“You Tube and I Find”—Personalizing Multimedia Content Access

To help users find material of interest in large multimedia collections, researchers are exploring techniques to allow authors to easily capture effective descriptive data, and methods for modeling user preferences.

By Svetla Venkatesh, Senior Member IEEE, Brett Adams, Dinh Phung, Chitra Dorai, Senior Member IEEE, Robert G. Farrell, Lalitha Agnihotri, and Nevenka Dimitrova

ABSTRACT | Recent growth in broadband access and proliferation of small personal devices that capture images and videos has led to explosive growth of multimedia content available everywhere—from personal disks to the Web. While digital media capture and upload has become nearly universal with newer device technology, there is still a need for better tools and technologies to search large collections of multimedia data and to find and deliver the right content to a user according to her current needs and preferences. A renewed focus on the subjective dimension in the multimedia lifecycle, from creation, distribution, to delivery and consumption, is required to address this need beyond what is feasible today. Integration of the subjective aspects of the media itself—its affective, perceptual, and physiological potential (both intended and achieved), together with those of the users themselves will allow for personalizing the content access, beyond today’s facility. This integration, transforming the traditional multimedia information retrieval (MIR) indexes to more effectively answer specific user needs, will allow a richer degree of personalization predicated on user intention and mode of interaction, relationship to the producer, content of the media, and their history and lifestyle. In this paper, we identify the challenges in achieving this integration, current approaches to interpreting content creation processes, to user modelling and profiling, and to personalized content selection, and we detail future directions. The structure of the paper is as follows: In Section I, we introduce the problem and present some definitions. In Section II, we present a review of the aspects of personalized content and current approaches for the same. Section III discusses the problem of obtaining metadata that is required for personalized media creation and present eMediate as a case study of an integrated media capture environment. Section IV presents the MAGIC system as a case study of capturing effective descriptive data and putting users first in distributed learning delivery. The aspects of modelling the user are presented as a case study in using user’s personality as a way to personalize summaries in Section V. Finally, Section VI concludes the paper with a discussion on the emerging challenges and the open problems.

KEYWORDS | Information retrieval; multimedia analysis; personalization; user modeling

I. INTRODUCTION

Much of multimedia information retrieval (MIR) is aimed at intrinsic properties of media, often because this is all an indexing system has access to. The context of the media and its extrinsic (or derived) properties are just as important. With the explosive growth in mobile devices and the attendant multimodal data (GPS, Bluetooth, persistent audio, etc.), we have richer and ever-increasing sources of information with which to associate the context of the media. However, the fundamental problem remains: “What is the content about?” and “What is the context?”
To effectively answer the question for a specific user need, there needs to be an examination of both the content and the intention of the content. For the latter question, especially for personalized media consumption, social context seems to provide an answer; however, the “situational” aspect of the problem remains, and it leads to another challenge: how to include the user in the loop when he/she is critical. For example, given a set of media items and a parallel stream of GPS traces for a user, how do we leverage the knowledge of the meaning of the signatures in those traces to influence the search and browsing behavior of the user? Or how can we infer what the user will like and dislike from a particular video in order to show only the sections that will interest him/her? The user here is the person who creates and consumes the media data.

Traditional MIR indexes are a subset of intrinsic properties of media, such as shot indexes, color features, and motion parameters. These indexes are insufficient for more than trivial access requirements as they do not bridge the so-called semantic gap between the encoded description of media and the user’s conception of it. In addition to these traditional indexes, there are intrinsic properties that are richer and more powerful because they approach the user’s conception of the media. They might include common sense descriptive terms of pictured content, such as children, dogs, sunsets, weather, etc. Extraction of these indexes has been a focus within the multimedia community for at least the last decade. However, there has been an increasing acknowledgement of the importance of extrinsic properties of media. The context, in the widest sense (physiological, historical, social, locational), of a piece of media or, derivatively, the user at capture time, can be just as important as the intrinsic properties, if not more so. For example, photos of your child are probably more significant to you than those of other children, but this familial relationship cannot be discovered from the photo itself. Extrinsic properties are difficult to come by precisely because they lie outside the media, requiring that they be either manually encoded by the user or else inferred from other sensory and/or historical data. Compounding the problem at present is that often only the media item itself is shared or accessed, and even if metadata exists, the lack or noninteroperability of standards means it does not get transported or used. That is the media item itself—for example, image, video—is the currency of exchange, and when shared is ripped from its original web of context and transferred alone before being consumed.

Further, the meaning of information lies in the eyes of the beholder. We all perceive things differently. Usually we want most “relevant” and “appealing.” For example, ratings for different TV shows appeal to different demographics of users as analyzed by Nielsen Media. The user in this case is consuming data produced by others, such as content on television, online on www.YouTube.com, etc. It is also known that gender, age, culture, and social influences play a role in the kinds of information that people like. Personalization has been explored from various angles in the literature. Recommender systems are widely in use for recommending all sorts of merchandise to individual consumers. There are a number of web sites that recommend music, movies, and restaurants among other products based on a user profile (www.launch.com online music, entrée restaurant recommenders, movie recommenders, etc.). When designing content searching systems, it is important to present the summary that will be appealing to the user rather than have a single summary for all users.

In this paper, we will explore the different questions of helping the users personalize content capture, annotate content meaningfully, and to retrieve content created by others in a personalized way.

II. REVIEW OF MULTIPLE ASPECTS OF PERSONALIZED CONTENT AND CURRENT APPROACHES

This section details the current state of art in the related fields of personalized media capture, user modeling, and the current state of understanding of media creation processes.

A. What Are Some Current Methods for Personalized Media Capture and Access?

The user here is primarily the content creator, and the following sections explore methods to help this personalized capture to facilitate efficient access to this media. Effective search and access requires rich annotation of both intrinsic and extrinsic properties of media. However, users are usually loath to do manual annotation, and automated analysis is unable to, and herein lies the impasse. Bulterman [17] notes that “manual [annotation] simply doesn’t get done because creating useful metadata descriptions… is not in the critical path of content creation…” Consequently, some researchers have endeavored to get humans back in the loop in various ways. Any solution to be effective needs to optimize a mixture of when to solicit user input and in what form. Models of the user and the authoring task will need to indicate points at which the user is most likely to be willing to supply annotation; the form of metadata solicited ought to be simple to input and at the same time be of maximal utility for authoring or sharing. The following examples demonstrate a variety of “authoring” paradigms, mixes of computation and interaction, aimed at extracting annotation from a user.

