may come from some interesting combination of taxonomy and folksonomy.

5 Conclusion

Personalized multimedia access, retrieval, and analysis are emerging research areas that received growing attention in the research community over the past decade. Though modeling and indexing techniques for content-based image indexing and retrieval domain have reached reasonable maturity [85], content-based techniques for personalized multimedia data, particularly those employing spatio-temporal concepts, are at the infancy stage. In this survey, techniques have been presented from all along the chain of research involved in designing useful multi-modal information portals for users that are adaptable and customizable. At the beginning of application chain is the ability to decipher atomic pieces of information—a scene in a movie or article in a news paper. From there, these pieces of information have to be annotated (preferably in an automated fashion) and several techniques and models were presented demonstrating some current methodologies. More models have been presented illustrating how data is then organized ontologically and related categorically to each other. They key step from here presentation. IObjects and some xml based techniques were used to describe presentation layout. Layout and context adaptability round out the discussion with several ideas on how to understand what a user actually wants. In each stage of the presentation chain, we find ample opportunities to develop new techniques for making really strong, contextually-aware and preference-based multimedia presentations a common reality.

Acknowledgements We would like to thank Dick Bulterman, Stavros Christodoulakis, Chabane Djeraba, Daniel Gatica-Perez, Thomas Huang, Alex Jaimes, Ramesh Jain, Mike Lew, Andy Rauber, Pasquale Savino, Arnold Smeulders, and the whole FACS consortium for excellent suggestions and discussions. The work of Nicu Sebe has been supported by the FP7 IP GLOCAL european project and by the FIRB S-PATTERN project. The work of Yijuan Lu was supported in part by the Research Enhancement Program (REP) and start-up funding from the Texas State University.

References

1. Agarwal S, Fankhauser P, Gonzalez-Ollala J, Hartman J, Hollfelder S, Jameson A, Klink S, Lehti P, Ley M, Rabbidge E, Scharzkopf E, Shrestha N, Stojanovic N, Studer R, Stumme G, Walter B, Weber A (2003) Semantic methods and tools for information portals. Proceedings of INFORMATIK 2003 - Innovative Informatikanwendungen, pp 116–131
2. Agius H, Angelides M (2007) Closing the content-user gap in MPEG-7: the hanging basket model. Multimed Syst 13(2):155–176
3. Ahn LV, Liu R, Blum M (2006) Peekaboom: a game for locating objects in images, SIGCHI Conference. Human Factors in Computing Systems, pp 55–64
4. Aizawa K, Tancharoen D, Kawasaki S, Yamasaki T (2004) Efficient retrieval of life log based on context and content. ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, pp 22–31
5. Arifin S, Cheung PYK (2007) A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information. ACM Multimedia, pp 68–77
6. Arthur GM, Harry A (2008) Video summarization: a conceptual framework and survey of the state of the art. J Vis Commun Image Represent 19(2):121–143
7. Battelle J (2005) The search: how Google and its rivals rewrote the rules of business and transformed our culture, Portofolio Hardcover
8. Belloti R, Decurtins C, Grossniklaus M, Norrie M, Palinginis A (2004) Modeling context for information environments, ubiquitous mobile information and collaboration systems. Lect Notes Comput Sci 3272:43–56
9. Blei D, Jordan M (2003) Modeling annotated data. ACM SIGIR, pp 127–134
10. Brewer E et al (2005) The case for technology in developing regions. IEEE Computer 38(6):25–38

