Informedia Experience-on-Demand: Capturing, Integrating and Communicating Experiences across People, Time and Space

HOWARD D. WACTLAR, MICHAEL G. CHRISTEL, ALEXANDER G. HAUPTMANN, YIHONG GONG
Carnegie Mellon University, Pittsburgh, PA {wactlar, christel, hauptmann, ygong}@cs.cmu.edu

ABSTRACT

The Informedia Experience-on-Demand system uses speech, image, and natural language processing combined with GPS information to capture, integrate, and communicate personal multimedia experiences. This paper discusses an initial prototype of the EOD system.

KEYWORDS

Multimedia information systems; audio, video and location content analysis

INTRODUCTION

The Informedia Experience-on-Demand Project (EOD) develops tools, techniques and systems allowing people to capture a record of their experiences unobtrusively, and share them in collaborative settings spanning both time and space. Users may range from rescue workers carrying personalized information systems in operational situations to remote crisis managers in coordinating roles. Personal EOD units record audio, video, Global Positioning System (GPS) spatial information, and other sensory data, which can be annotated by human participants. The EOD environment synthesizes data from many EOD units into a "collective experience" – a global perspective of ongoing and archived personal experiences. Distributed collaborators can be brought together over time and space to share meaning and perspectives.

Each constituent EOD unit captures and manages information from its unique point of view, as illustrated in Figure 1. This information is transferred to a central site where the integration of multiple points of view provides greater detail for decision-making and event reporting. A longer term goal, dependent on advances in communication technology, is for each portable EOD unit to be not only a data collector but also a data access device, interoperating with the other EOD units and allowing audio and video search and retrieval.

The foundation for this work, the Informedia Digital Video Library Project [Christel *et al.* 1996], has demonstrated the applicability of speech, image, and natural language processing in automatically creating a rich, searchable multimedia information resource holding over 1000 hours of video. We have built a prototype EOD system that builds on these technologies by addressing continuously captured, unstructured, unedited video in which location data is added as another information dimension. This system will be discussed in the context of gathering personal experiences in the Pittsburgh, PA area, querying and displaying those results geographically, and combining multiple views into common perspectives.

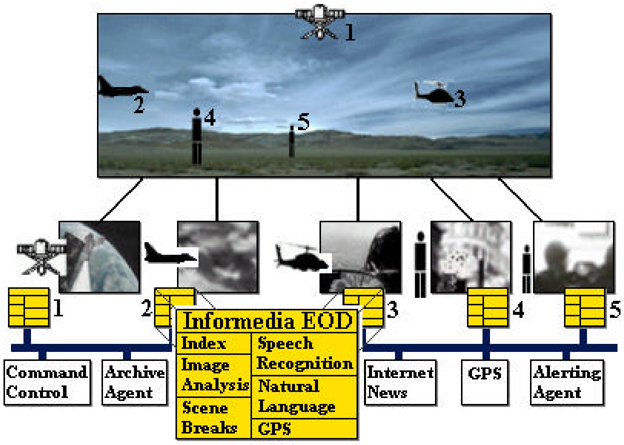


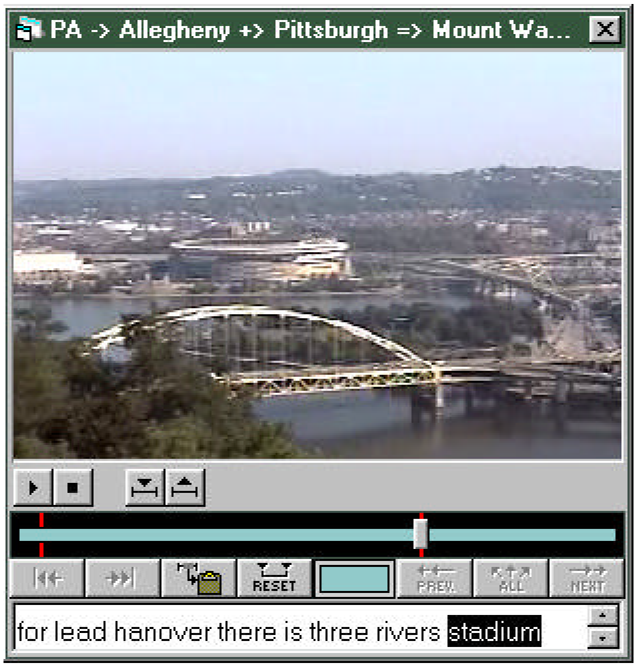
Figure 1. Five unique viewpoints in EOD environment

MULTIMEDIA PERSONAL EXPERIENCES

We have experimented with portable microphones, cameras and GPS units wired into a wearable vest and a hat for capturing audio and video with corresponding time and location in the field. We have collected over 40 hours for our initial trials. While we have not focused on improving the form factor of the recording devices – making them less cumbersome and visible – we have instead concentrated on determining whether the quality of the resulting audio and video was suitable for subsequent automatic processing.

