# Personalized Multimedia Retrieval: The New Trend?

Nicu Sebe
University of Amsterdam
The Netherlands
nicu@science.uva.nl

Qi Tian
University of Texas at San Antonio
USA
qitian@cs.utsa.edu

## ABSTRACT
The aim of this paper is to present an overview of the current situation in this hot topic of Multimedia Information Retrieval: Personalization. We are considering several aspects of this problem. On one hand, the user will want to have a personalized access to his image/video collections and this can be achieved by providing intuitive and natural browsing capabilities and customized features. Furthermore, the system is required to perform user profiling and to adapt the existing parameters of the system to the user needs and interest. On the other hand, it is also important to consider the devices and applications in which this technology is going to be deployed. Mobile media is a high growth area but the state-of-the-art technologies are lagging behind the consumers expectations. We are addressing in this paper precisely these three important aspects.

## Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Information Filtering; H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms
Algorithms, Management, Human Factors.

## Keywords
Personalization, Information Access, Multimedia.

## 1. INTRODUCTION
The explosion of multimedia content in databases, broadcasts, streaming media, etc. has generated new requirements for more effective access to these global information repositories. Content extraction, indexing, and retrieval of multimedia data continue to be some of the most challenging and fastest-growing research areas. A consequence of the growing consumer demand for multimedia information is that sophisticated technology is needed for representing, modeling, indexing, and retrieving multimedia data. In particular, we need robust techniques to index/retrieve and compress multimedia information, new scalable browsing algorithms allowing access to very large multimedia databases, and semantic visual interfaces integrating the above components 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. 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 travel 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 on 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 production quality, including segment transitions and music. 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 affect on tourism and accommodation prices across the Mediterranean next summer. Some presentations may be synchronized with a map display which may be accessed interactively. 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" 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


*MIR'07*, September 28–29, 2007, Augsburg, Bavaria, Germany.


[1] Courtesy to the FACS Consortium.

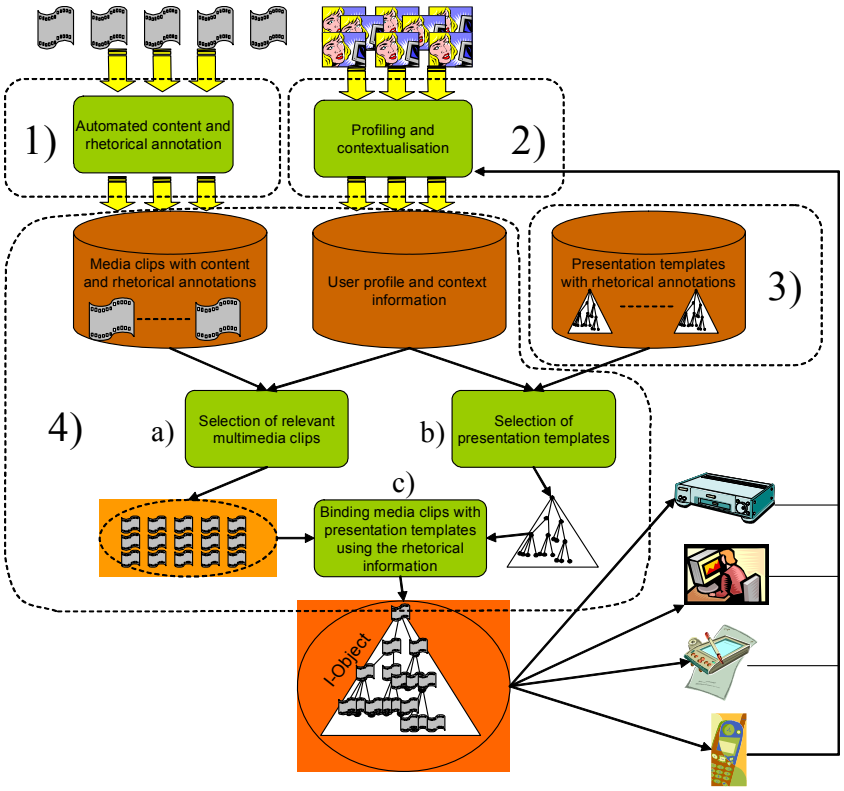


**Figure 1. Automated generation of I-Objects**

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.

Several important research directions may be derived from this vision:

- Representation of multimedia content descriptions
- Representation of user preferences and context
- Automated content annotation
- Automated presentation authoring
- 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 context, and his 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 Figure 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 [28]. 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. **Acquisition of user profiles.** 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.