11. Bruno D, Denis L, Sharon O (2009) Multimodal interfaces: a survey of principles, models and frameworks, human machine interaction. Lect Notes Comput Sci 5440:3–26
12. Bulterman D, Rutledge L (2004) SMIL 2.0: Interactive multimedia for web and mobile devices. Springer-Verlag, Heidelberg
13. Bulterman D, Hardman L, Jansen J, Mullender K, Rutledge L (1998) GRiNS: A GRaphical interface for creating and playing SMIL documents. Comput Netw ISDN systems 10:519–529
14. Chen L, Sycara K (1998) WebMate: personal agent for browsing and searching. Int. Conf. on Autonomous Agents, pp 132–139
15. Chen H, Zheng NN, Liang L, Li Y, Xu YQ, Shum HY (2002) PicToon: a personalized image-based cartoon system, ACM Multimedia, pp 171–178
16. Crystal D (1991) A dictionary of linguistics and phonetics. Blackwell, Oxford
17. Deng J, Dong W, Socher R, Li J, Li K, Li FF (2009) ImageNet: a large-scale hierarchical image database. IEEE Conf. on Computer Vision and Pattern Recognition, pp 248–255
18. Dimitrova N (2003) Multimedia content analysis: the next wave, Int. Conf. on Image and Video Retrieval, pp 415–420
19. Dimitrova N, Zhang HJ, Shahraray B, Sezan I, Huang T, Zakhor A (2002) Applications of video-content analysis and retrieval. IEEE Multimedia 9(3):42–55
20. Dorai C, Farrell R, Katriel A, Kofman G, Li Y, Park Y (2006) BMAGICAL demonstration: system for automated metadata generation for instructional content. ACM Multimedia, pp 491–492
21. eHealth Workshop 2010, http://research.microsoft.com/en-us/collaboration/global/asia-pacific/programs/ehealth.aspx
22. Eynard D (2008) Using semantics and user participation to customize personalization, HP Laboratories Technical Report HPL-2008-197
23. Fergus R, Perona P, Zissermann A (2003) Object class recognition by unsupervised scale invariant learning, IEEE Conf. on Computer Vision and Pattern Recognition, pp 264–271
24. Foote JT (1997) Content-based retrieval of music and audio. SPIE Multimed Storage Archiving Syst II 3229:138–147
25. Gevers T, Smeulders A (1999) Color based object recognition. Pattern Recogn 32:453–464
26. Ghidini C, Giunchiglia F (2001) Local models, semantics, or contextual reasoning = locality + compatibility. Artif Intell 127(2):221–259
27. Giunchiglia F, Serafini L (1994) Multilanguage hierarchical logics, or how can we do without modal logics. Artif Intell 65(1):29–70
28. Guerts J, van OssenBruggen J, Hardman L (2001) Application-specific constraints for multimedia presentation generation. Int. Conf. on Multimedia Modelling, pp 247–266
29. Guerts J, van OssenBruggen J, Hardman L, Rutledge L (2003) Towards a multimedia formatting vocabulary. Int. Conf. on WWW, pp 384–393
30. Hanjalic A (2005) Adaptive extraction of highlights from a sport video based on excitement modeling. IEEE Trans Multimedia 7(6):1114–1122
31. Hanjalic A (2006) Extracting moods from pictures and sounds: towards truly personalized TV. IEEE Signal Process Mag 23(2):90–100
32. Hanjalic A, Xu LQ (2005) Affective video content representation and modeling. IEEE Trans Multimedia 7(1):143–154
33. Hirsh H, Basu C, Davison B (2000) Learning to personalize. Commun ACM 43(8):102–106
34. Hori T, Aizawa K (2003) Context-based video retrieval system for the Life Log applications. ACM Multimedia Information Retrieval Workshop, pp 31–38
35. Hori T, Aizawa K (2004) Capturing life log and retrieval based on context. IEEE Conf. on Multimedia and Expo, pp 301–304
36. http://www.oratrix.com/GRiNS/
37. Hua XS, Lu L, Zhang HJ (2004) P-Karaoke: personalized karaoke system, ACM Multimedia, pp 172–173
38. Infomedia Project, http://www.informedia.cs.cmu.edu
39. Isbister K, Hook K, Sharp M, Laaksolahti J (2006) The sensual evaluation instrument: developing an affective evaluation tool. SIGCHI Conf. on Human Factors in Computing Systems, pp 1163–1172
40. Jaimes A, Sebe N (2007) Multimodal human-computer interaction: a survey. Comput Vis Image Underst 108(1–2):116–134
41. Jaimes A, Sebe N, Gatica-Perez D (2006) Human-centered computing: a multimedia perspective, ACM Multimedia, pp 855–864
42. Jaimes A, Gatica-Perez D, Sebe N, Huang T (2007) Human-centered computing: toward a human revolution. IEEE Computer 40(5):30–34

