♦ Multimedia Interactive Services Automation Based on Content Indexing

Julien Royer, Hang Nguyen, Olivier Martinot, and Françoise Prêteux

On fixed or mobile television (TV), interactive services embedded in multimedia are key technologies as they forge links among the media, telecom, and services worlds. In a context where content evolves rapidly, the challenge is to be able to propose value-added services on live programs in real time. In this paper, we propose a solution that automates the generation of interactive TV applications, providing TV viewers with additional content in context with the original TV program. The proposed architecture is based on plug-in multimedia analyzers which generate a contextual description of the media and on an interactive scene generator to dynamically create related interactive scenes. Principles, architecture, implementation, and experiments are described; they are based on an application case related to interactive services added to a session of a French parliament TV program like "Les travaux de l'Assemblée Nationale" on the French television channel La Chaîne Parlementaire (LCP). © 2008 Alcatel-Lucent.

Introduction

Providing multimedia content to end users is a key market driver for service providers [1, 2, 15, 16], whatever way it is delivered, via television (TV), radio, or video-on-demand. Personal content is becoming a part of that multimedia world, as evidenced by products like Alcatel-Lucent's 5900 My Own TV Application and services like YouTube*. In this context, the added value for service providers relies not only on simple delivery of multimedia content, but also on additional services and content that can be proposed to the end user, e.g., short message service (SMS), voting, or enhanced content. One way to propose these additional services is to embed them within the video content itself, using technologies commonly called interactivity. At present, the standard approach is to generate manually interactive content which is not always linked to the multimedia content. This is due to a lack of real-time multimedia analyzers to provide an exhaustive content description of complex scenes including objects, places, and people in order to build dynamically the semantics of the scene (i.e., interactions of those people, places, and objects). Instead, each type of media, e.g., video or audio, is analyzed individually, with heterogeneous performances [1, 6, 20].

This paper proposes a new architecture to insert interactive content into the media automatically. The basic idea is to "split" a complex multimedia analyzer into several simple analyzers. **Figure 1** illustrates the global architecture overview of the proposed system. The system is designed as a modular plug-in platform consisting of multimedia analyzers to generate the

Bell Labs Technical Journal 13(2), 147–154 (2008) © 2008 Alcatel-Lucent. Published by Wiley Periodicals, Inc.
Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/bltj.20310

multimedia description, and an interactive scene generator to create related interactive scenes and new multimedia services dynamically.

The interactive service templates are supplied by service providers in order to be embedded into multimedia content at appropriate moments depending on the semantic description of the media. The media analyzer (plug-in modular architecture) generates the media description. Then, the service selector selects and implements service(s) which best fit the current context of the scene from the available services database.

This paper first describes the general architecture of the proposed system. Next, the different parts of the architecture are detailed. Finally, we introduce a new interactive mobile TV application to validate the proposed architecture.

General Architecture

The architecture of the proposed system is based on plug-in multimedia analyzers and on an MPEG-7 scene context database from Motion Picture Experts Group (MPEG) standards [4]. The database is filled by plug-in multimedia analyzer results and then used to select and deploy interactive scenes corresponding to generated media descriptions.

Panel 1. Abbreviations, Acronyms, and Terms

BIFS—Binary format for scene
MPEG—Motion Picture Experts Group
SMS—Short message service
TV—Television
XML—Extensible Markup Language

Multimedia Analyzers Plug-In Architecture

Multimedia analyzer plug-ins generate descriptions of current multimedia context in real time. Using multimedia analyzer plug-ins as “filters” makes it possible to analyze each media type separately and to reduce a complex mechanism like a face recognition analyzer [22] or speech-to-text analyzer into “simpler” algorithms. Plug-ins could also be added or removed from the multimedia analyzer depending on the required semantic description. In this context, as the system gets more “simple” analyzers, “complex” analyzers will have less unknown parameters to consider and will then be able to work in known context conditions. For example, by leveraging known results from the other analyzers, a face recognition analyzer could receive as a parameter the position of the face to identify, and additional details such as whether the