3. **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.
4. **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 will 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. MPEG-7 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 contextualization 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. **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 [39].
4. **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.
5. **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

## 3.1 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) [1]. 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 [39]. Progress is currently being made in known object retrieval [2][3], while promising results are reported in object category discrimination [4], all based on the invariance paradigm of computer vision. Significant solutions to access the content and knowledge contained in audio/video documents are offered by StreamSage [5] and Infomedia [6]. 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 [7], in combination with the use of machine learning techniques.

Wold et al. [8] presented 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. [9]) 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 [10], where characteristics of frequency spectra are extracted and transformed according to psychoacoustic models focusing on the rhythmic characteristics.

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 importance only recently [11][12][28] as conventional genre categories are seen as too limited, inflexible, and not applicable to the wide range of individual perceptions.

## 3.2 Personalization and Contextualization

McCarthy [13] 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 [14], [15], [16], [17]. 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 [18]. In information systems, context has been described in terms of context types [19] that relate to application types. Context types proposed in [20] include organizational, domain-based, personal, and physical context. Each is subdivided into more specific types like workflow 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 [21].

Personalization in multimedia content is provided by MPEG-7/21 via user profiles and usage history. However, the personalization of MPEG-7/21 is limited and cannot account for the domain knowledge described in specific ontologies and thus it cannot exploit combined MPEG-7/domain-specific metadata descriptions.

## 3.3 Automated Generation of Presentations from Intelligent Objects

There are several existing technologies available that can be used to define storage containers for objects. Formats such as MPEG4 [22] 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 codecs. XML languages such as SMIL [24] 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 [23] [25] 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 codecs) 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. [26], who used a method based on relational grammars, and Zhou [27], 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) [29], but all of these approaches are tailored to generate text descriptions based on a structured markup at the text encoding level.

# 4. CHALLENGES

Despite the considerable progress of academic research in multimedia information retrieval [30], there has been relatively little impact of multimedia information retrieval (MIR) research on commercial applications with some niche exceptions such as video segmentation. 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. 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 [31].

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. 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.

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 [35]. 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 [32].

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) [40].

## 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 [33],[34],[41]. 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., [42] 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 [35] 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 [36] 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 [35][37] 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 [38]. 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.

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.

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. Interesting game-like approaches (see, e.g., www.espgame.org) 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. CONCLUDING REMARKS

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 [39], content-based techniques for multimedia data, particularly those employing spatio-temporal concepts, are at the infancy stage. Content representation through low-level features has been addressed fairly, and there is a growing trend towards bridging the semantic gap. Monomodal approaches have proven successful to a certain level, and more efforts are being put for fusion of multiple media. As visual databases grow bigger with advancements in visual media creation, compaction, and sharing, there is a growing need for storage-efficient and scalable search systems.

# 6. ACKNOWLEDGMENTS

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.

# 7. REFERENCES

[1] C. Tsinaraki, P. Polydoros, F. Kazasis, S. Christodoulakis, Ontology-based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content, *Multimedia Tools and Applications*, 26(3), 299-325, 2005.

[2] T. Gevers, A. Smeulders, Color based object recognition, *Pattern Recognition*, 32*,* 453-464, 1999.

[3] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, *Int. J. Comp. Vis,* 60, 63-86, 2004.

[4] R. Fergus, P. Perona, A. Zissermann, Object class recognition by unsupervised scale invariant learning, *IEEE Conf. on Computer Vision Pattern Recognition,* 2003

[5] StreamSage, http://www.streamsage.com

[6] Infomedia Project, http://www.informedia.cs.cmu.edu

[7] C. G. M. Snoek, M. Worring, J. Geusebroek, D. Koelma, F. Seinstra, A. Smeulders, The Semantic Pathfinder: Using an Authoring Metaphor for Generic Multimedia Indexing, *IEEE Trans. Patt. Anal. Machine Intell*, *28*(10):1678-1689, 2006.

[8] E. Wold, T. Blum, D. Kreislar, J. Wheaton. Content-based Classification, Search, and Retrieval of Audio, *IEEE Multimedia* 3(3):27-36, 1996.

[9] J.T. Foote. Content-based Retrieval of Music and Audio. *SPIE Multimedia Storage and Archiving Systems II*, 3229:138-147, 1997.

[10] A. Rauber, E. Pampalk, D. Merkl, The SOM-enhanced JukeBox: Organization and Visualization of Music Collections based on Perceptual Models. *Journal of New Music Research (JNMR),* 32(2):193-210, 2003.

[11] T. Li, O. Mitsunori, Detecting Emotion in Music. *Int. Conference on Music Information Retrieval (ISMIR)*, 239-240.

[12] D. Liu, L. Lu, H.J. Zhang, Automatic Mood Detection from Acoustic Music Data. *Int. Conference on Music Information Retrieval (ISMIR)*, 81-87, 2003.