We found that microphone type and placement with respect to the speaker greatly affected the accuracy of follow-up speech recognition, with usable results of 10% word error rate or less capable from mobile talkers. Through the use of high accuracy GPS and digital cameras we could reliably recreate panoramic views from various perspectives. Field experiences captured in audio and video as a form of personal memory are hence suitable for subsequent processing and use in collaborative settings.

By tailoring speech recognition for mobile, active talkers, we hope to improve the quality of the resulting text transcript that is used to index material in the multimedia experience database. For example, a search on "stadium" would bring up the video where "stadium" is mentioned, as shown in Figure 2. The scrolling transcript at the bottom of the figure reflects the current state of the recognizer, where the output does not exactly match the spoken dialogue of "<wind noise> and over there is Three Rivers Stadium." We are exploring enhancements to existing speech recognition algorithms to filter out the background noise typical in audio collected in outdoor environments. By optimizing language models for information retrieval accuracy rather than transcript word error rate [Witbrock and Hauptmann 1997], we hope to further improve the utility of the speech recognition output for indexing the experience database.



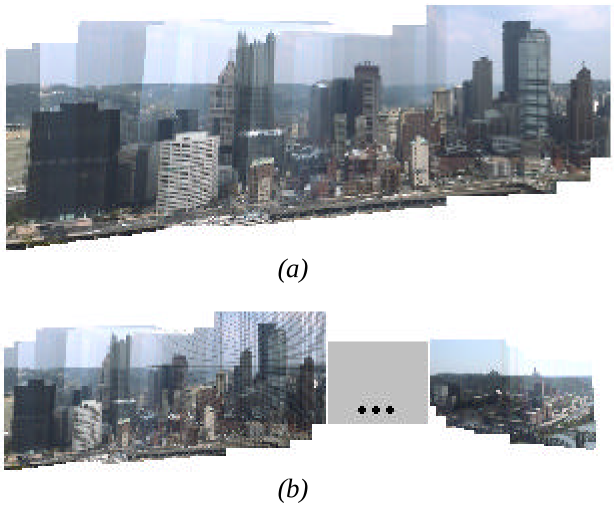
for lead hanover there is three rivers stadium

Figure 2. Finding video through search on text transcript generated via speech recognition

Similarly, we are modifying Informedia image processing modules to better work with field-captured motion video. Our current object detectors, for recognizing and matching faces and overlaid text, work well on broadcast news given certain assumptions, such as a well-lit face looking directly at the camera. These assumptions are less likely to be met with field video, and so we are investigating more robust techniques for object detection within video having varying shades of lighting and where the object of interest may appear at varying resolutions. One example is a face detector that will recognize profiles as well as full frontal shots of faces [Schneiderman and Kanade 1998]. A longer-term goal is to extend these image processing modules so that detection is coherent over time, enabling object tracking.

SYNTHESIS OF PERSONAL EXPERIENCE DATA

Our trial runs with mobile EOD systems have collected highly redundant video data. Long sequences of video contain little or no audio, with overlapping visual imagery. Filtering across space for these shots can be accomplished via image processing techniques that exploit EOD location data acquired through GPS. One strategy is to generate a 2-D panoramic view of the environment by combining personal views based on their time, location, and viewing angle. Consider multiple EOD systems recording city impressions at various viewing points; one system's output is shown in Figure 3a. GPS for each system is used to merge the visuals so that a panoramic view as shown in Figure 3b can be constructed. The shade variations have been left in to show that the panorama was generated from individual shots captured at different times with varying amounts of sun and clouds; these shadings could be filtered out to produce a smoother panorama. The box area labeled "..." in Figure 3b indicates a portion of the cityscape for which no viewer has yet contributed information.