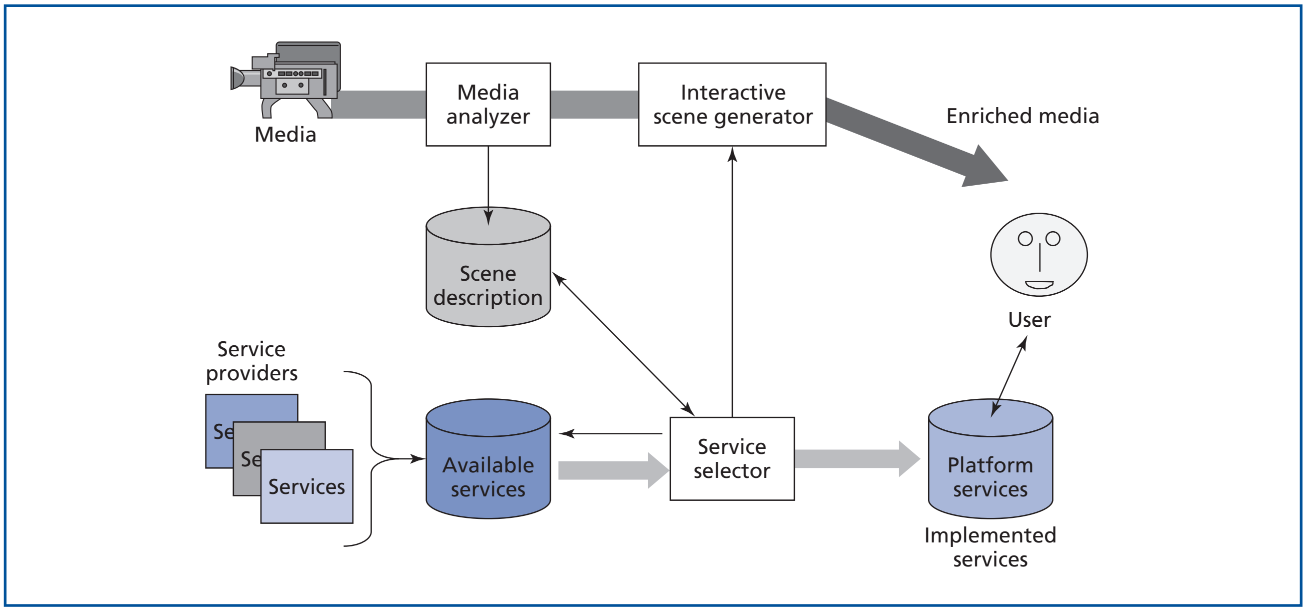

Figure 1.
System architecture overview.