Nack and Putz target news production and use a smart camera to record capture parameters that would otherwise be difficult or impossible to extract automatically, such as camera lens movement and state, and camera distance and angle [43]. A handheld device provides the user the ability to annotate in–out points and conceptual dependencies.
between scenes at the point of capture when the material and context is most fresh in the user’s mind. Automatic synchronization between the camera and the handheld ensures metadata is attached to the correct scene. Conceptual dependencies are used to automatically group scenes in an editing suite, and simple video production rules are used to clip and juxtapose shots. The editing tools further motivate the user to perform annotation during capture.

An active capture paradigm is proposed in [23] casting media capture as a “control process with feedback,” in order to obtain media assets of suitable quality and annotation. The system assumes the role of director and interacts with the user in one of four modes, ranging from directed performance (for example, the user is directed to scream), to agit prop (for example, the user’s response to an unexpected stimulus, such as Boo!, is recorded). Active media templates (AMTs) are used to create personalized content from these known entities. Media properties are able to be automatically assessed by virtue of the restricted context provided by the active capture’s direction. For example, assuming the user follows instructions and screams, rather than, say, laughs, the length and volume of the scream can be verified automatically using its audio envelope.

A computation at capture time to provide creative input and video production expertise is used in [14]. Manual, free-text annotation attached to a shot is used to seed the generation of shot suggestions by means of commonsense reasoning (Open minded common-sense database—for example, “Resting is an activity that follows running”) and video syntax rules (For example, “Shoot events that catalyze other actions in extreme close up then get a shot of the action triggered”). Accepted suggestions are attached to shots as metadata, and the resulting material has greater narrative potential come edit time.

In the context of the MyLifeBits project, Gemmell et al. [31] provide opportunities for lightweight mark-up of photos at likely opportunistic points, such as at capture time, photo import, search and browse tools, and even a screensaver. Mark-up includes voice/text annotation, thumbs up/down ranking, and whether a photo should be available for public or private consumption. This mark-up, together with usage logging, can then aid storytelling by automatically arranging the most popular media items or events in timelines or exporting them in an attractive format to a blog.

B. What Are Current Methods of Modeling the User—In the Wild?

The previous section included capture paradigms that involve the user “in the wild,” i.e., where the user is moving and interacting within their daily, embodied life. It raises the question of how best to model the user in various physical settings. Recently, location and social interactions among users have been identified as important dimensions to be characterized and discovered. At the simplest level this includes individual signatures of users such as where they are, what they are doing, and patterns of these signatures in daily life. At a more complex level, the inter-penetration of our lives in social interactions, both pair-wise and in groups, are useful signatures: Who do we spend time with? Where do we meet? What sorts of activities do we do together? These aspects form a pivotal basis for media sharing and creation, useful for modeling the user in the roles of content creator and content consumer.

Location is an important piece of metadata, as it allows positioning media in space when the user is a content creator, enhancing the simple temporal positioning of media and facilitating better access capability. Much work has been directed towards the recovery of user locations and the understanding of behavior, and one of the most commonly used sensors is GPS. GPS data has the advantage of being collectable over large time periods with low storage cost but suffers from several issues such as noise, especially in urban spaces, and missing data, in particular when users switch off devices, devices run out of power, or fail to get a fix of the coordinates. Ashbrook and Starner [11] performed initial work on extracting locations using K-means clustering of GPS data, and Kang et al. [35] improve this by using a time threshold-based filtering to enhance the clustering performance. The work of Zhou et al. [53] is interesting in dealing with the aspects of noisy GPS data and uses density and spatio-temporal clustering to discover personal paths. Adams et al. [2] first discover places a user visits in a day and then stays within days using a density-based clustering method to deal with the noisy GPS data. In particular, they use interpolation to deal with the missing GPS data that occurs indoors, precisely at places of interest. Heuristics are used to automatically label places as one of home, work, or other. A simple, static measure of the strength of relationship between two people, termed social tie, is also formulated, which can be weighted by the nature of the locations a pair of users shared. For example, time spent together at the location labeled home could be weighted higher than time spent elsewhere. This is an initial attempt at extracting what is increasingly recognized as an important aspect of media ecosystems: interactions among people.

Other sensors used to recover location-based information include Bluetooth, cell phone tower ID, WIFI, and fusions of these sensors. Locations are extracted by clustering cell data in the work of Nurmi and Koolwaaij [45]. A richer set of sensors is used in the work of Clarkson [18], who extracts more complex “life patterns” by clustering and then classifying audio and video data. Handheld GPS data is used by Hariharan et al. [33] to extract location data by using both Markov and non-Markov models.

All the above work has concentrated on extracting the simplest information about a user: where they are. However, encoding of interactions between users arguably
opens the way for even richer applications. For example, it can serve as the basis for interest metrics for sharing geo-positioned media between two users who are co-present, or alternatively, two users who have a history of interaction, but are not co-present at an event.

Early work presents interesting results in this area. Choudhury and Basu [19] use a sensor package consisting of infrared, speech, and motion sensors to understand interactions between people, in particular turn taking in conversation. The group from MIT, in particular Eagle [28], examines human behavior and its dynamics using a rich array of sensors: Bluetooth, cell tower ID, calls logs, phone status, and application usage in conjunction with self surveys. They propose an interesting concept called Eigen behaviors, which examines a sequence of locations a user has visited to find useful and repeatable patterns. At a much larger scale, Ratti et al. [46], examine signatures of cell phones and related services to understand patterns of migration across cities. Adams et al. [4] present algorithms for extracting social rhythms, which are defined as “Latent pursuits of daily life give rise to repeated occurrences along the dimensions of people, place and time.” These activities leave footprints in the sensors data, such as location (via GPS) and co-presence (via Bluetooth), and projections onto these different dimensions uncover inherent to activities. For example, projections on time reveal activities bound by institutional timetables or the structure of their content; projections on place reveal the presence of location-bound elements vital to an activity; and projections on people can uncover activities that are constituted by who must be present. Complex spatio-temporal constructions like rhythms are required to tease apart activities occurring at the same location, as location alone is an insufficient query filter for many application domains (for example media capture hotspots like the home).

C. What Are Current Methods in Automatically Modeling the User and What Is the Current State of the Art in Automatic Generation of Personalized Summaries?

The previous sections examined methods to aid the user as a content creator and model their behaviors in the “real-world.” This section explores the user as a content consumer. For example, YouTube is a perfect example of a great but not fully usable resource. Once the user enters a keyword and retrieves so many videos back there is no way to preview all these hours of video. They may spend hours downloading and when watching one, decide that it was a miss. A powerful personalized summarization method is desperately needed to save user’s time and network bandwidth.

