Who's That Actor? The InfoSip TV Agent

Nevenka Dimitrova[1] Angel Janevski[1] Dongge Li[2] John Zimmerman[3]

[1]Philips Research
345 Scarborough Rd.
Briarcliff Manor, NY 10510
nevenka.dimitrova@philips.com

[2]Motorola Labs
1301 East Algonquin Road
Schaumburg, Illinois 60196
dongge.li@motorola.com

[3]HCI I, Carnegie Mellon,
5000 Forbes Avenue
Pittsburgh, PA, USA
johnz@cs.cmu.edu

ABSTRACT

We present a content augmentation application that enhances the primary video watching experience by providing related supplemental information. Our goal is to explore the value of cross-referencing related content among different media such as TV and Web. The cross-referencing is based on metadata, which is automatically extracted from the content. Metadata extraction can make a great difference for personalized user experience. In addition, annotations that provide title, genre, description, and cast can be greatly enriched with detailed information at the subprogram level, i.e. at the clip and scene level. We developed a system called InfoSip that performs person identification and scene annotation based on actor presence. The system links this information with actors' filmographies and biographies and produces an enriched viewing experience.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Algorithms, Indexing methods, Video analysis*

General Terms

Algorithms, Measurement, Theory.

Keywords

Person identification, content enhancement, content augmentation, hypervideo, Interactive TV services. talking head analysis

1. INTRODUCTION

Ambient intelligence applications offer context sensitive, multimedia content and information at any time and any place with effortless interaction. Within this research direction, Interactive TV services provide enriched user experience by enhancing traditional broadcasts. The challenge is to provide technology that will select the related pieces of multimedia content and deliver it in an appropriate manner. Here we present an application that bridges two fundamentally different paradigms: TV and Web.

Conference '03, November 2-8, 2003, Berkeley, CA, USA.

2. RELATED WORK

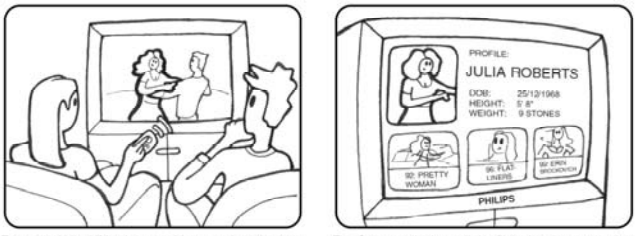
Context sensitive content delivery was explored in the web reconnaissance agent Letzia [6] and Project Aura [4]. Commercial applications include DVDs with supplemental data as well as ABC enhanced broadcasts.

Letzia is a computer agent that assists users by profiling the users behavior to learn preferences and then assessing links in the area of that user's current web activity [6]. Like InfoSip, Letzia knows the user's context by design, in this case web browsing. Unlike InfoSip, Letzia does not extract or summarize the information, but instead provides users with a prioritized list of associated links in the side of the browser window. Project Aura focuses on inferring a user's context using a variety of sensors and then tailoring information so as to eliminate unnecessary distraction. Unlike InfoSip, Aura does not already know the user's context, but similar to InfoSip, Aura is expected to extract and summarize relevant data [4].

Consumer DVDs that provide supplemental information are quite similar in purpose to InfoSip. However, InfoSip has some advantages such as up-to-date fresh information and personalization. ABC's Enhanced TV broadcasts, like InfoSip, combine web and TV content in a TV-centric viewing experience. These broadcasts synchronize webcasts with TV broadcasts. Users can get supplemental sports statistics while watching football games, participate in live interactive polls during talk shows, or even play along with game shows. Unlike InfoSip, these broadcasts divide user's attention on two screens: a PC (possibly laptop or webpad) and the TV screen. This division of attention is much more intrusive to the viewing experience than InfoSip.. Also, these enhanced broadcasts combine content from the same source (ABC) while InfoSip can collect user specific data from a variety of web source to supplement the narrative, video content. Finally, the webcasts themselves are a "production" and require considerable human effort to put it together while InfoSip is designed to run autonomously in the background.