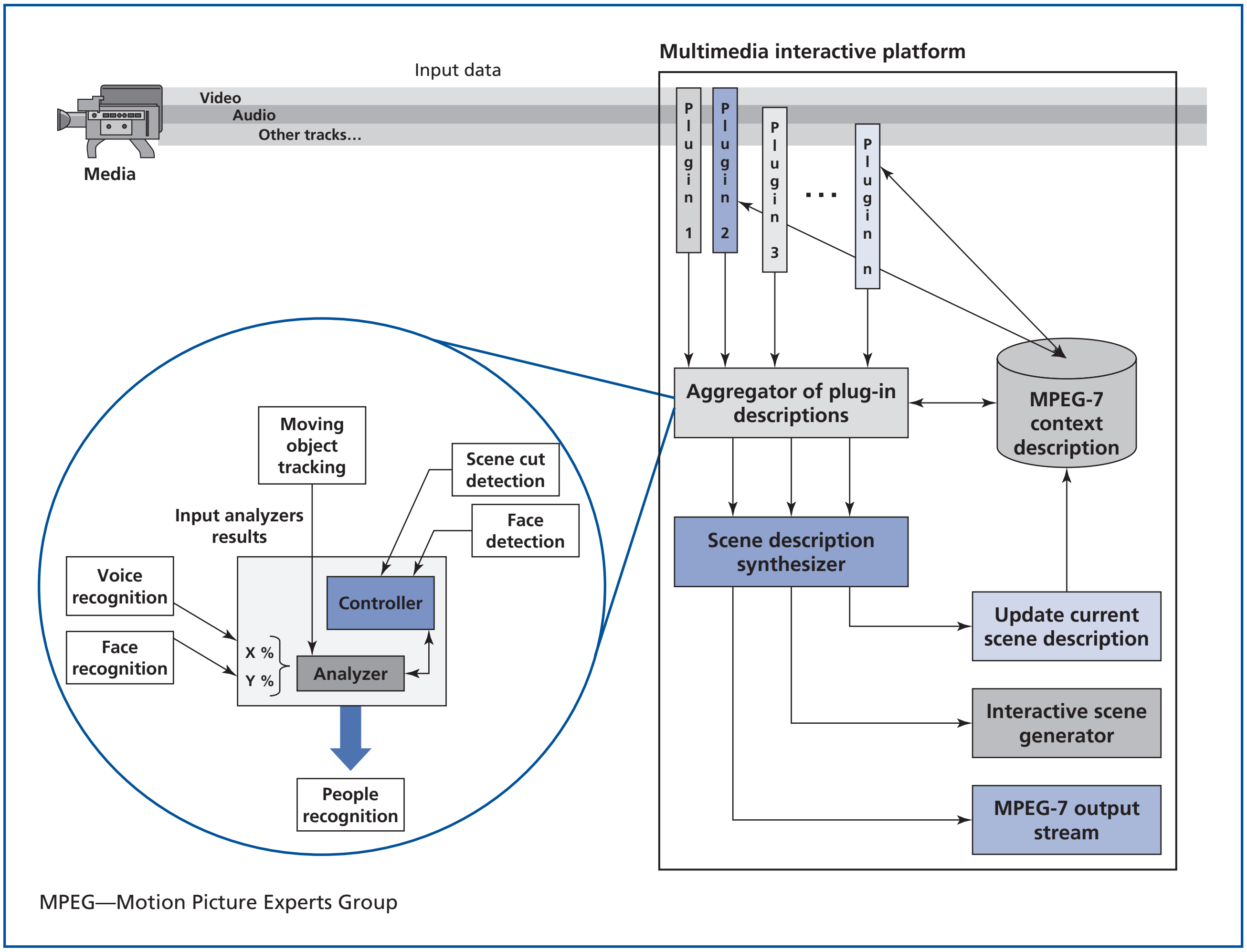

Figure 2.
Details on semantic analysis architecture.

person is facing the camera or is masked. This proposed architecture improves the performance of the face recognition analyzer. The system is also extensible, as we can select which plug-in media analyzers are to be inserted in our system. It is upgradable by adapting the number and the versions of the different plug-ins. Moreover, in our system, the multimedia stream can be analyzed in either asynchronous or synchronous mode, allowing the results to be serialized or parallelized to increase the overall performance. The media description is implemented according to the MPEG-7 standard [3, 12, 18], which was chosen because it allows a complete description of multimedia contents by presenting a standardized set of descriptors.

Combining Plug-In Results

The aggregator of plug-in descriptions, illustrated in **Figure 2**, combines the plug-in results. These results can also be combined with the existing descriptions of the media, if any. The MPEG-7 description generated by individual plug-ins can be exchanged as “mutual information” to improve the analysis performance of every plug-in analyzer. Two kinds of mutual information are considered: instantiated descriptors and events. Events are used to initialize, reset, and synchronize the different media analyzers. As an example, a scene cut detector will send an alert when a scene cut is detected. This alert will reset the moving object analyzer and initialize the module used to save the context. Instantiated descriptors provide details on the current scene description, e.g., location or people on-scene. The platform combines the descriptions generated from the plug-ins using combination rules. For example, in our system, the MPEG-7 plug-in results for moving object tracking and face detection are used to set up the face recognition plug-in, if the face was not previously recognized from earlier frames. Those combination rules can be settled by plug-in designers or manually added through a user interface. Moreover, weighting plug-in results serves to increase media description reliability, since face recognition results should be weighted as more reliable for scenes where people are facing the camera.

Figure 2 zooms out the aggregator of plug-in descriptions module as an example of combination rules implementation. This example implements combination rules for five simple media analyzers to identify people in the multimedia stream. First, the controller function starts the recognition analyzers as soon as the face detector plug-in detects a face. Then, the moving object tracking analyzer tracks the detected face. The recognition analyzer can then use that tracking information (voice and face) to improve its own performance. Last, as soon as a new scene cut event is set by the scene cut detector, the system is restarted.