Although the literature abounds on video summarization, little attention has been given to personalized video summarization and modeling the users who consume the content. As we know, content producers make different trailers of the same movie in order to appeal to specific audiences in different demographics. General video summarization is becoming unsatisfactory as the amount of content data grows beyond our ability to search for the right content and consume it quickly and effortlessly. According to a panel that looked into summarization issues, it was agreed that as different people like different things, summaries need to be personalized [7]. Personalization requires implicitly or explicitly collecting user information and leveraging that knowledge in the content delivery framework to select what information is presented to the users. A powerful approach for summarization involves:

1) the personalization of subject matter (syntax and semantics);
2) how it is presented (morphology);
3) where and when it is presented (context).

An approach for personalizing summaries (syntax and semantics) based on a user profile and multimedia content features is needed.

Personalization of multimedia summaries is a fairly new topic and is getting attention in this day and age of personalization and customization. Personalization can be performed at either category level or content level. Category level filters incoming content at the level of the category classification of a video segment, for example, a news segment on weather, an entire segment of a guest on a talk show, etc. Content level takes pieces within these segments to present to the user, for example, segments containing pitches in a baseball game, segment containing faces larger than a certain threshold, etc.

An approach for category level personalization of indexed content is explored by Merialdo et al. [39]. They present a way for personalizing the presentation of indexed content for a user. The authors segment a news program using video analysis and then manually check it and add annotations for the category and importance of the news story. The users can adjust the value of the interest they have in the predefined categories. A simplified solution to the “knapsack” problem is then provided which enables the system to automatically select news stories that for a user based on their importance, duration, and the user interest. The personalization, however, is on a very coarse level and requires manual annotation. A method for automatic personalization at content level is presented by Dimitrova et al. in a system called Video Scout [24]. The output of the system annotates the content (guests in a talk show, category in a new program, etc.) to enable the user of home video storage (like TiVo) to view only segments and/or guests they are interested in, specified by keyword “magnets.”

For a content level personalization, Babaguchi et al. [13] present personalized abstracts of sports videos based on favorite teams, players, and events. They first detect significant events in the video stream by matching textual overlays appearing in an image frame with the descriptions of gamestats in which highlights of the game are described.
Then, they select highlight shots which should be included in the video abstract from those detected event, s reflecting on their significance degree and personal preferences, and generate a video clip by connecting the shots augmented with related audio and text. An hour-length video can be compressed into a minute-length personalized abstract. This is a good start for creating personalized sports summaries. However, most genres do not lend themselves to generation of a profile that can be specified easily by the user.

Hanjalic et al. propose affective video content analysis [32]. They model the intensity and type of feeling or emotion expected to arise while watching a clip, and portions with high affect values are presented as summary. However, this is bound to be different for different people and a way to model this is not yet explored.

D. What Is the Current State of Understanding of Media Creation Processes and Their Role in More Effective Content Interpretation?

The previous sections examined the state of the art from the perspective of a user as a creator and a consumer. In this section, we analyze the role of an intermediary between the creator and the consumer that allows content to be processed and annotated in a meaningful way that reflects the intent of the creator as well as aids the consumer to efficiently access the right content. The intermediary can be people who tag the video content at varying differing levels of details as needed, such as video annotators and librarians at the CNN studios who make content available for various programs broadcast by CNN.

On the other hand, in the presence of increasing volumes of media data, the intermediary is often desired to be automated video analysis and annotation systems which can handle the scaling of the data efficiently and consistently. A video, from the perspective of this intermediary, therefore can be considered to be a structured visual and aural presentation comprising frames at the lowest display level, shots at the fundamental capture level, with scenes and segments occupying increasingly higher levels of story semantics as depicted in the video. Long videos (for example, movies) sometimes have threads at an even higher level embodying multiple story segments which are shown using parallel film editing techniques. For the analysis purposes of the automated system, a shot is an unbroken sequence of frames captured between a camera’s record and stop operations and thus forms a basic unit amenable for algorithmic analysis and for annotation.

There is a plethora of techniques in [22] to detect shots or sequences of homogeneous frames as a first level segmentation of the video, using pixel differences, color or gray level statistics such as histograms of frames, or even motion.

Regardless of whether the event presented in the video is produced professionally (a scripted story as in a movie or a live broadcast sport event), an underlying structure of presentation of shots is exploited to convey the meaning or the mood of the event that is larger than the interpretation of individual shots [10]. Accepted rules and techniques are used during film/video production to solve problems presented by the task of communicating a story to the audience [10]. These rules relate to different cinematic aspects including editing, cameras, motion, and action. An overall impression or mood of the video is brought about by using a very specific pattern of shot arrangement. For example, when a sudden freeze of the image on the screen occurs interrupting the flow of the motion, it is used to either convey the termination of the shot/scene or the conclusion of the story itself. Thus, to understand the semantics of the video for better annotation, characteristics across different shots need to be analyzed. To accomplish the shift in focus from analyzing within-shot characteristics to those between shots requires more than frame level processing and more than simple frame or shot level features.

For the automated video analysis intermediary to successfully annotate, high-level understanding interpretation needs to be derived from the way the shots are strung together and from higher order features of the content. As a case in point, a characterization of motion syntax and semantics (motion patterns that stand in as shortcuts in understanding, and their rules of composition) [25] can provide useful vocabulary to the end user and assist in articulating content descriptions that she has in her mind to aid in her search and even crafting the style, mood, or content descriptions that she has in her mind to aid in his/her search.

The most promising approach for the automated intermediary has been to take a top-down approach, keeping user expectations within the design loop, and finding the means to transform them into computable attributes and measures from video. This approach builds on top of previous systems that describe, for example, only observed motion and object motion, by investigating the question of what higher level of constructs exist in which motion plays a crucial role enabling better video appreciation and concise description. For example, building layers of meaning arising out of motion in video is motivated by the manner in which motion is used in different contexts to imply multiple things. At the lowest level of description of a video, motion can be used to derive an overall sense of what the movement was say, in a contiguous set of frames. A camera swiftly tracks the basketball from the left end of the court to the right basket in a sports video and this camera pan can be detected and described with both direction and magnitude. At the next level, this can be further refined into individual objects and their motion if the system expects to handle queries relating to specific object movements, for example, a person running in a given scene. This gives rise to another layer of descriptions where we target our attention to extraction of individual objects and to tracking their movements across frames and activities associated with those identifiable motion.
patterns: walking, running, etc. The next layer of semantics includes events that denote coordinated sequences of specific activities in a given domain, for example, a three-point shot in basketball. Their descriptions stem from a domain-specific, finite, commonly accepted vocabulary for a specific task such as sports event annotation. For detailed annotation, one can envision a sports database of stored motion event models based on detection of coordinated sequences of discrete motion activities, and video search and browsing involves indexing into the stored models of movements.