3. FAQ FOR NARRATIVE VIDEO

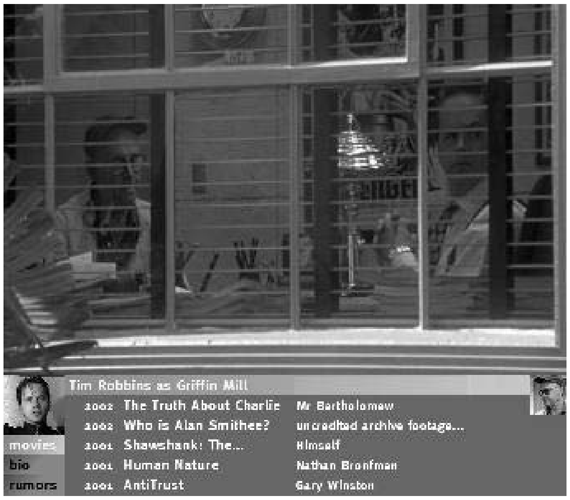
We conducted a focus group in order to evaluate various content augmentation concepts. Participants really enjoyed a concept involving easy access to supplemental information about actors for movies and TV shows (Figure 1). However, they did not want its use to interrupt viewing. In addition, they wanted to get more than just actor information. They wanted supplementary information such as the name of the song being played or more detailed information about locations. Our system allows users to select a specific query on the remote control. The users can potentially ask: Who's that actor? What's that song? Where are they? What kind of shoes are those? etc. In general terms, InfoSip uses predefined categories of questions/buttons such as "who", "where", "what", "when", "why", and "how much". Focus group participants did not want links to web sites. Instead, they wanted much more digested and summarized information that appears immediately as an overlay, allowing them to continue watching the movie/TV show (Figure 2).



Tommy and his girlfriend are watching a movie and see an actress on the screen that looks familiar but they can't remember who she is...

The System tells them all the films they have seen that this actress has been in.

Figure 1. Actor Info Concept, shown to focus group



Tim Robbins as Griffin Mill

2002 The Truth About Charlie Mr Bartholomew
2002 Who is Alan Smithee? uncredited archive footage...
2001 Shawshank: The... Himself
2001 Human Nature Nathan Bronfman
2001 AntiTrust Gary Winston

Figure 2. Sample screen from InfoSip (Who)

InfoSip is an example of a "frequently asked questions" answering application. It unobtrusively serves actor information related to the scene. Users press the "who" button to ask "who's that actor?" The system displays a list of all of the actors in the current scene using annotated data from person identification (see section 4) and supplemental data about each actor obtained through Web Information Extraction. The Information Extraction method automatically finds pertinent information in Web pages [5]. Filmography information can be personalized, based on the user's viewing history. Highlighting the movies in which users have seen this actor increases the chances that they will remember why this person looks familiar. The design of the menus on the overlays is also reconfigurable, based on a personal profile. For example, "bio", "filmography", and "rumors" are the menus available for person interested in gossip, but "bio", "filmography", and "references" are menus available for people more interested in references this movie is making to other movies.

4. AUGMENTING NARRATIVE CONTENT

A rich "frequently asked question"-answering application relies on manual annotation or automatic extraction of high-level information from video. In Figure 3 we show a high level diagram of the content augmentation process. The process is conceptually divided in two major stages: server and client. The server stage receives a program, which may or may not contain annotations and/or augmentation. In principle, we assume that the metadata needs to be either extended or updated. Based on the feature extraction and annotation, the server will execute one or more information extraction tasks (WebIE), which will retrieve current data from Web sites. For example, such data can be an actor's recent filmography and current biography. This information is integrated with any other program annotation and the program is augmented with metadata. Finally, the result is formatted and delivered to the client. In the client stage, the content is stored and can optionally be augmented with Internet content based on locality and/or user preferences.