It is possible to implement as many combination rules as needed for required descriptions and available plug-in analyzers. The selection and combination of media analyzers increase multimedia analysis performance since we can optimize the description with respect to the requirements. Finally, the output synthesizer combines the generated descriptions by automatically removing redundant descriptors and completes the general scene description and the current scene context using semantic analyzers.

MPEG-7 Current Scene Context

The MPEG-7 database illustrated in Figure 2 contains the current scene description, the "mutual information" described earlier. The plug-in analyzers and aggregator of plug-in descriptions are both linked to the MPEG-7 database to read the instantiated MPEG-7 description values if needed [14]. This allows the analyzer's algorithms to receive and leverage previously analyzed results in conjunction with its own media analysis. The platform manager, illustrated in **Figure 3**, sets the smallest incremental unit of the media (e.g., video frames per second, audio frequency) to synchronize the plug-in analyzers. The MPEG-7 database is then backed up with new data after each analysis.

Interactive Services

As detailed previously, multimedia content is analyzed and described semantically in order to identify pertinent interactive services to be inserted. Therefore, we can generate corresponding interactive services.

Generating Interactive Scenes and Services to Enrich Media Content

Interactive scenes are implemented by using Part 11 of the MPEG-4 standard [7, 10, 11, 17], called the binary format for scenes (BIFS). MPEG-4 BIFS is an object-oriented system which provides mechanisms and protocols to compose, animate, and describe interaction between media objects in terms of time and space in the scene. In the context of scene generation, interactive scenes can be modulated using scene templates, for example, a list of BIFS update commands or interactive scenarios. The interactive scene templates generated would then be implemented or withheld accordingly, based on input from the service selector. Therefore, service providers have to deliver interactive scene template(s) for the interactive services they would like to deploy on user terminals along with the required scene context (descriptors) to activate them. The service selector also evaluates the interactive template(s) selected from the database both spatially and temporally in order to ensure homogeneity with the current scene, e.g., to harmonize a scene or avoid a service that would mask another. Finally, the service selector can also "classify" proposed advertisements according to end user profiles. The end user terminal then will select which advertisements to display according to the user profile and user preferences.

The interactive scenes are built using an initial static scene composed of "inactive" objects such as a checkbox or media links. Once interactive scenes to be inserted on the media have been chosen, objects will be dynamically implemented as required by the