43. Jain R (2003) Folk computing. Communications ACM 46(4):27–29
44. Jameson A (2001) Systems that adapt to their users. Tutorial presented at IJCAI 2001, www.dfki.de/~jameson
45. Jameson A (2001) User-adaptive and other smart adaptive systems: possible synergies. The First EUNITE Symposium, pp 13–14
46. Kadlek T, Jelenik I (2008) Semantic user profile acquisition and sharing, Int. Conf. on Computer Systems and Technologies and Workshop for PhD students in Computing
47. Kang HB (2002) Analysis of scene context related with emotional events. ACM Multimedia, pp 311–314
48. Klemke R (2000) Context framework—an open approach to enhance organizational memory systems with context modeling techniques, Int. Conf. on Practical Aspects of Knowledge Management, pp 14-1–14-12
49. Lang PJ (1993) The network model of emotion: motivational connections. In: Advances in social cognition. Lawrence Erlbaum Associates, Hillsdale, NJ, pp 109–133
50. Lavrenko V, Feng S, Manmatha R (2003) Statistical models for automatic video annotation and retrieval. Int. Conf. on Acoustics, Speech and Signal Processing, pp 17–21
51. Lee M, Wilks Y (1996) An ascription-based approach to speech acts, Int. Conf. on Computational Linguistics, pp 699–704
52. Lew M, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: state-of-the-art and challenges. ACM Trans Multimed Comput Commun Appl 2(1):1–19
53. Li T, Mitsunori O (2003) Detecting emotion in music. Int. Conf. on Music Information Retrieval (ISMIR), pp 239–240
54. Li X, Yan J, Fan WG, Liu N, Yan SC, Chen Z (2009) An online blog reading system by topic clustering and personalized ranking. ACM Trans. on Internet Technology 9(3) Article 9
55. Liu D, Lu L, Zhang HJ (2003) Automatic mood detection from acoustic music data. Int. Conf. on Music Information Retrieval (ISMIR), pp 81–87
56. Liu B, Gupta A, Jain R (2005) MedSMan: a streaming data management system over live multimedia, ACM Multimedia, pp 171–180
57. Liu D, Hua G, Viola P, Chen T (2008) Integrated feature selection and higher-order spatial feature extraction for object categorization. IEEE Conf. on Computer Vision and Pattern Recognition, pp 1–8
58. Lu L, Liu D, Zhang HJ (2006) Automatic mood detection and tracking of music audio signals. IEEE Trans Audio Lang Process 14(1):5–18
59. Magnini B, Strapparava C (2004) User modeling for news web sites with word sense based techniques. User Model User-Adapt Interact 14(2–3):239–257
60. Mann W, Matthiesen C, Thompson S (1989) Rhetorical structure theory and text analysis, technical report ISI/RR-89-242, November
61. Marszalek M, Schmid C (2006) Spatial weighting for bag-of-features. IEEE Conf. on Computer Vision and Pattern Recognition, pp 2118–2125
62. Maybury MT (1997) Intelligent multimedia information retrieval, AAAI/MIT Press
63. McCarthy J (1987) Generality in artificial intelligence. Commun ACM 30(12):1030–1035
64. Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr Psycho 14(4):261–292
65. Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. Int J Comp Vis 60:63–86
66. Moncrieff S, Dorai C, Venkatesh S (2001) Affect computing in film through sound energy dynamics. ACM Multimedia, pp 525–527
67. MPEG—Moving Picture Expert Group, http://www.chiariglione.org/mpeg/
68. Naphade, Huang TS (2001) A probabilistic framework for semantic video indexing, filtering and reieval. IEEE Trans Multimedia 3(1):141–151
69. Naphade MR, Huang TS (2002) Extracting semantics from audiovisual content: the final frontier in multimedia retrieval. IEEE Trans Neural Netw 13(4):793–810
70. Naphade MR, Kristjansson T, Frey B, Huang TS (1998) Probabilistic multimedia objects (Multijects): a novel approach to video indexing and retrieval in multimedia systems. Int. Conf. on Image Processing, pp 536–540
71. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree, IEEE Conf. on Computer Vision and Pattern Recognition, pp 2161–2168
72. Oviatt S (2003) User-centered modeling and evaluation of multimodal interfaces. Proc IEEE 91 (9):1457–1468
73. Parsons S, Sierra C, Jennings NR (1998) Agents that reason and negotiate by arguing. J Log Comput 8 (3):261–292