An even higher semantic layer deals with the sequencing of shots of certain motion together to portray a certain effect, a mood, or an impression that make an impact on the viewer, and hence are memorable enough to be searched for later. A good example of this is the description of tempo of a movie. Motion in a sequence of shots portrayed in a certain way indicates the pace of the movie, for example, sluggish, steady, or fast. This vocabulary of expressive elements is particularly employed by movie critiques to concisely summarize a movie’s appeal in their reviews. They follow the motion constructs embodying tension or different kinds of action scenes that aid and describe the story flow, thus occupying higher levels of our semantic descriptions layers.

This computational framework serves in designing a smart intermediary that analyzes and understands the content of a video and its form using a layered approach, resulting in multiple benefits: First, it can provide useful high level vocabulary to content consumers in multiple contexts. Second, even if a user simply provides a video clip or a segment and requires that the system retrieve all stored sequences that best resemble the query, this query can be processed in a multiresolution fashion to derive possible content descriptions at multiple levels. This multilayered description is likely to succeed in securing the most similar video from the database, thus allowing for multiple intents behind the search. The layers of motion descriptions and their meanings also provide mechanisms for smart “visualization” of shots. Further, this study of flow and dynamics across shots assists in the classification and labeling of shots based on actual content as well as their conveyed meanings, useful in personalizing the results from a search.

1) Computational Media Aesthetics and Why It Can Help: Following this motivation of basing automated analysis of video on sound media production principles, Dorai and Venkatesh [26] proposed an algorithmic framework called the Computational Media Aesthetics that allows for a automated understanding of the dynamic nature of the narrative structure and techniques via analysis of the integration and sequencing of audio–visual elements and is targeted at bridging the semantic gap and building effective content management systems at higher levels of abstraction and meaning. Zettl [52] defined media aesthetics as a study and analysis of media elements such as lighting, motion, color, and sound both by themselves and their roles in synthesizing effective productions. Computational Media Aesthetics is the algorithmic study of a number of image and aural elements in media and the computational analysis of the principles that have emerged underlying their use and manipulation, individually or jointly, in the creative art of clarifying, intensifying, and interpreting some event for the audience.

The Computational Media Aesthetics approach, guided by the broad rules and conventions of content creation, has used media production knowledge to elucidate the relationships between the many ways in which basic visual and aural elements are manipulated in video and their intended meaning and perceived impact on content users. It analyzes videos to understand the production grammar in particular and uses the set of rules that are followed during the narration of a story, to assist in deriving the high level description of video content effectively. A system built using this approach where videos are analyzed guided by the tenets of film grammar will be effective in providing high-level concept-oriented media descriptions that can function across many contexts and in enhancing the quality and richness of descriptions derived.

Several papers studies have explored the workings of Computational Media Aesthetics when applied to extraction of meaning using many of the aesthetic elements introduced by Zettl: Time, sound, and color. Adams et al. [1] explored film tempo, and showed us that though descriptive and sometimes fuzzy in scope, film grammar gives us rich insights into the perception of subjective time and its manipulation by the makers of film for drama. Moncrieff et al. [40] have presented a study of the sound element, particularly sound energy dynamics and its manipulation to affect drama in film. Ba Tu et al. [49] applied this approach to extract structures in video based on color that bring out the interweaving of different themes and settings in a film. Subsequent research [30], [41] has applied this approach widely from extracting mood in music to adding musical accompaniment to videos.

2) User-Centered Design of Multimedia Content Analysis and Metadata Generation Systems: In this section, we explore the role of humans during content tagging and annotation. The traditional creators of metadata are professional librarians and indexers. However, with the growth of web-based multimedia collections, largely due to more egalitarian participation, users without professional expertise are now routinely tagging videos, images, web sites, and other resources. However, these users typically lack the expertise of professionals, especially for the task of organizing large numbers of multimedia objects for effective browsing. More research is needed to determine what kind of expertise professional librarians and indexers possess and whether their knowledge can be captured in automated metadata generation systems. Clearly, the
usability of tools to support manual and semi-automated metadata tagging is becoming increasingly important [21].

Popular Web sites such as YouTube and Flickr use collaborative tagging. Users can reuse existing tags or make up their own. However, one study showed that as the user population grows, the efficiency of information retrieval based on user generated tags tends to decrease due to content diversity [20]. In addition, the quality of user-generated tags from one community may make the content unusable for another [15].

For more controlled collections of corporate and educational materials, metadata tagging processes may need to be more formalized. A first step is to identify users with the right expertise. For example, descriptive tags may be best supplied by the author, whereas taxonomic tags may require more domain broad expertise. Contextual tags, such as those identifying an intended audience, may need to be supplied by those with experience in using the content, such as a teacher or an instructor.

In many cases it is not operationally feasible for people to supply all of the metadata. Legacy databases may contain content that is disorganized, out of date, or inaccessible. Authors may be too busy or no longer available to supply metadata for this content. Even when proper processes are in place, some metadata may still be too costly to collect. For example, Farrell et al. [29] found that domain experts took between two and four hours to extract and tag 200 30-minute learning objects with a subset of IEEE Learning Object Metadata.

Automated metadata generation systems hold promise for reducing the amount of human effort involved in creating effective metadata. IBM’s MAGIC system [27] is able to generate critical learning object metadata that would otherwise take a large amount of manual effort (see Section IV). Liddy et al. [38] created an automated metadata generation system that compared favorably with human taggers, but most systems have not been evaluated in real settings.

Perhaps the most promise is with systems that combine content analysis with collaborative tagging. For example, Aurnhammer et al. [12] show how visual features from content analysis can be used to improve collaborative tagging. Automated systems may also be trained to filter out tags that are not useful for the community, reducing “tag spam.” Further research is needed to determine when and how collaborative tagging and automated metadata generation can be combined.

III. Obtaining Metadata for Personalized Media Creation and Access

This section presents alternate ways of acquiring and using metadata for personalized media creation and access. We begin by considering a case study in three sections: Integrated Media capture environments, then survey user-centered design of multimedia content analysis and metadata generation systems (Section IV), and finally consider automatic metadata generation for personalized media search using focused modeling of personal traits (Section V).