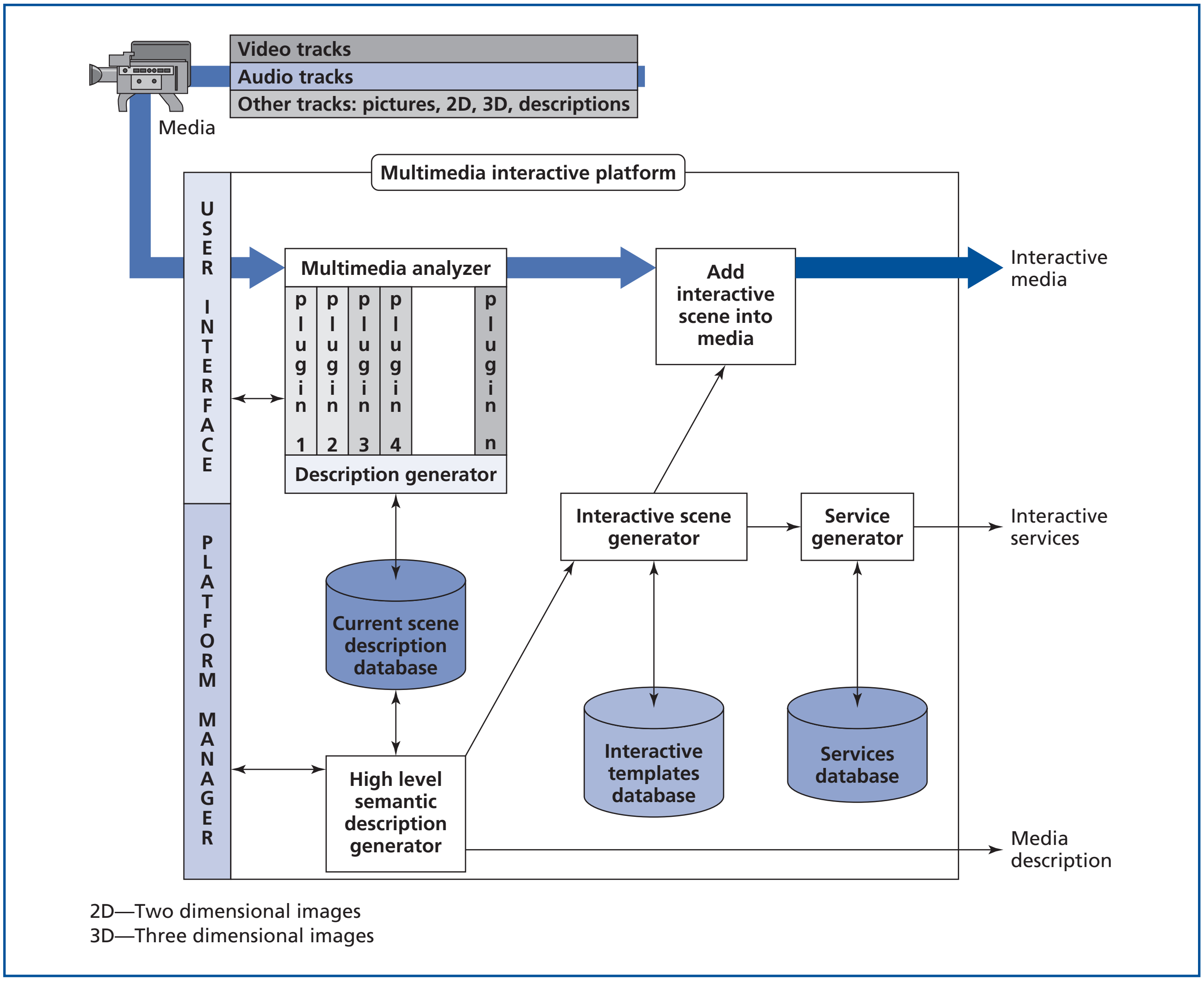
Video tracks
Audio tracks
Other tracks: pictures, 2D, 3D, descriptions
Media
Multimedia interactive platform
USER INTERFACE
PLATFORM MANAGER
Multimedia analyzer
plugin 1
plugin 2
plugin 3
plugin 4
plugin n
Description generator
Add interactive scene into media
Interactive media
Interactive scene generator
Service generator
Interactive services
Current scene description database
Interactive templates database
Services database
High level semantic description generator
Media description
2D—Two dimensional images
3D—Three dimensional images

Figure 3.
Plug-in multimedia analyzer.

interactive scene templates using BIFS-update commands. Finally, the interactive scene generator inserts generated interactive scenes into the multimedia content to create a rich media and deploys the corresponding service logic on the service platform if needed.

Interactive Parliament TV Session Application to Validate the Proposed Architecture

Figure 3 illustrates the different parts of the system that have been developed. In order to demonstrate the capabilities of the platform, we developed two kinds of interactive services for a TV broadcast session of the French Parliament on the LCP television channel. "Static" interactive services like quiz or chat were developed along with dynamic interactive services. The dynamic services served as the main points of interest to evaluate the system, since the interactive content is dynamically adapted to the media content description. As a first step, we designed a complementary information service about people present on the scene, as it is a contextual interactive service. The interactive service has to update proposed complementary information in real time, in order to fit with the current scene. As advantages, we consider that the environment is well known with respect to the camera's position and people's position. As disadvantages, the environment colors (wood furniture and decors) are similar to skin color components [19], and even if people are facing the camera, they are very rarely looking at it. Hence, we implemented a face detector [21] and utilized face recognition and face tracking analyzers to detect, recognize, and follow people that have been recognized. The analyzer's results are exchanged as mutual information using the MPEG-7 Extensible Markup Language (XML) database. Moreover, we developed a set of simple plug-in analyzers to generate descriptors at lower levels of abstraction. We analyzed video frames in the compressed domain [9] in order to generate temporal segmentation (i.e., scene cut detection based on histogram analysis [8, 13]) and spatial segmentation (people's location in the scene using background registration [5]). Finally we extracted people's faces in order to identify them.