[13] J. McCarthy, Generality in Artificial Intelligence, *Communication of the ACM*, 30(12), 1030-1035, 1987

[14] C. Ghidini, F. Giunchiglia, Local Models, Semantics, or Contextual Reasoning = Locality+Compatibility, *Artificial Intelligence* 127(2), 221-259, 2001

[15] M. Lee, Y., Wilks, An ascription-based approach to speech acts, *Int. Conf. in Computational Linguistics*, 1996

[16] F. Giunchiglia, L. Serafini, Multilanguage hierarchical logics, or how can we do without modal logics, *Artificial Intelligence* 65(1), 29-70, 1994

[17] S. Parsons, C. Sierra, N.R. Jennings, Agents that reason and negotiate by arguing, *Journal of Logic and Computation,* 8(3), 261-292, 1998

[18] D. Crystal, A Dictionary of Linguistics and Phonetics, Blackwell, Oxford, UK, 1991

[19] R. Belloti, C. Decurtins, M. Grossniklaus, M. Norrie, A. Palinginis, Modeling Context for Information Environments,

*Ubiquitous Mobile Information and Collaboration Systems*, 2004

[20] R. Klemke, Context Framework – An Open Approach to Enhance Organizational Memory Systems with Context Modeling Techniques, *Int. Conference on Practical Aspects of Knowledge Management*, 2000

[21] B. Schilit, N. Adams, R. Want, Context-aware computing applications. *IEEE Workshop on Mobile Computing Systems and Applications,* 85-90, 1994

[22] MPEG - Moving Picture Expert Group, http://www.chiariglione.org/mpeg/

[23] D. Bulterman, L. Hardman, J. Jansen, K. Mullender, and L. Rutledge, GRiNS: A GRaphical Interface for Creating and Playing SMIL Documents, *Computer Networks and ISDN systems,* 10, 519-529, 1998..

[24] D. Bulterman, L. Rutledge, SMIL 2.0: Interactive Multimedia for Web and Mobile Devices, Springer-Verlag, Heidelberg, 2004.

[25] http://www.oratrix.com/GRiNS/

[26] L. Weitzman, K. Wittenberg, Automatic Presentation of Multimedia Documents Using Relational Grammars, *ACM Multimedia,* 443-451, 1994.

[27] M. Zhou, Visual Planning: A Practical Approach to Automated Presentation Design, *Int. Joint Conference on Artificial Intelligence* 634-641, 1999.

[28] A. Hanjalic, L-Q. Xu, Affective Video Content Representation and Modeling, *IEEE Trans. on Multimedia*, 7(1), 143-154, 2005.

[29] W. Mann, C. Matthiesen, and S. Thompson. Rhetorical Structure Theory and Text Analysis, technical report ISI/RR-89-242, November 1989.

[30] M. Lew, N. Sebe, C. Djeraba, and R. Jain, Content-based Multimedia Information Retrieval: State-of-the-art and Challenges, *ACM Transactions on Multimedia Computing, Communication, and Applications*, 2(1), 1-19, 2006.

[31] J. Battelle, The Search: How Google and its rivals rewrote the rules of business and transformed our culture, Portofolio Hardcover, 2005.

[32] N. Dimitrova, Multimedia Content Analysis: The Next Wave, *Int. Conference on Image and Video Retrieval*, 2003

[33] A. Jaimes, N. Sebe, and D. Gatica-Perez, Human-Centered Computing: A Multimedia Perspective, *ACM Multimedia*, 2006.

[34] A. Jaimes and N. Sebe, Multimodal Human-computer Interaction: A Survey, *Computer Vision and Image Understanding*, 2007.

[35] M. T. Maybury, Intelligent Multimedia Information Retrieval, AAAI/MIT Press, 1997.

[36] B. Liu, A. Gupta, and R. Jain, MedSMan: A streaming data management system over live multimedia, *ACM Multimedia*, 2005.

[37] G. Wei, V. Petrushin, and A. Gershman, From Data to Insight: The Community of Multimedia Agents, *Int. Workshop on Multimedia Data Mining*, 2002.

[38] D. Roy and A. Pentland, Learning Words from Sights and Sounds: A Computational Model, *Cognitive Science*, 26(1), 2002.

[39] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content based image retrieval at the end of the early years, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12), 2000.

[40] N. Dimitrova, H-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, Applications of Video-Content Analysis and Retrieval, *IEEE Multimedia*, 9(3), 42-55, 2002.

[41] A. Jaimes, D. Gatica-Perez, N. Sebe, and T. Huang, Human-centered Computing: Toward a human revolution, *IEEE Computer,* May 2007.

[42] E. Brewer et al., The Case for Technology in Developing Regions, *IEEE Computer*, June 2005.

[43] R. Jain, Folk Computing, *Communications ACM*, 46 (4), 2003.