A. eMediate—Case Study of Integrated Media Capture Environments

As pointed out in Section II-A, obtaining rich metadata requires placing humans in the loop of media capture in a more involved manner than simply as device carriers. This brings potentially competing objectives into play: minimal disruption to the user’s modus operandi with extraction of maximally useful information from them. Integrated media capture environments (IMCEs) attempt to find an optimal balance between these criteria by taking a holistic perspective. An IMCE can be defined as a system and methodology that supplies computation to most or all of the media authoring process—from conception or capture, through editing, to publishing, and even repurposing; and it is potentially augmented by sensors in addition to the primary media capture apparatus (for example, some digital cameras now have integrated GPS sensors). With this setting, IMCEs attempt to use computation to support the user by automatically extracting what metadata they can, propagating user supplied metadata using context-primed analysis techniques, dynamically determining likely times to solicit information from the user, and even supplying timely help in the form of encoded domain knowledge (for example, editing rules for video production or composition rules for photography).

Adams and Venkatesh [2] present eMediate (see Fig. 1), an IMCE embodied in a smart camera, aimed at creating home movies of high quality and rich annotation. We will consider it as an example of an IMCE and examine how it attempts to balance the competing objectives mentioned above.

Video creation with eMediate (and subsequent development) involves a number of simple interactions.

1) Elicitation of the user’s filming context—eMediate seeks to gather information about what the user is filming, including where they are and who and what they are interested in. This information is used both to offer useful suggestions and annotate recorded video. Input from the user is solicited in a passive manner: they may ignore the system and film as if the camera were dumb, choose from default suggestions for any or all of the metadata types, based on device history or other contextual clues such as location, or input via voice or text, which is filtered via part-of-speech tagging into atoms suitable to be selected from a list. They can also indicate the desired style of video from a selection of commonly understood genres.

2) Issue of shot suggestions—At any point during filming following the input of any information (even default), the system can begin offering shot
suggestions. A suggestion consists of a configuration of a number of primitive cinematic elements, such as subjects, camera motion, duration, framing type, and so on. This is an example of a suggestive interface [34], a good paradigm for putting humans in the loop using computation to remove as much effort as possible. The availability of suggestions is indicated discreetly, leaving the user free to ignore them. Suggestions are generated on the fly as the output of attempting to optimize a measure that takes account of all available information (who, what, where, intended style), video captured already, and video production rules. This is akin to active capture [23], in the sense that the camera issues directions, with the difference being they are not critical to the authoring process. From the user's point of view, the value eMediate adds at this point is either as a shepherd helping them to achieve an explicit goal or more simply as creative input (in a weaker sense to Barry [14], which uses commonsense reasoning to augment shot suggestions). The richness of suggestions degrades gracefully depending on the amount of information the user is willing to input.

3) **Shot capture**—The user may capture a shot as per a dumb video camera or attempt to follow a suggestion. In the latter case, a one-click protocol is used to attach the metadata embodied by the suggestion to the captured shot: the user indicates whether they performed the shot as suggested or not. A positive response is similar to a thumbs-up, ala Gemmell [31], and includes the shot in the final video, with the added action of associating the structured metadata of the suggestion (such as who or what is in it) with the shot. If the user fails to verify successful capture, verification of low level features, such as motion type and direction, is still performed. Extraction of higher level features, such as a confidence value that a certain presence is pictured in the shot, can be bootstrapped by human verification of other shots if they have done so.

4) **Preview (Edit)**—Editing is performed automatically, and the video being authored can be previewed at any point during filming. Effective editing is enabled by the metadata obtained through the use of shot suggestions and includes appropriate shot transitions, audio overlay, and frame-level edits that attempt to satisfy properties such as the desired duration or motion level, or higher level aesthetic or narrative properties, such as movie tempo, rhythm, or dramatic structure.

5) **Archival and sharing**—Movies are saved in a flexible proprietary format that preserves all of the metadata. Reparameterization allows movies of, say, a different style to be produced. Movies can also be exported to a final, easily shared format, such as MPEG-2. The smartcamera, being also a handheld computer, has wireless communication, and authored videos can be uploaded to a server for, say, publishing in a blog, even when partially complete.

How does an IMCE like the one described above obtain rich metadata by putting humans in the loop? The kind of metadata this IMCE captures is event and cinematic-centric—who, what, where, how (style). In addition to being useful information for media search, it is information immediately useful for the authoring task itself. While it is not on the critical path of authoring (the user can ignore suggestions and the verification protocol), it is yet relevant and useful to the activity the user is engaged in. This is one inducement for getting the human into the loop (casting annotation as a game is another, for example, www.espgame.org or Peekaboom [50]). Moreover, the computation available in an IMCE setting can augment the information the user does provide. For example, if the user verifies that a couple of shots contain a specific person, this information can seed a subject-detection algorithm, which can then verify other shots automatically with higher performance than context-less analysis. (For example, Naaman et al. [42] seek to aid manual annotation of people in photo collections by providing short lists based on existing partial manual annotation of the collection. Short
listing is based on confidence values which can be assigned as fuzzy annotation in its own right.) This ability to constrain the context to the point where algorithms become tenable for extraction of high-level features is akin to how active capture [23] functions. Computation making a “little go a long way” is further inducement to the user to enter the loop, as are immediate and automated editing at any time, and, on a longer scale, organization of media via metadata and better quality video (with derivative inducements via feedback from others, from a simple thanks to explicit popularity ranking of media items etc.). For example, Adams and Venkatesh [3] present a personal media browser that indexes photos, videos, and movies generated by eMediaTE, in a 3-D spatio-temporal context. Information entered as part of authoring eMediaTE movies becomes available in this case as a rich index for searching and filtering (for example, objects or people of interest).

B. Automatic Metadata Generation for Personalized Media Summary Using Focused Modeling of Personal Traits

A way to generate metadata automatically for content that is created by others is desirable. The presence of such metadata can enhance the user experience in getting the right content recommended to them. This section presents the metadata required for generating personalized summaries. Agnihotri et al. [8] present that it is possible to use personality traits in order to generate summaries that are personalized for users. This is useful in this YouTube world where people find more content then they can consume. It is known from commercial media research performed that different TV shows appeal to different demographics of users. Further, people relate to one another differently based personality. Reeves and Nash [47] present the Media Equation states that people react to media the same way they interact with other people. Thus, the underlying assumption is that there exists a mapping of the personality traits to multimedia content features and content features influence preferences for summarization.