74. Quiroga L (1999) Empirical evaluation of explicit vs implicit acquisition of user profiles in information filtering systems, ACM Conf. on Digital Libraries, pp 238–239
75. Rauber A, Pampalk E, Merkl D (2003) The SOM-enhanced jukebox: organization and visualization of music collections based on perceptual models. J New Music Res JNMR 32(2):193–210
76. Rigo S, Jose O (2008) Advanced in conceptual modeling—challenges and opportunities: ER 2008 Workshops CMLSA, ECDM, FP-UML, M2AS, RIGiM, SeCoGIS, WISM. Lect Notes Comput Sci 5232
77. Roy D, Pentland A (2002) Learning words from sights and sounds: a computational model. Cogn Sci 26(1):113–146
78. Russell J, Mehrabian A (1977) Evidence for a three-factor theory of emotions. J Res Pers 11:273–294
79. Savarese S, Winn J, Criminisi A (2006) Discriminative object class models of appearance and shape by correlatons. IEEE Conf. on Computer Vision and Pattern Recognition, pp 2033–2040
80. Schilit B, Adams N, Want R (1994) Context-aware computing applications. IEEE Workshop on Mobile Computing Systems and Applications, pp 85–90
81. Schlosberg H (1954) Three dimensions of emotion. Psychol Rev 61(2):81–88
82. Sebe N, Tian Q (2007) Personalized multimedia retrieval: the new trend? ACM Multimedia Information Retrieval Workshop, pp 299–306
83. Zhang S, Huang Q, Jiang S, Gao W, Tian Q (2010) Affective visualization and retrieval for music video. IEEE Trans Multimedia, Special Issue on Multimodal Afftective Interaction 12 (6):510–522
84. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos, Int. Conf. on Computer Vision, pp 1470–1477
85. Smeulders A, Worring M, Santini S, Gupta A, Jain R (2000) Content based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
86. Snoek CGM, Worring M, Geusebroek J, Koelma D, Seinstra F, Smeulders A (2006) The semantic pathfinder: using an authoring metaphor for generic multimedia indexing. IEEE Trans Patt Anal Mach Intell 28(10):1678–1689
87. Song Y, Hua XS, Dai LR, Wang M (2005) Semi-automatic video annotation based on active learning with multiple complementary predictors. ACM Int. Workshop on Multimedia Information Retrieval, pp 97–104
88. StreamSage, http://www.streamsage.com
89. Sullivan DO, Smyth B, Wilson DC, McDonald K, Smeaton A (2004) Improving the quality of the personalized electronic program guide. User Model User-Adapt Interact 14(1):5–36
90. Torralba A, Fergus R, Freeman WT (2008) 80 million tiny images: a large dataset for non-parametric object and scene recognition. IEEE Trans Pattern Anal Mach Intell 30(11):1958–1970
91. Tseng BL, Lin CY, Smith JR (2004) Using MPEG-7 and MPEG-21 for personalizing video. IEEE Trans Multimedia 11(1):42–52
92. Tsinaraki C, Christodoulakis S (2005) Semantic user preference descriptions in MPEG-7/21. The 4th Hellienic Data Managerment Symposium (HDMS)
93. Tsinaraki C, Christodoulakis S (2006) A multimedia user preference model that supports semantics and its application to MPEG 7/21. Int. Conf. on Multimedia Modelling, pp 35–42
94. Tsinaraki C, Polydoros P, Kazasis F, Christodoulakis S (2005) Ontology-based semantic indexing for MPEG-7 and TV-anytime audiovisual content. Multimed Tools Appl 26(3):299–325
95. Venkatesh S, Adams B, Phung D, Dorai C, Farrell RG, Agnihotri L, Dimitrova N (2008) “You Tube and I Find”-personalizing multimedia content access. Proc IEEE 96(4):697–711
96. Wang HL, Cheong LF (2006) Affective understanding in film. IEEE Trans Circuits Syst Video Technol 16(6):689–704
97. Wang FS, Lu W, Liu J, Shah M, Xu D (2008) Automatic video annotation with adaptive number of key words, Int. Conf. on Pattern Recognition, pp 1–4
98. Wang F, Jiang YG, Ngo CW (2008) Video event detection using motion relativity and visual relatedness. ACM Multimedia, pp 239–248
99. Webb GI, Pazzani MJ, Billsus D (2001) Machine learning for user modeling. User Model User-Adapt Interact 11(1–2):19–29
100. Wei G, Petrushin V, Gershman A (2002) From data to insight: the community of multimedia agents, Int. Workshop on Multimedia Data Mining
101. Weitzman L, Wittenberg K (1994) Automatic presentation of multimedia documents using relational grammars. ACM Multimedia, pp 443–451