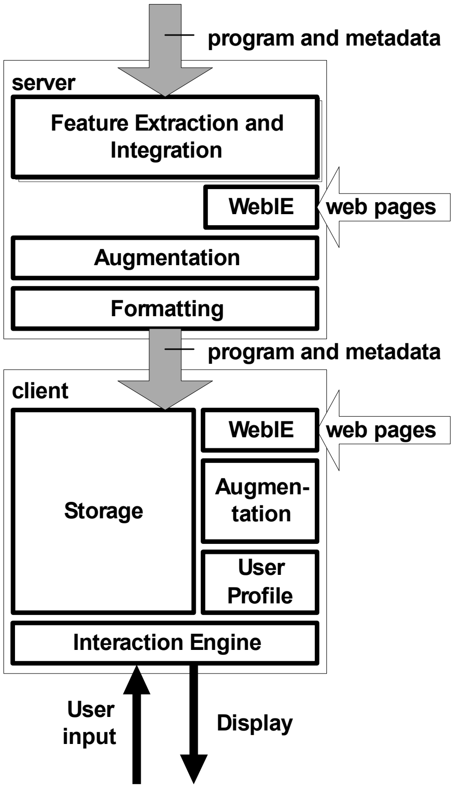


Figure 3: Content augmentation process

5. ACTOR RECOGNITION

Metadata extraction for FAQ type of applications requires either automatic or manual annotation. For example, to answer the "who is this person?" question in a movie, documentary, or home video we need to know who are the people/actors that are present in the scene for each individual scene. The major challenge is to robustly identify persons from different views, distance, lighting conditions, and in various background noise conditions. We used automatic face and voice identification methods for this task [2]. Also name spotting on textual information extracted from closed caption data provide an additional channel for our analysis. Figure 4 shows the structure of our InfoSip person identification system.

A person identification approach is constructed based on the joint use of visual, audio, and textual information. First, we perform visual analysis for detection, tracking, and recognition of faces in video. Face trajectories are first extracted and the Eigenface method is used to label each face trajectory as one of the known persons in database. Our face detection system can detect both front-view and side-view faces [8]. While the side-view face detection is important for generating face trajectories, only the front-view faces are used in the recognition process. Due to the limitation of existing face recognition techniques and the complex environmental factors in our experimental data, the visual recognition accuracy is not high.

Next, we employ audio segmentation and classification to find the speech segments. Film often has music background or environmental noise in the soundtrack and these factors make the audio identification a challenging process. We developed an audio classification system that can classify audio segments into seven categories including speech, music, noise, silence, speech with noise, speech with speech, and speech with music [1]. Speaker identification using Gaussian Mixture Models is then applied on those segments with speech components.

Both audio and visual analyses have their advantages under different circumstances, and we studied how to exploit the interaction between them for improved performance. In the fusion phase we employ two strategies. In the first, *audio-verify-visual fusion* strategy, speaker identification verifies the face recognition result. In the second, *visual-aid-audio fusion* strategy, we use face recognition and tracking to supplement speaker identification results. The first strategy has a slightly lower recall than the face recognition and best precision which is good for surveillance type of applications. The second strategy generates the best overall identification performance and is suitable to TV content analysis.

For the textual information extracted from closed caption or video caption, we have a name spotting process that extracts role names that appear in each video scene and assign a score for each detected role name according to the frequency of its appearance as well as those that closely relate to it. These scores together with our audiovisual detection results are used in a final voting process to decide which role(s) appear in the scene. The integration is based upon the belief values of different candidates using a single layer Bayesian network. The ones with highest integration belief will then be justified as top characters appearing in the scene.

Some earlier work integrates visual and audio analysis results while disregarding the face and speech correspondence. While these systems work well for applications where there is only one talking face on the screen each time, such assumption does not hold for narrative content such as movies. We resolve the problem through the use of a talking head detection process, which automatically detects the face(s) on the screen that has corresponding speech in the synchronized soundtrack. Such information can then be used in the fusion process to integrate the speaker identification results with the corresponding face trajectory. A cross-modal association method called Cross-modal Factor Analysis (CFA) is proposed and used for our talking head detection [7][3]. CFA achieves 91.1% detection precision in our experiments, while our two other implementations based on Latent Semantic Indexing (LSI) and Canonical Correlation Analysis (CCA) achieve 66.1% and 73.9% detection precision respectively using the same set of testing data.