In order to generate personalized summaries, the multimedia content needs to be mapped into their “personality types.” This is the metadata that is required for each multimedia segment. Fig. 2 shows the flow of a personalized summarization algorithm. Agnihotri et al. [9] present an algorithm that starts with the selected video genres and audio, video, and text features that are extracted automatically. A user test is first done in order to find a mapping between the users’ personalities to the multimedia features that they chose as the most appropriate summary. This mapping between the features and personality traits varies for different genres. The mapping is used to transform the multimedia content features into personality traits and a personality classification vector for video segments is obtained. This vector is now used to generate personalized multimedia summaries. It can additionally be used for generating recommendations based on user’s personality and for retrieving and indexing media according to user’s personality type. For generating personalized summaries the user profile is projected on to personality classification to obtain the importance value of the video segments for the user’s personality. Segment selection constraints are then applied on the importance values to select the summary segment that is personalized for the user.

IV. MAGIC—CASE STUDY IN PUTTING USERS FIRST IN DISTRIBUTED LEARNING DELIVERY

A system called MAGIC (“Metadata Automated Generation for Instructional Content”) [27], [37] was developed by IBM to assist learning content authors and course developers in creating modular learning content with rich metadata. Many organizations are creating repositories of learning resources that conform to a standard, typically the SCORM [5]. SCORM (Sharable Content Object Reference Model) specifies how media should be modularized into sharable content objects with associated run-time behavior. SCORM sharable content objects can then run on any SCORM compliant system. SCORM incorporates the IEEE Learning Object Metadata (LOM) standard to enable searching and browsing using a common set of attributes defined by the international learning technology community. Currently, values for these attributes are entered manually as metadata on each learning object, a labor-intensive process.

In building MAGIC, we attempted to extract useful metadata from thousands of documents including web pages and videos from U.S. Department of Homeland Security (DHS) web sites. Our goal was to enable DHS
agencies to better retrieve, aggregate, and exchange web-based learning content under quickly changing conditions. The MAGIC system automatically generates SCORM metadata sufficient to register and describe assets for distributed learning applications. It provides a web-based user interface enabling authors to review and edit metadata at the time when the content is created, but also when it is used for browsing or retrieval or packaged for use in courseware. Using MAGIC, course developers assign metadata so that they or other users of the system can later find and aggregate across repositories. Automatic generation of metadata improves consistency and completeness and enables more timely usage. Automated content analysis with user augmentation improves the quality of metadata when compared with manual processes.

The MAGIC system incorporates several software tools to analyze training documents, instructional videos, and other learning assets. Based on our prior efforts in creating and deploying manual learning object metadata creation and editing tools [29], we first identified three LOM elements that are time consuming to enter, not typically provided by editing tools, and are amenable to automation: the description, keywords, and classification of the learning object in a subject taxonomy. We initially applied our text, audio, and video content analysis algorithms on a sample of over 100,000 web pages, 1,000 documents, and seven videos, all drawn from public DHS web sites and other government agencies to evaluate the accuracy of the metadata generated and to identify areas of improvement. We integrated the improved tools into a common architecture.

A. MAGIC System

The MAGIC system (Fig. 3) consists of a Metadata Generation Environment (MGE) and a Metadata Editor. The MAGIC user (a content author or course developer) interacts with the system via the Metadata Editor. The user accesses a training document or an instructional video by entering a uniform resource locator (URL) or the name of a local file. If needed, the converter converts documents to text. MAGIC then loads the content into the content cache where it is processed by the Metadata Generator, which creates metadata and stores them in the Metadata repository DB. The user can view and correct the metadata. After generating and correct metadata for a collection of instructional videos or training documents, the user can request the Packager to create a SCORM-compliant package for export to Web sites, SCORM-compliant authoring systems, learning management systems, or learning content repositories.

The MGE consists of a set of text and video processing tools integrated through a common set of application program interfaces.

1) Text Analysis Tools—These tools generate a title, keywords that include important people, places, and organizations found in the document, and a summary description. These tools leverage high-speed natural language processing techniques to parse the text into linguistic tokens, identify sentence boundaries, and determine the part of speech (for example, noun or verb) for each token. These same tools are applied to closed caption transcripts of video.

2) Text Categorizer Tool—This tool includes a taxonomy—a hierarchical organization of topic categories, such as “Anthrax” and “Suicide Bombings”—and a text classifier component for automatically and precisely assigning documents to categories in this taxonomy. We have extended the taxonomy by developing new categories for Web pages from public DHS web sites.

3) Audio and Video Analysis Tools—These tools segment videos to recover narrative structure (for example, instructor speaks, a slide with information is shown, the instructor speaks again) using image and signal processing algorithms and machine learning techniques. These tools identify visual features (for example, a person’s face) and aural features (for example, music playing). The structural segments obtained are further annotated with text features extracted by the text analysis tools from time-stamped closed caption text present in the video.

We refer the reader to [25], [29], and [37] for details of these content analysis tools. The system is implemented in Java as a multitier architecture which includes a client, a web application server, and a relational database, and is delivered via the Web.

B. Creating Metadata With MAGIC

To create learning object metadata, MAGIC first accesses the content over the Web and generates short title, a summary, and ten important words or phrases.
Fig. 4 shows the metadata generated by MAGIC for a web page. MAGIC does not store the media content, only the URL and the metadata.

Users can browse the topic taxonomy to see how the system classified a particular document. Each document is assigned to one category, which can have multiple paths through the hierarchy. An example is shown in Fig. 5. Anthrax is the lowest level category under Terrorism, which appears under Terrorism and Counterterrorism.

MAGIC can generate metadata for a video using a close caption transcript, if one is available. However, it can also segment the video at two different levels of granularity, macro level and micro level.

For videos with a well-planned plot and distinct topic structure, the macro-segments can capture their high-level content structure very well. Moreover, the keywords identified for each macro-segment seem to capture the various subjects of discussion. Fig. 6 shows one such example, where a macro-segment is identified and displayed along with its thumbnail images and component segments. The thumbnail images provide users a visual overview of the macro-segment. The macro-segment’s topical keywords are also listed, which include anthrax, anthrax bomb, and offensive biowarfare research program. This 2 minute and 17 second long segment (from 02:29 to 04:46) makes a suitable learning object about the use of anthrax as a weapon. We can then select this segment as a video clip (which is marked as blue) and export it into an edit decision list (EDL) for use with video editing software. The “micro” segments reflect points where the visual or auditory characteristics of the video shift (see Fig. 6, in the middle of the screen). Keywords are associated with each micro segment.