Conclusion and Perspectives

As we developed multimedia analyzers based on very simple algorithms, the goal in this paper was not to present quantitative results regarding the performance of the system in terms of speed and accuracy. As it is possible to integrate dedicated media analyzers to enhance the system, our preliminary objectives were to verify the feasibility and to implement a generic architecture able to combine results from multimedia analyzers.

Preliminary results obtained in prototyping interactive TV Parliament sessions demonstrate the capabilities of this architecture. The plug-in architecture ensures a rapid systems adaptation to multimedia analysis requirements and significant interactive services diversification through use of the MPEG-7 description standard.

Further work should be done on semantic analyzers to generate semantic attributes from descriptors, and vice versa, as the system needs to estimate how close the current scene description is to each required service's context description. In order to increase system performance, MPEG-7 description schemes as well as descriptors should be created to reduce the set of descriptors available to current media content.

*Trademarks

YouTube is a trademark of Google, Inc.

References

[1] E. L. Andrade, E. Khan, J. C. Woods, and M. Ghanbari, "Player Classification in Interactive Sport Scenes Using Prior Information Region Space Analysis and Number Recognition," Proc. Internat. Conf. on Image Process. (ICIP '03) (Barcelona, Sp., 2003), vol. 3, pp 129–132.

[2] T. Bara, E. Papaioannou, and N. Ioannidis, "MELISA Multiplatform E-Publishing for Leisure and Interactive Sports Advertising," Proc. Melisa Workshop (Athens, Gr., 2005).

[3] J. J. Burred, A. Röbel, and X. Rodet, "An Accurate Timbre Model for Musical Instruments and Its Application to Classification," Proc. 1st Workshop on Learning the Semantics of Audio Signals (LSAS, '06) (Athens, Gr., 2006).

[4] L. Chiariglione, "MPEG," <http://www.chiariglione.org/leonardo/standards/mpeg/index.htm>.

[5] S.-Y. Chien, S.-Y Ma, and L.-G. Chen, "Efficient Moving Object Segmentation Algorithm Using Background Registration Technique," IEEE Trans. Circuits and Syst. for Video Technol., 12:7 (2002), 577–586.

[6] Cleveland State University, "The Content Analysis Guidebook Online: An Accompaniment to The Content Analysis Guidebook by Kimberly A. Neuendorf," 2007, <http://academic.csuohio.edu/kneuendorf/content/>.

[7] C. Concolato and J. Le Feuvre "MPEG-4 BIFS and XMT Tutorial," 2005, <http://gpac.sourceforge.net/tutorial/bifs_intro.htm>.

[8] A. M. Ferman, A. M. Tekalp, and R. Mehrotra, "Robust Color Histogram Descriptors for Video Segment Retrieval and Identification," IEEE Trans. Image Process., 11:5 (2002), 497–508.

[9] L. Gu, Video Analysis in MPEG Compressed Domain, Ph.D. Thesis, University of Western Australia, 2002.

[10] C. Herpel and A. Eleftheriadis, "MPEG-4 Systems: Elementary Stream Management," Signal Process.: Image Commun., 15:4–5 (2000), 299–320.

[11] International Organization for Standardization and International Electrotechnical Commission, "Coding of Audio-Visual Objects: Systems," Final Draft Internat. Standard, MPEG-4

Systems, ISO/IEC 14496-1, ISO/IEC JTC1/SC29/WG11 N2501, Oct. 1998.

[12] International Organization for Standardization and International Electrotechnical Commission, "MPEG-7 Overview (Version 10)," ISO/IEC JTC1/SC29/WG11 N6828, Oct. 2004.

[13] R. A. Joyce and B. Liu, "Temporal Segmentation of Video Using Frame and Histogram Space," IEEE Trans. Multimedia, 8:1 (2006), 130–140.

[14] A. J. Perrott , A. T. Lindsay, and A. P. Parkes, "Real-Time Multimedia Tagging and Content-Based Retrieval for CCTV Surveillance Systems," Proc. SPIE Conf. on Internet Multimedia Management Syst. III (Boston, MA, 2002), vol. 4862, pp. 40–49.