(a)

(b)

Figure 3. Panorama for: (a) one view, and (b) wider pan generated from multiple views for user-selected viewpoint

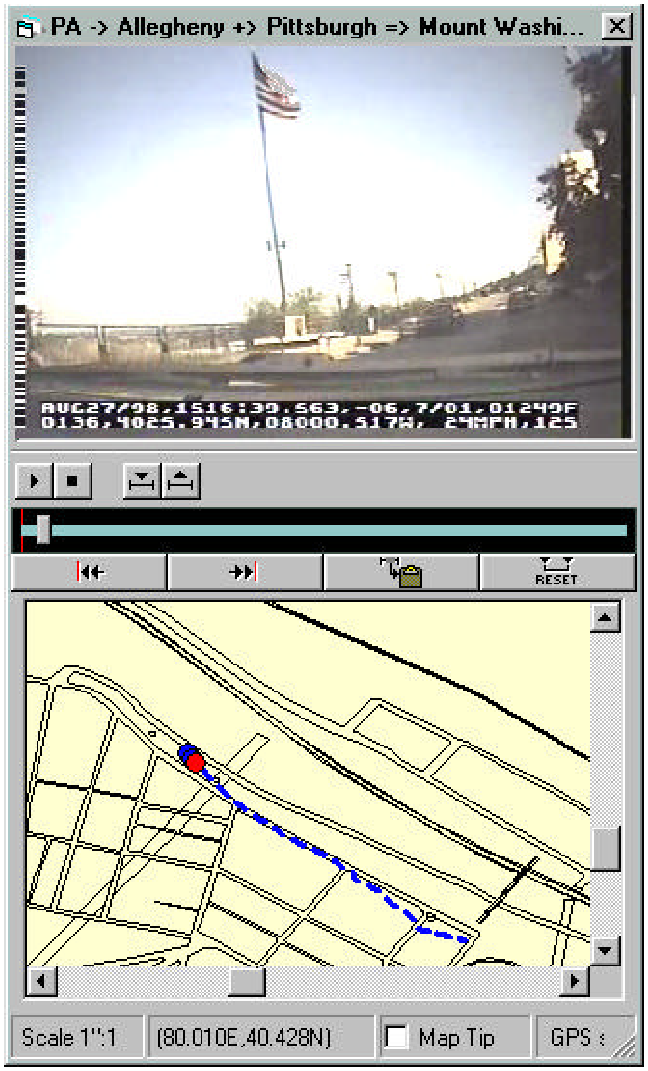
Cooperative users can help address deficiencies in the automatic EOD processing. A user can speak into a microphone and tag his or her experience data with “who”, “what,” “when,” and “where” descriptors. These descriptors can be then used to segment data into episodes.

A longer term goal is to enable the user to direct which interesting events should be automatically flagged in the data as requiring further inspection. Gong and his colleagues at Singapore have developed a Scene Description Language (SDL) for video [Gong *et al.* 1995]. The SDL can be extended into an advanced experience parser. A user directs the image processing analysis via the SDL: which video features to be focused (such as images of flying birds), the events to be tracked (period when a bird is in a particular area), and the changes to be monitored (when a bird enters the field of view for a particular EOD unit). When user-defined significant events are detected, alerts are communicated across the EOD environment.

INTERFACE CONSIDERATIONS

Figure 4 shows how continuously recorded GPS data can be used to tightly synchronize a playing video to a map. As the video plays, the location for the displayed video image is highlighted on the map, with the area covered within the video segment shown on the map as well. The map is both *active*, changing as the video plays, and *interactive*, allowing the user to modify the map display and use it to issue spatial queries to locate experiences in specified areas. A reasonable next step is to augment the map with the panoramic realism shown in Figure 3. Future interface work will incorporate information filtering, 2-D panoramic views and 3-D reconstructed views to enable “pan-and-zoom” access to experience data organized by temporal, spatial, and viewpoint attributes.

Information filtering based on a usage context will be used to identify the particular data attributes that need to be visualized. Semantic zooming will be employed, where objects change their appearance as the amount of display real estate given to them changes [Furnas and Bederson 1995]. As a fully zoomed out view, the experience space is fully aggregated and a select set of attributes may be shown concerning the whole space. By drilling down, detail is added to either the visualization as a whole (more attributes are shown), to the attributes in the view (e.g., text labels are added), or to a subset of the full experience space (e.g., opening up a particular artifact in the space and visualizing its detail). This concept of "drilling down" or zooming in to add dimensions or detail to the display and zooming out or "rolling up" to aggregate and hide detail has been implemented successfully in systems like Visage [Roth *et al.* 1996] and Pad++ [Furnas and Bederson 1995]. The importance of preserving context while zooming into a document space has been empirically validated [Schaffer *et al.* 1996]; such context preservation will be a design constraint for our work.