MAGIC’s metadata repository is open to all users for collaborative work, much like social tagging systems. Users can search or browse media objects in the repository, examine or edit the metadata, and organize learning objects into collections for use in courses. Collections are groups of resources, usually pertaining to a specific topic. Users can create collections by adding or removing resources that are registered in the MAGIC system. After collections are created, they can be packaged for distribution. The SCORM packaging is a standard way of distributing learning content. Using SCORM, MAGIC supports three types of packaging: 1) learning content only; 2) metadata only, and 3) learning content and metadata. For packages containing content, the URL for each resource is accessed and the latest content is downloaded at the time the package is generated.

C. System Evaluation

MAGIC has been used on a trial basis by IBM customers and is being evaluated by University of North Carolina’s Department of Information and Library Science. Based on our initial experiences, we anticipate that automatically generated titles, summaries, and keywords will not rise to the level of human performance.

Fig. 4. Viewing a learning object.

Fig. 5. Viewing DHS taxonomy.

Fig. 6. Browsing results of video segmentation.
for sense-making tasks. However, the metadata tags, while different, appear to be both understandable and consistent. We anticipate that the cost of manually correcting the machine-generated metadata will be significantly less than inputting metadata completely by hand. Many IBM customers have legacy databases with a large amount of resources that would otherwise be completely inaccessible without a system like MAGIC to describe and organize the material.

The MAGIC categorization is nearly as good as human performance, as long as the MAGIC taxonomy is used. The system removed the need for humans to browse and consistently identify categories in a taxonomy of several thousand nodes, which is extremely time consuming and can result in inaccurate assignments. However, in some cases, companies required a different taxonomy than was supplied by MAGIC but were not willing to invest the time or cost to extend and organize the MAGIC taxonomy, since it requires 80 or more carefully selected training documents per new category.

Our initial experiences indicate that the greatest value from video segmentation comes from being able to locate contextually relevant portions within long videos. For example, there may be a useful portion about a Japanese leader in a war film that could be repurposed for a history course. This segment might not be identified by the segmentation algorithm because the video was primarily about different styles of war, but since the leader’s name is a key phrase, MAGIC users can use it to navigate to relevant portions of the video.

V. METHODOLOGIES FOR AUTOMATICALLY GENERATING USER PROFILES FOR BETTER TUBING AND FINDING

Personalization of any sort requires knowledge of the users’ preferences. In order to personalize the user’s experience with the videos that are retrieved in a search, a user profile is needed. Eliciting the user profile in an easy and unobtrusive manner is a big challenge as users do not want to spend time giving their preferences. Further, the content level preferences are something that the users may not even be able to articulate. The extraction of the user profile should be based on a structured methodology that is proven to capture the essence of the person’s likes and dislikes. The users’ preferences are expressed in a personal profile that specifies various multimedia content features that a particular user wants to see in a summary. A user profile has to meet two requirements. First, the system needs to generate a profile with minimal interruption—users want to spend their time watching TV, not data input. Second, the match between the profile and multimedia features needs to produce summaries that are informative and useful.

In order to meet these challenges, Agnihotri et al. [8] tested the hypothesis that the personality traits determine the summaries users prefer. The underlying assumption was that there exists a mapping of the personality traits to multimedia content features and content features influence preferences for summarization. This method requires input only once from the user. If a good mapping between personality and multimedia features can be found by surveying a number of people, meaningful summaries can be generated. This mapping gives varying importance to video segments based on the user profile and the multimedia content features in the video segment. A methodology to extract and apply this underlying user-centric mapping for automatically generating personalized summaries is explained.

A methodology that can be followed for generating personalized and validating these summaries is presented in Agnihotri et al. [9]. The methodology (Fig. 7) for multimedia personalization starts with a user study to extract the user-centric mapping between personality traits and multimedia content features. Multimedia content has inherent features: face presence, text presence, anchor segment, keywords, etc., as presented in Agnihotri [6]. Users also have inherent properties, which are governed by their personality types. There are many possible ways to extract personality traits by using different tests. In order to minimize the dependence of our results on a specific personality test, three approaches were employed: Myers
Briggs Type Indicator [36], Maximizing Interpersonal Relationships [51], and Brain.exe [48]. The goal of the user study was to explore and to establish a user-centric mapping between these two spaces. For the mapping user study, users were shown a series of segments of news, talk shows, and music videos. The user’s personality traits were obtained by giving personality tests to the users. For multimedia preferences, for each of the segments, users chose their preferred audio, video, and text summaries. A factor analysis was performed to extract the predominant directions in which different personality traits and multimedia content features vary together.

Once the mapping is established, it is used in the personalization algorithm to generate personalized summaries. The users take a simple personality test and using the user centric mapping the algorithm generates a personalized multimedia summary [9] that contain segments that are appealing to users. For example, extroverts get segments with faces, while introverts get segments where the host is elaborating on a topic. Before a searched video is downloaded the user can request to see short summaries based on their own personal profile generated based on their personal traits. This allows you now to tube and find easily the content that you are interested in.

VI. EMERGING CHALLENGES AND OPEN PROBLEMS

Some of the emerging challenges and open problems in personalizing multimedia content access in the YouTube era are presented here.

A. Nonsymmetric Actor Detection

One of the issues in presence or co-presence detection is the detection of other actors based on multiple modalities. This problem is far from solved. For example, even with audio, although speaker identification problems have been solved in controlled settings, the problem of speaker identification in uncontrolled settings such as cafes, where the environmental conditions are difficult and the microphone is not close to the speaker, is still an open problem. Much work is also required to obtain position and location information from impoverished, but multimodal data, for example sparse GPS, Bluetooth, and WIFI ids.

B. Capturing Personality Traits and Sensitivity Issues

To have an accurate mapping of personality traits to multimedia features, we need to conduct user testing across a huge number of users, across all types of contents, and across all possible multimedia features that can be computed. Further, this mapping only gives an idea of the person, the context of the user still needs to be explored as it is based on the task (consuming content, reviewing content, etc.), so the summaries provided need to be different. Also the issue of multi-user scenario is difficult as it is hard to identify the person who is sitting in front of the TV. Will we need people to “log-in” before starting to watch TV? Will we adjust the profiles to accommodate for multi-user cases (give more and more generic summaries)?