[15] L. A. Rowe, "The Future of Interactive Television," University of California Berkeley, Comput. Sci. Division, Electrical Engineering and Comput. Sci. (EECS), Aug. 30, 2000.

[16] J. Shin, D. Y. Suh, Y. Jeong, S. H. Park, B. Bae, and C. Ahn, "Demonstration of Bidirectional Services Using MPEG-4 BIFS in Terrestrial DMB Systems," ETRI J., 28:5 (2006), 583–592.

[17] J. Signes, Y. Fisher, and A. Eleftheriadis, "MPEG-4's Binary Format for Scene Description," Signal Process.: Image Commun., 15:4-5 (2000), 321–345.

[18] T. Sikora, "MPEG-7-Based Audio Annotation for the Archival of Digital Video," 2005, <http://www.nue.tu-berlin.de/forschung/projekte/mpeg7/>.

[19] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A Survey on Pixel-Based Skin Color Detection Techniques," Proc. GraphiCon (Moscow, Rus., 2003), pp. 85–92.

[20] D. Weinland, R. Ronfard, and E. Boyer, "Free Viewpoint Action Recognition Using Motion History Volumes," Comput. Vision and Image Understanding, 104:2–3 (2006), 249–257.

[21] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," IEEE Trans. Pattern Analysis and Machine Intelligence, 24:1 (2002), 34–58.

[22] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips, Face Recognition: A Literature Survey, University of Maryland, UMD Tech. Report CAR-TR-948, College Park, MD, Aug. 2002.

Additional Resources

— Akond, <http://www.akond.net>.

— R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," Proc. IEEE, 83:5 (1995), 705–740.

— Fedora Commons, <http://www.fedoracommons.org>.

— R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video Abstracting," Commun. ACM, 40:12 (1997), 54–62.

— M. Lux, "Caliph and Emir," <http://www.semanticmetadata.net/features/>.

— U. Park, H. Chen, and A. K. Jain, "3D Model-Assisted Face Recognition in Video," Proc. 2nd Canadian Conf. on Comput. and Robot Vision (CRV '05) (British Columbia, Can., 2005), pp. 322–329.

— Riya, "Riya Visual Search," <http://www.riya.com>.

— Yahoo!, "Flickr," <http://www.flickr.com>.

(Manuscript approved March 2008)

JULIEN ROYER is a Ph.D. student in the Conventions Industrielles de Formation par la REcherche (CIFRE) program working at Alcatel-Lucent as part of a collaboration with Institut TELECOM / TELECOM & Management SudParis, Advanced Research TEchniques for Multidimensional Imaging Systems (ARTEMIS). He is based in Villarceaux, France. His Ph.D. studies focus on the automatic generation of interactivity for multimedia services.

HANG NGUYEN is an associate professor at the Institut TELECOM/TELECOM & Management SudParis. She has worked on radio mobile network architecture (Universal Mobile Telephone Service [UMTS], General Packet Radio Service [GPRS], Enhanced General Packet Radio Service [EGPRS]), the radio mobile physical layer, and the MPEG standards family on scalable video codecs and transmission. She obtained her Ph.D. for work on video transmission over wireless networks and has written over 15 papers and registered 15 patents.

OLIVIER MARTINOT is the research manager of the HyperMedia Applications Domain at Alcatel-Lucent Bell Labs in Villarceaux, France. He is responsible for the Residential Networked Application Project (ReNA), and his team's objective is to address convergence—fixed/mobile, multimedia/telecom, or other. He has written several papers and registered over 20 patents.

FRANÇOISE PRÊTEUX is head of the ARTEMIS

(Advanced Research TEchniques for Multidimensional Imaging Systems) Project Unit at the Institut TELECOM/TELECOM & Management SudParis. She graduated from the Ecole des Mines de Paris and received her Ph.D. degree in mathematics from the University of Paris VI. Dr. Prêteux is a member of the Editorial Board of the Journal of Electronic Imaging and a co-chair of the SPIE Conference on Mathematical Modeling and Estimation Techniques in Computer Vision. She is the author or co-author of over 60 scientific papers within the field of stochastic modeling and mathematical morphology with applications to pattern recognition, medical imaging, and non-destructive testing. ◆