PA -> Allegheny +> Pittsburgh => Mount Washi...

Scale 1":1 | (80.010E,40.428N) | □ Map Tip | GPS
AUG27/98,1516:39.563,-06.7/01,01249F
0136,4025.945N,08000.517W, 24MPH,125
RESET

Figure 4. Video with map showing trajectory of motion (as dotted line) for full video segment and location (upper left of dotted line) associated with the shown video frame

The Informedia Digital Video Library Project defined numerous abstractions for structured broadcast-quality video, including text titles, thumbnail images, filmstrips, and skims [Wactlar *et al.* 1996, Christel *et al.* 1998]. These abstractions are now being extended to address the unbounded continuous nature of experience video, and to move from words as the primary information and indexing source to audio/image interdependence. Our goal for EOD interfaces is to allow the information to be quickly and effectively accessed, queried, viewed, abstracted, navigated, summarized, and annotated along dimensions of time, space, and user perspective.

CONCLUSION

Experience-on-Demand addresses collaboration and summarization of multiple simultaneous information generators integrated across people, time, and space. A wealth of information will be collected through the Informedia EOD environment. This data, collectively referred to as a "personal experience," has potential value in the following situations:

- It serves as a useful memory reference for that particular point of view.
- It fills in missing information for another point of view.
- It aids in strategic planning at a command center view.
- It better documents an unfolding situation when combined with other points of view.
- It archives details, which may be of interest in future training and simulation-building exercises.

The data is by nature voluminous in size yet sparse in information content, with tremendous redundancy along the temporal and spatial dimensions and across points of view. To help analysts rapidly and effectively derive the relevant meaning from this large body of data, the EOD environment is being developed to support intelligent information analysis, organization, and manipulation techniques.

ACKNOWLEDGEMENTS

This material is based on work supported by the Defense Advanced Research Projects Agency and NRaD under contract number N66001-97-C-8517. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of DARPA, the Office of Naval Research, or the U.S. Government. Special thanks go to Bryan Maher, Mark Dambacher and Ricky Houghton for their efforts with the initial EOD trials.

REFERENCES

CHRISTEL, M., STEVENS, S., KANADE, T., MAULDIN, M., REDDY, R., and WACTLAR, H. Techniques for the Creation and Exploration of Digital Video Libraries. *Multimedia Tools and Applications*, B. Furht, ed. Boston, MA: Kluwer Academic Publishers, 1996, Chapter 8.

CHRISTEL, M., SMITH, M., TAYLOR, C.R. and WINKLER, D. Evolving Video Skims into Useful Multimedia Abstractions. *Proc. ACM CHI '98* (April 1998), 171-178

FURNAS, G. and BEDERSON, B. Space-Scale Diagrams: Understanding Multiscale Interfaces. *Proc. ACM CHI '95* (May 1995), 234-241.

GONG, Y., SIN, L.T., CHUAN, H.C., ZHANG, H.J., and SAKAUCHI, M. Automatic Parsing of TV Soccer Programs. *Proc. 2nd IEEE Conf. Multimedia Computing and Systems* (May 1995), 167-174.

ROTH, S., LUCAS, P., SENN, J., GOMBERG, C., BURKS, M., STROFFOLINO, P., KOLOJEJCHICK, J., and DUNMIRE, C. Visage: A User Interface Environment for Exploring Information. *Proc. IEEE Information Visualization* (October 1996), 3-12.

SCHAFFER, D., ZUO, Z., GREENBERG, S., BARTRAM, L., DILL, J., DUBS, S., and ROSEMAN, M. Navigating Hierarchically Clustered Networks through Fisheye and Full-Zoom Methods. *ACM Trans Computer-Human Interaction* 3 (1996), 162-188.

SCHNEIDERMAN, H., and KANADE, T. Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition. *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)* (June, 1998),

WACTLAR, H., KANADE, T., SMITH, M., and STEVENS, S. Intelligent Access to Digital Video: Informedia Project. *IEEE Computer* 29, 5 (1996), 46-52.

WITBROCK, M. and HAUPTMANN, A. Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents. *Proc. ACM Digital Libraries '97* (July 1997), 30-35.