Beyond the Informedia Digital Video Library: Video and Audio Analysis for Remembering Conversations

Alexander G. Hauptmann and Wei-Hao Lin

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

The Informedia digital video library project pioneered the automatic analysis of television broadcast news and its retrieval on demand. Building on that system, we have developed a wearable, personalized Informedia system, which listens to and transcribes the wearer's part of a conversation, recognizes the face of the current dialog partner and remembers his/her voice. The next time the system sees the same person's face and hears the same voice, it can retrieve the audio from the last conversation, replaying in compressed form the names and major issues that were mentioned. All of this happens unobtrusively, somewhat like an intelligent assistant who whispers to you: "That's Bob Jones from Tech Solutions, two weeks ago in London you discussed solar panels." This paper outlines the general system components as well as interface considerations. Initial implementations showed that both face recognition methods, and speaker identification technology have serious shortfalls that must be overcome.

1. INTRODUCTION

The foundation for this work, the Informedia Digital Video Library (DVL) Project [23], has pioneered the successful application of speech, image, and natural language processing in automatically creating a rich, indexed, searchable multimedia information resource. The Informedia Digital Video Library [24] was the only NSF DLI project focusing specifically on information extraction from video and audio content. Over a terabyte of online television video data was collected, with automatically generated metadata and indices for retrieving videos from this library. Speech recognition is used to create text transcripts, and image analysis finds shot breaks, faces, and overlaid text for OCR processing. The video is segmented based on video and audio properties, and natural language analysis creates topics, titles, named entities mentioned in the video. The basic Informedia system allows a user to ask a query and receive a set of relevant video clips as results. The user may then browse through automatically generated abstractions of these video clips in the form of summary key frames for the complete clip, text titles, filmstrips with a representative frame for each shot, topics, and spatial maps of geographic entities occurring in the transcript. During the querying and browsing phase, the user may at any time select and play any clip from the result set.

We build on these technologies, moving beyond a digital video library for broadcast video into new information spaces of personal memory and real-time collaboration, with unedited, continuously captured video and audio from devices worn by individuals.

2. A VISION FOR PERSONAL INFORMEDIA

Suppose digital storage was infinite and *everything* that you did, saw and heard could be saved for posterity. In an era of ever-increasing data collection and archiving, the existence of yet more information stored somewhere is of no use to you *unless* you could access it easily and efficiently.

Our research is developing tools, techniques, and systems allowing people to capture and retrieve from a complete audio, video and textual record of their personal experiences and electronic communications.

This vision assumes that within ten years technology will be in place for creating a continuously recorded, digital, high fidelity record of one's whole life in video form [14]. Personal Informedia units will record audio, video, GPS and electronic communications. This data will eventually be incorporated into a structured multimedia resource as in containing multiple synchronized streams of information. This research fulfills the vision of Vannevar Bush's personal Memex [13], capturing and remembering whatever is seen and heard, and quickly returning any item on request, but recognizing that automated search, presentation, and summarization are technologies key to its realization.

The capture and abstraction of personal experiences recorded through audio, video, GPS and electronic communication can serve as a form of personal memory. A personal Informedia system becomes the substrate for an intelligent assistant that can provide memory refreshers as needed. In the following sections we describe a first implementation of a personal Informedia system for remembering people and conversations.

3. PROTOTYPE SYSTEM DESCRIPTION

While our vision outlines a research program expected to last for a number of years, we have reduced certain aspects of this

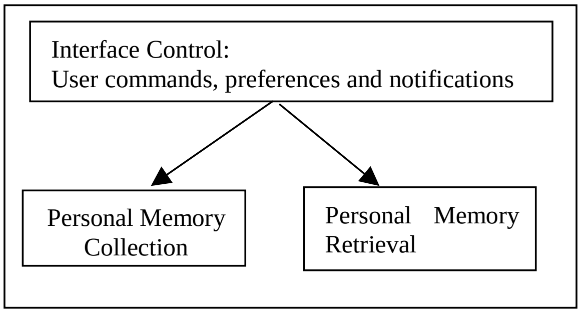
Interface Control:
User commands, preferences and notifications
Personal Memory Collection
Personal Memory Retrieval

Figure 1. The basic architecture of the Personal Memory system.

vision into an operational personal memory prototype that remembers the faces and voices associated with a conversation and can retrieve snippets of that conversation when confronted with the same face and voice. The system currently only combines face detection/recognition with speaker identification, and audio recording and analysis. The face detection and speaker identification enables the storing and retrieval of the audio conversation associated with a face and a voice. The face recognition and speaker identification allows automatic retrieval of the audio associated with the face or voice. Audio analysis and speech recognition is used to compact the conversation, retrieving only important phrases. All of this happens unobtrusively, somewhat like an intelligent assistant who whispers relevant personal background information to you before a meeting.