102. Winn J, Criminisi A, Minka T (2005) Object categorization by learning universal visual word dictionary. Int. Conf. on Computer Vision, pp 1800–1807
103. Wold E, Blum T, Kreislar D, Wheaton J (1996) Content-based classification, search, and retrieval of audio. IEEE Multimedia 3(3):27–36
104. Xu D, Chang SF (2008) Video event recognition using kernel methods with multilevel temporal alignment. IEEE Trans Pattern Anal Mach Intell 30(11):1985–1997
105. Xu M, Chia LT, Jin J (2005) Affective content analysis in comedy and horror videos by audio emotional event detection. IEEE Int. Conf. on Multimedia and Expo, pp 622–625
106. Yang L, Meer P, Foran DJ (2007) Multiple class segmentation using a unified framework over mean-shift patches. IEEE Conf. on Computer Vision and Pattern Recognition, pp 1–8
107. Yu B, Ma WY, Nahrstedt K, Zhang HJ (2003) Video summarization based on user log enhanced link analysis. ACM Multimedia, pp 382–391
108. Zeng ZH, Pantic M, Roisman GI, Huang T. A survey of affect recognition methods: audio, visual and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58
109. Zhang S, Tian Q, Hua G, Huang Q, Li S (2009) Descriptive visual words and visual phrases for image applications. ACM Multimedia, pp 75–84
110. Zhou M (1999) Visual planning: a practical approach to automated presentation design. Int. Joint Conference on Artificial Intelligence, pp 634–641
111. Zhou XS, Huang TS (2003) Relevance feedback in image retrieval: a comprehensive review. Multimed Syst 8(6):536–544
112. Zhou M, Houck K, Pan S, Shaw J, Aggarwal V, Wen Z (2006) Enabling context-sensitive information seeking, Int. Conf. on Intelligent User Interfaces, pp 116–123
113. Zhou X, Zhuang XD, Yan SC, Chang SF, Johnson MH, Huang T (2008) SIFT-Bag kernel for video event analysis. ACM Multimedia, pp 229–238
114. Von AL (2006) Games with a purpose. IEEE Computer 39(6):96–98

Yijuan Lu is an Assistant Professor in the Department of Computer Science, Texas State University. She received her Ph.D. in CS in 2008 from the University of Texas at San Antonio. During 2006, 2007, 2008, she was a summer Intern Researcher at FXPAL lab, Web Search & Mining Group, Microsoft Research Asia (MSRA), National Resource for Biomedical Supercomputing (NRBSC) at the Pittsburgh Supercomputing Center (PSC), Pittsburgh. She was the Intern Researcher at Media Technologies Lab, Hewlett-Packard Laboratories (HP) 2008, and research fellow of Multimodal Information Access and Synthesis (MIAS) Center at University of Illinois at Urbana-Champaign (UIUC) 2007.