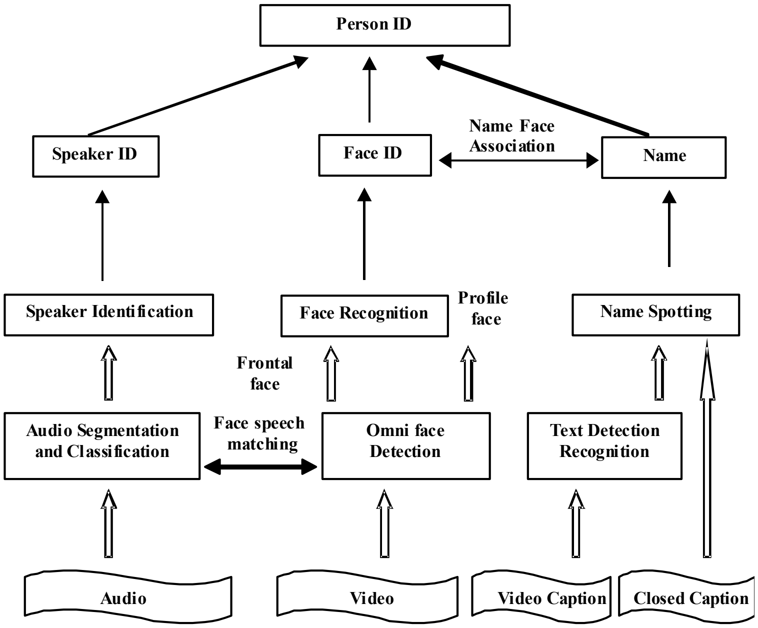


Figure 4. Architecture of InfoSip person identification system

6. SUMMARY

We presented a content augmentation application that relies on automatic meta-data extraction and provides a new interaction paradigm for consumers. The content analysis and meta-data linking work to provide seamless information delivery based on the TV program context. The application uses automatic person identification for providing links and pulling actor biography and filmography information from Web sources. The initial feedback from the users indicates that this is a desirable application in the context of introducing new interactive TV experiences.

7. ACKNOWLEDGMENTS

We would like to thank Mingkun Li, Gang Wei, Ishwar Sethi for various discussions and aspects of person identification project.

8. REFERENCES

[1] Li, D., Sethi, I.K., Dimitrova, N., and McGee, T., “Classification of general audio data for content-based retrieval,” Pattern Recognition Letters, 22, (2001) 533-544

[2] Li, D., Wei, G., Sethi, I.K., Dimitrova, N., “Person Identification in TV programs,” Journal of Electronic Imaging, Vol. 10, Issue. 4, pp. 930-938, October 2001.

[3] Li, D., Dimitrova N., Li, M., and Sethi, I.K., “Multimedia content processing through cross-modality association,” ACM Multimedia, November 2-5, Berkeley

[4] Garlan, D., Siewiorek, D., Smailagic, A., Steenkiste, P. Project Aura, “Toward Distraction-Free Pervasive Computing,” IEEE Pervasive Computing 1(2). Spring 2002. 22-31

[5] Janevski, A. and Dimitrova, N. “Web Information Extraction for Content Augmentation,” in Proc. IEEE Int. Conference on Multimedia and Expo, Switzerland, Aug., (2002), pp 389-392.

[6] Lieberman, H., Fry, C., Weitzman, L. “Exploring the Web with Reconnaissance Agents,” Communications of the ACM 44(8). August 2001. 69-75.

[7] Li, M., Li, D., Dimitrova N., and Sethi, I.K., “Audio-visual talking face detection,” Proc. ICME 2003, Baltimore, MD, July 2003.

[8] Wei, G., Sethi, I.K., "Omni-Face Detection for Video/Image Content Description," International Workshop on Multimedia Information Retrieval, in conjunction with ACM Multimedia Conference 2000, (MIR2000), Los Angeles, November 2000.