There are currently two modes of system operation: data collection (learning) and memory retrieval.

The basic hardware components of the system are a wearable miniature digital video camera, 2 microphones (1 close-talking and 1 omni-directional), headphones to receive system output and a laptop computer for processing. The software modules involved in the system are: A module for speaker identification, a speech recognition module, a face detection and recognition modules, a database, and an interface control manager module. The basic architecture of the system, as outlined in Figure 1, shows the interface control manager selecting between the acquisition and the retrieval module as needed.

3.1. Interface Control

The interface control module determines which of 3 states the system is currently in: Idle, Collecting Memory or Retrieving Conversations from Personal Memory. The current user input interface to this control module consists of a wearable mouse, with one button to start collecting conversation memory, one button to start retrieving the conversations (or to skip to the next relevant one) and one button to set the system into an idle state where neither collection nor retrieval takes place.

On the output side, it is assumed that all output will be provided through the headphones to the user wearing the unit. A visual display interface is available only for test and demonstration purposes. The interface control module has the ability to provide a set of short audio notifications to inform the user of what the system is currently doing. These notifications include, among others:

- Memory collection is starting
- Memory collection is ending
- Memory retrieval is starting
- Memory retrieval is ending
- Face was detected in video stream
- Failure to detect faces
- Various internal system failure states

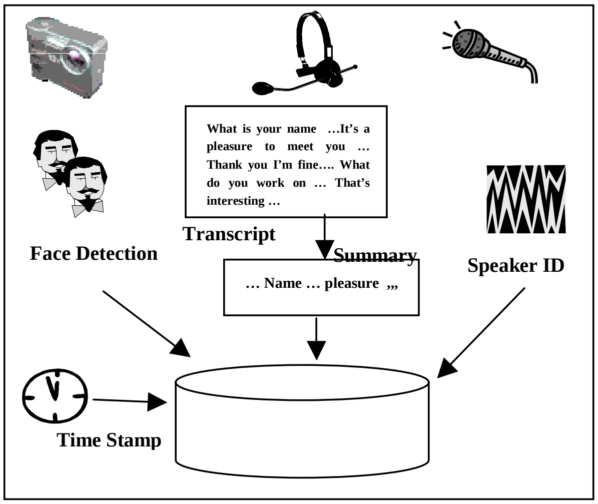
What is your name …It's a pleasure to meet you … Thank you I'm fine…. What do you work on … That's interesting …
Transcript
Face Detection
Summary
… Name … pleasure ,,,
Speaker ID
Time Stamp

Figure 2. Process for Personal Memory Collection

Our experience has shown that these user notifications were extremely important to keep the user informed of what state the system is in, since there is no visual feedback in normal use.

Future versions of the system will likely have the interface manager module controlled through a limited vocabulary, command speech interface (e.g. "Record this to memory" or "Who is this person?"). We envision that eventually the interface system will be triggered entirely by context-dependent recognized dialogue phrases, for example, ``Nice to meet you'', "My name is", etc., instead of mouse clicks or specific commands by the user.

3.2. Personal Memory Collection

The system hardware is designed as a wearable device consisting of a miniature ‘spy’ camera, a cardioid lapel microphone and an omni-directional microphone all attached to a laptop computer. The system works by detecting the face of the person you are talking to in the video stream, and listening to the conversation from both the close-talking (wearer) audio tracks and the omni-directional (dialogue partner) audio track. An overview of the ‘learning’ system for memory collection is shown in Figure 2.

The close-talking audio is transcribed by a speech recognition system to produce a rough, approximate transcript. The omni-directional audio stream is processed through a speaker identification module. An encoded representation of the face of your current dialog partner, the dialog partner speaker characteristics, and the raw audio of the current conversation is saved to a database. The next time the system sees the same person (by detecting a face and matching it to the stored faces in the database), it can retrieve and play back the audio from the last conversation.

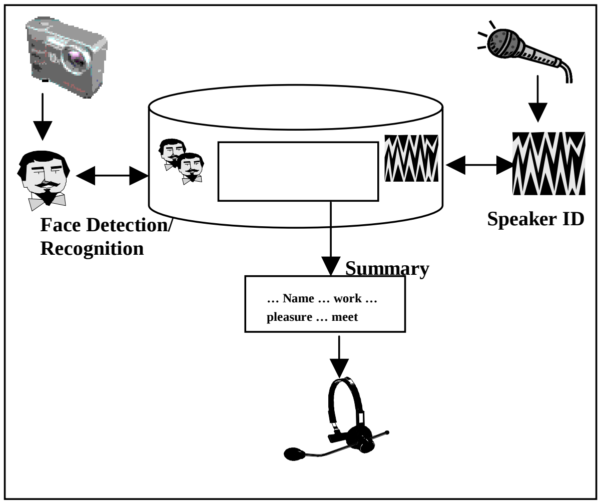

Figure 3. Process for Personal Memory Retrieval

The audio can optionally be processed through audio analysis (silence removal, emphasis detection) and general speech recognition to efficiently replay only the person names and the major issues that were mentioned in that conversation.