Her current research interests include Multimedia Information Retrieval, Computer Vision, Machine Learning, Data Mining, and Bioinformatics. She has published extensively and serves as reviewers at top conferences and journals. She is the 2007 Best Paper Candidate in Retrieval Track of Pacific-rim Conference on Multimedia (PCM) and the recipient of 2007 Prestigious HEB Dissertation Fellowship, 2007 Star of Tomorrow Internship Program of MSRA, She is a member of IEEE and ACM.

Nicu Sebe received his PhD degree in 2001 from University of Leiden, The Netherlands. He is with the Faculty of Cognitive Sciences, University of Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human-computer interaction in computer vision applications. He is the author of Robust Computer Vision—Theory and Applications (Kluwer, April 2003) and of Machine Learning in Computer Vision (Springer, May 2005). He was involved in the organization of the major conferences and workshops addressing the computer vision and human-centered aspects of multimedia information retrieval, among which as a General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference, FG 2008, ACM International Conference on Image and Video Retrieval (CIVR) 2007, and WIAMIS 2009 and as one of the initiators and a Program Co-Chair of the Human-Centered Multimedia track of the ACM Multimedia 2007 conference. He is the general chair of ACM CIVR 2010 and a track chair of ICPR 2010. He has served as the guest editor for several special issues in IEEE Computer, Computer Vision and Image Understanding, Image and Vision Computing, Multimedia Systems, and ACM TOMCCAP. He has been a visiting professor in Beckman Institute, University of Illinois at Urbana-Champaign and in the Electrical Engineering Department, Darmstadt University of Technology, Germany. He was the recipient of a British Telecomm Fellowship. He is the co-chair of the IEEE Computer Society Task Force on Human-centered Computing and is an associate editor of Machine Vision and Applications, Image and Vision Computing, Electronic Imaging and of Journal of Multimedia.

Ross Hytnen is an Engineering Scientist at the Applied Research Labs of the University of Texas and will soon take a position as lead senior software developer at NanoMetrics in Austin, Texas. He is currently completing his masters in CS in 2011 from Texas State University. His current research interests include image recognition, multimedia information retrieval, and sonar image construction.

Qi Tian is currently an Associate Professor in the Department of Computer Science, the University of Texas at San Antonio (UTSA). During 2008–2009, he took one-year Faculty Leave at Microsoft Research Asia (MSRA) in the Media Computing Group (former Internet Media Group). He received his Ph.D. in 2002 from UIUC. Dr. Tian's research interests include multimedia information retrieval and computer vision. He has published about 110 refereed journal and conference papers in these fields. His research projects were funded by NSF, ARO, DHS, HP Lab, SALSI, CIAS, CAS and Akiira Media System, Inc. He was the co-author of a Best Student Paper in ICASSP 2006, and co-author of a Best Paper Candidate in PCM 2007. He was a nominee for 2008 UTSA President Distinguished Research Award. He received 2010 ACM Service Award for ACM Multimedia 2009. He is the Guest Editors of IEEE Transactions on Multimedia, ACM Transactions on Intelligent Systems and Technology, Journal of Computer Vision and Image Understanding, Pattern Recognition Letter, and EURASIP Journal on Advances in Signal Processing and is the associate editor of IEEE Transaction on Circuits and Systems for Video Technology and in the Editorial Board of Journal of Multimedia. He is a Senior Member of IEEE (2003), and a Member of ACM (2004).