C. Collective Tagging, New Models of Metadata, and Implications for Enhancing Metadata Quality and Personalization

The use of mobile devices that are equipped with a plethora of sensing modalities such as Bluetooth, WIFI, and GPS makes the extraction of metadata possible in new and interesting ways. If one were to couple this with new media, blogs, and other forms of online media, it is possible to extract signatures about people at a much higher semantic level: What are they currently interested in? With whom are they discussing this? Does this group change over time?

Multimedia research, standing as it does at the cross roads of many disciplines relevant to this technology—signal processing and encoding, computer vision, and HCI—is well placed to take on the challenge. Much of the media these devices produce, and the new media genres that have evolved, are socially motivated. They are used for self-expression, memory preservation, relationship creation, etc. Not surprisingly, the result of the confluence of all these technologies in the presence of Internet, looks like an ecosystem to be developed! Multimedia researchers need to create or adapt existing frameworks that are cognizant of the sociological dimension to their quiver. Social Geometry [16] is an example of a candidate framework. It enumerates social dimensions that might offer a starting point to a specifically multimedia-aware framework, capable of cataloguing and arranging important aspects from the confusion, and how they might be extracted from sensors and media.

Acknowledgment

The authors would like to thank S. Gates, A. Katriel, G. Kofman, Y. Li, Y. Park, Y. Ravin, and W. Teiken at IBM Research for their collaboration and active participation in the MAGIC project.

REFERENCES

Personalizing Multimedia Content Access

Venkatesh et al.: “You Tube and I Find”—Personalizing Multimedia Content Access


ABOUT THE AUTHORS

Svetha Venkatesh (Senior Member, IEEE) is the Director of the institute for Multi-Sensor Processing and Content Analysis, Curtin University of Technology, Perth, Australia. She has extensive experience in low-level vision, pattern recognition, and multimedia content analysis and has made substantial contributions to computer science over the past decade, most recently in her development of the field of computational media aesthetics and media creation.

Dr. Venkatesh was the recipient of the John Curtin Distinguished Professorship and was elected a Fellow of the Australian Academy of Technological Sciences and Engineering and a Fellow of the International Association of Pattern Recognition.

Brett Adams received the B.S. (first class honours) degree in engineering from the University of Western Australia, in 1996, and the Ph.D. degree in computer science from Curtin University of Technology, Perth, Australia, in 2003.

He currently holds a Curtin Research Fellowship in the Department of Computing, Curtin University. His research interests include computational media aesthetics-enabled media creation, management and interaction, and the representation and discovery of social context from both online and ubiquitously sensed data and their interplay.

Dinh Phung received the B.S. (first class honours) and Ph.D. degrees in computer science from Curtin University of Technology, Perth, Australia, in 2001 and 2005, respectively.

He is currently a Research Fellow in the Department of Computing, Curtin University. His research interests include computational media aesthetics-enabled media creation, management and interaction, and the representation and discovery of social context from both online and ubiquitously sensed data and their interplay.

Chitra Dorai (Senior Member, IEEE) received the B. Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, the M.S. degree in electrical engineering from the Indian Institute of Science, Bangalore, and the Ph.D. from the Department of Computer Science, Michigan State University.

She is a Senior Research Manager at the IBM T.J. Watson Research Center, Hawthorne, New York. Her research interests include the areas of distributed stream processing systems, e-learning media management, mobile imaging, multimedia content analysis, computer vision, pattern recognition, and machine learning. She has published over 85 technical papers in premier IEEE and ACM conferences and journals, and has been granted multiple patents. She was the Editor of the book Computational Media Aesthetics (2002) and has contributed chapters to various books and edited collections in multimedia.

Dr. Dorai served as an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA and as the Associate Editor of the Pattern Recognition Letters journal. From Michigan State University, she received the Distinguished Academic Achievement Award. She is a member of the ACM.

Robert G. Farrell received the B.S. degree in mathematics from Carnegie-Mellon University, Pittsburgh, PA, and M.S. and M.Phil. degrees in computer science from Yale University, New Haven, CT.

He is a Research Staff Member in the Social Computing Group, IBM Research, Hawthorne, NY. His research focuses on user interfaces to support information organization, collaborative action, and contextual learning. He was co-PI of the MAGIC project, a government-funded initiative to develop tools to generate metadata for digital media. Prior to IBM, he worked as a member of Technical Staff at Bell Communications Research, Morristown, NJ. He is the author of two books and over 40 articles in the fields of cognitive science, human-computer interaction, information science, and artificial intelligence.

Lalitha Agnihotri received the B.S. degree in engineering from New Delhi University, India, in 1996, the M.S. degree from Pennsylvania State University, College Park, in 1998, and the Ph.D. degree from Columbia University, New York, in 2005, both in computer science.

She is a Senior Member of Research Staff at Philips Research, Briarcliff Manor, NY. She joined Philips Research in 1998. Her research activities over the years have covered wide range of information management in important areas such as multimedia content analysis and personalization as well as medical decision support systems. She has over 50 publications in peer-reviewed conferences and journals and 16 patents issued.

Dr. Agnihotri has provided her professional service to the program committees of several conferences including ACM Multimedia and IEEE Conference on Multimedia and Expo. She has reviewed many articles for leading IEEE journals.

Nevenka Dimitrova received the B.S. degree in mathematics and computer science from the University of Kiril and Metodij, Skopje, Macedonia, in 1984, the M.S. degree in computer science from Arizona State University, in 1991, and the Ph.D. degree in 1995.

She is a Research Fellow in the Department of Reliable Care Solutions, Philips Research Asia, Bangalore, and Domain Owner of the Bioinformatics Program at Philips Research. She has been with Philips Research USA since 1995 and with Philips Research India since July 2006. At the same time, she is a Visiting Scientist at Cold Spring Harbor Laboratory and at Columbia University. Her research interests include diverse areas such as video signal processing, bioinformatics, epigenomics, and DNA computing, but they all have a common theme: data and pattern mining. Her current research activities are in bioinformatics and biomarker discovery and how to enable decision support systems for personalized medicine. She has over 40 issued patents and over 120 publications.

Dr. Dimitrova was recently awarded the Philips Silver Medal for Innovation. She has given keynote presentations at CIVR, Med@net, IEEE ITCC. She actively participates in IEEE, ACM, and SPIE conferences, has chaired and served on 30+ different program committees, and is currently serving on IEEE Genomic Signal Processing Conference Program Committee as well as three editorial boards: ACM MULTIMEDIA SYSTEMS JOURNAL, IEEE MULTIMEDIA and ACM TRANSACTIONS ON INFORMATION SYSTEMS. She was a Special Sessions Co-chair for ICME 2004 and General Chair of ACM MM 2004, in New York.