3.3. Personal Memory Retrieval

In the retrieval (remembering) mode, the system searches for a face in the video stream and performs speaker identification on the omni-directional audio stream. Once a face is detected, the face and speaker characteristics will be matched to instances of faces and speaker characteristics stored in the memory database. A simple linear interpolation is used to combine the score of both faced and speaker matches. When a sufficiently high scoring match is found, the system will return a brief summary of the corresponding recorded conversation with the person. Figure 3 shows the process of personal memory retrieval.

4. VIDEO AND AUDIO ANALYSIS

This section gives a brief overview of the individual components that enable the multimedia retrieval of conversations.

4.1. Face Detection and Recognition

Extensive work in face detection has been done at CMU by Rowley [17, 18] and Schneidermann [20]. Face matching was used at CMU in the NameIt system [19] using the ‘eigenface’ approach. Meanwhile there have been several commercial systems offering face detection and identification, such as Visionics [25]. In our implementation we have been using the Visionics FaceIt toolkit for both face detection and matching. However, earlier versions of the prototype system have used [20] for detection and [19] for matching.

4.2. Speaker Identification

Speaker identification is done through our own implementation of Gaussian Mixture Models as described in [26]. The speaker identification system also uses the fundamental pitch frequency to eliminate false alarms. Generally, about 4 seconds of speech are required to get reliable speaker identification under benign environmental conditions.

4.3. Speech Recognition

For speech recognition, we use a near real-time version of CMU’s Sphinx-III speech recognition system [22]. Fortunately, since we are only processing selected snippets of conversation during the ‘memory collection’ phase, the speech recognition needs to be accurate, but does not have to be performed in real time. A language model of conversational English was adapted to include key phrases for commands. Speech recognition is only performed on the close-talking audio stream, since the omni-directional stream was found to be too noise for useful transcription. As a result, we only capture the wearer/speaker’s side of the conversation in transcript form.

4.4. Language Processing

The Phoenix [27] phrase-based parsing approach is used to attempt to identify selected ‘carrier’ phrases that indicate a name was likely to have been mentioned in the temporal vicinity of the phrase. For example, if the speaker says “Nice to meet you”, it is a good assumption that the dialog partner has just been introduced by name. This introduction becomes a key part of the later summary. Similar analysis yields critical information from the (omni-directional) dialog partner stream immediately around the (close-talking) speaker phrases such as ‘hello my name is Alex’, or ‘I’m sorry, what is your name again’.

4.5. Information Summarization

The audio stream to be summarized is selectively compacted s based on power. Similar to the video skims [28] and audio summarization [15, 16], only selected portions of the audio are played. Silences are used to define ‘cuts’, and low signal-to-noise ratio segments are eliminated. We have also implemented a TF.IDF weighting scheme to rank segments based on transcript words. So far, no optimal solution for producing a concise summary of the conversation has been found. As a result, the system tends to play back too much of the conversation, forcing the user to actively interrupt the summary playback.

5. RELATED RESEARCH

5.1. Wearable Computing and Remembering Human Experiences

Wearable computing technologies are vital in the development of this system because both the experience collection and the memory recall must be available ``on the spot. Many wearable devices have been invented and designed both in academia and industry, such as EyeTap [2], or the EyeGlass Display [1l. In addition to standard laptops, wearable computers can be bought off the shelf from xybernaut [3], Charmed [4], or ViA [5].

Some earlier attempts at wearable computing for the integration of different video, audio and sensor sources are described in [2] and embodied in research demonstrations such as the ‘StartleCam’ [29], which attempted to record video if the wearer showed increased ‘excitement’ according to a biometric sensor attached to the skin. Various research projects have explored the direction of integrating information into daily life through augmented reality, such as the WorkBench [6], SmartRooms [9], and the Wearable Museum [7] projects.

Several projects have started investigating how to collect and utilize human experience, for example, in the classroom [8] or when people type on the computer [10]. Some devices, like the Memory Glasses [11] and Forget-me-not [12] have been designed to help people or patients remember previous events. However, these devices assume everybody is wearing a digital identification and do not rely on video or audio processing.

6. CONCLUSION AND FUTURE WORK

The main focus of our system is on the integration of multi-modal human experience. The novelty of the system is in using the face and the audio cues to help remember essential details about the previous meeting with the same person, automatically creating a database of information associated with the face, the voice and the words.

In the future, we expect to extend the system along the following lines: Location information (through GPS) can be combined with the audio and face data. (Video) OCR can also be used in conjunction with the face and audio, e.g. to recognize nametags or business cards. The control of the interface through spoken commands is also slated for expansion to allow many different ways of actively querying the memory. Eventually the speech transcript may also enable audio indexing and search.

An intelligent assistant drawing from annotated personal history could overcome age and other limits to mental capacity and help recall the details needed in a given situation.

7. REFERENCES

[1] MicroOptical, http://www.microopticalcorp.com/

[2] S. Mann, Intelligent Image Processing, JWiley Inc, 2001.

[3] Xybernaut, http://www.xybernaut.com/},

[4] Charmed, http://www.charmed.com/},

[5] ViA, http://www.flexipc.com/

[6] B. Leibe, T. Starner, W. Ribarsky , Z. Wartell, D. Krum, J. Weeks, B. Singletary and L. Hodges, Toward Spontaneous Interaction with the Perceptive Workbench, IEEE Computer Graphics and Applications, 2000.

[7] F. Sparacino, G. Davenport and A. Pentland, Media in performance: Interactive spaces for dance, theater, circus, and museum exhibits, IBM Systems Journal, 2000.

[8] G.D. Abowd, Classroom 2000: An Experiment with the Instrumentation of a Living Educational Environment, IBM Systems Journal, 1999.

[9] K. Russell, T. Starner and A. Pentland, Unencumbered Virtual Environments, Proc. of IJCAI'95 Workshop on Entertainment and AI/Alife, 1995.

[10] B. Rhodes and T. Starner, Rememberance Agent: A continuously running automated information retrieval system, Proc. of Pract. App. of Intelligent Agents and Multi-Agent Tech, 1996.

[11] R. W. DeVaul, MemoryGlasses, http://wearables.www.media.mit.edu/projects/wearables/mithril/memory-glasses.html.

[12] M. Lamming and M. Flynn, Forget-me-not: Intimate Computing in Support of Human Memory, FRIEND21, Symposium on Next Generation Human Interface, 1994.

[13] Bush, V., As we may think, Atlantic Monthly, Vol.176, No. 1; pages 101-108, 1945

[14] Gray, J., What next? A few remaining problems in Information Technology, *ACM Federated Research Computer Conference*, Atlanta, GA, May 1999.

[15] Arons, B.M., *Interactively Skimming Recorded Speech*, Ph.D. Dissertation, MIT, February 1994.

[16] Arons, B.M., *Pitch-Based Emphasis Detection for Segmenting Speech Recordings*, ICASSP-94, Vol. 4, Yokohama, Japan, September 18-22, 1994.

[17] Rowley, H., Baluja, S. and Kanade, T. Human Face Detection in Visual Scenes. Carnegie Mellon University, *Technical Report CMU-CS-95-158*, Pittsburgh, PA.

[18] Rowley, H.A., Baluja, S. and Kanade, T. Rotation invariant neural network-based face detection, *IEEE CVPR*, Santa Barbara, 1998.

[19] Satoh, S., and Kanade, T. NAME-IT: Association of Face and Name in Video. *IEEE CVPR97*, Puerto Rico, 1997.

[20] Schneiderman, H. and Kanade, T. Probabilistic Modeling of Local Appearance and Spatial Relationships of Object Recognition, *IEEE CVPR*, Santa Barbara, 1998.

[21] Virage Corporate Web Site http://www.virage.com.

[22] Seymore, K., Chen, S., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., Ravishankar, M., Rosenfeld, R., Siegler, M., Stern, R., and Thayer, E., The 1997 CMU Sphinx-3 English Broadcast News Transcription System, DARPA Workshop on Broadcast News Understanding Systems (BNTUW-98), Lansdowne, VA, February 1998.

[23] Wactlar, H.D., Kanade, T., Smith, M.A. and Stevens, S.M. "Intelligent Access to Digital Video: Informedia Project". *IEEE Computer*, **29**(5), 46-52.

[24] Wactlar, H.D., Christel, M.G., Gong, Y., and Hauptmann, A.G. "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library", *IEEE Computer* **32**(2): 66-73.

[25] Visionics FaceIt Developer Kit, http://www.visionics.com

[26] Schmidt, M., Golden, J., and Gish, H. "GMM sample statistic log-likelihoods for text-independent speaker recognition," *Eurospeech-9*, Rhodes, Greece, September 1997, pp.855 - 858.

[27] Ward, W. "Understanding Spontaneous Speech: the Phoenix System." ICASSP'91 (1991), pp. 365-367.

[28] Smith, M. and Kanade, T. Video skimming and characterization through the combination of image and language understanding techniques, *IEEE CVPR97*,(San Juan, Puerto Rico, 1997), 775 – 781.

[29] J. Healey and R. Picard**, "**StartleCam: A Cybernetic Wearable Camera," *ISWC '98*, October, 1998.