An end-to-end eChronicling System for Mobile Human Surveillance

Gopal Pingali, Ying-Li Tian, Shahram Ebadollahi, Jason Pelecanos, Mark Podlaseck, Harry Stavropoulos

IBM T. J. Watson Research Center

19 Skyline Drive
Hawthorne, NY 10532

1. Introduction

Rapid advances in mobile computing devices and sensor technologies are enabling the capture of unprecedented volumes of data by individuals involved in field operations in a variety of applications. As capture becomes ever more rich and pervasive the biggest challenge is in developing information processing and representation tools that maximize the utility of the captured multi-sensory data. The right tools hold the promise of converting captured data into actionable intelligence resulting in improved memory, enhanced situational understanding, and more efficient execution of operations. These tools need to be at least as rich and diverse as the sensors used for capture, and need to be unified within an effective system architecture. This paper presents our initial attempt at such a system and architecture that combines several emerging sensor technologies, state of the art analytic engines, and multi-dimensional navigation tools, into an end-to-end electronic chronicling solution for mobile surveillance by humans.

2. System Architecture and Overview

Figure 1 presents a conceptual overview of our system. The left part of the figure depicts support for a variety of wearable capture sensors and personal devices to enable pervasive capture of information by individuals. Supported sensors include digital cameras, microphones, GPS receiver, accelerometer, compass, skin conductance sensors, heart rate monitors etc. The user captures data through these sensors, which are either worn or carried. On-board processing on a wearable computer provides real-time control for data capture, and to some extent, local analysis of captured data. In addition to data captured through wearable sensors, the system also allows input of event logs and corresponding data from personal devices such as personal computers and PDA's. The user can also enter textual annotations both during wearable capture and while working on their PC/PDA. The data thus captured is stored with appropriate time stamps in a local "individual chronicle repository". This electronic chronicle, in short, represents a rich record of activity of the individual obtained both from wearable sensors while in the field and event loggers on their PC while at their desk.

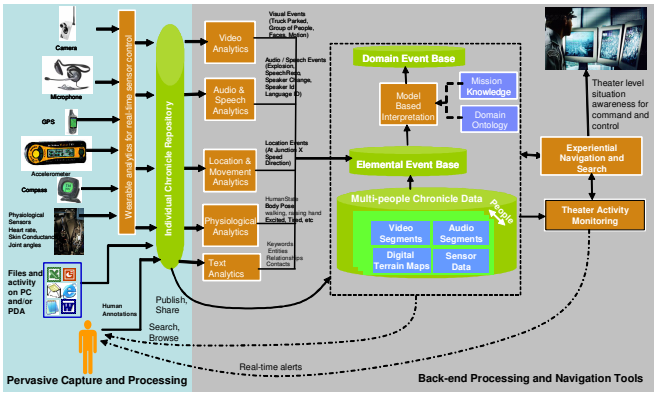

Figure 1 Overview of the end-to-end electronic chronicling system

The right portion of Figure 12.1 depicts the back-end processing and navigation tools to analyze, extract, manage, unify, and retrieve the information present in electronic chronicles. First, a variety of analytic engines process the chronicled data, including those that process image/video, speech/audio, text, location and movement, and physiological data. The results of this analysis are "elemental" events detected at the sensory data level. These represent meta-data or machine derived annotations representing events in the original captured data. Examples of such events include detection of a face, a moving vehicle, somebody speaking a particular word etc. The system also includes a data management component that stores the raw chronicle data, the elemental events, as well as domain-level events obtained from further analysis of the elemental events. The latter are events that are expressed in terminology specific to a domain based, for

1-4244-1180-7/07/$25.00 ©2007 IEEE

example, on mission knowledge, and specific ontology. For example, detection of an "explosive sound" or "visible fire" is an elemental event in our terminology while detection of a "fuel gas incident of type propane involving toxic release with high severity" is a domain specific event.

The right-most portion of Figure 1 indicates tools for navigation and retrieval of the chronicled data and associated events and metadata. These tools communicate with the data management component and allow the user to search, explore, and experience the underlying data. These tools apply at an individual level for browsing personal data and also across data shared by multiple individuals. In the latter case, they enable overall situation awareness based on data from multiple people and enable theater-level search and planning.

Plug-in architecture: An important challenge in this kind of a system is flexibility to adapt the base architecture to different domains with differing sensing, analytics, and navigational needs. In order to provide such flexibility we provide a clean separation between the sensors, the analytics, the data management, and the navigation/retrieval components. Standardized XML interfaces define ingestion of data from the sensors into the database. Similarly, XML interfaces are defined between the analytics components and the database. The analytics components, which can be distributed on different servers, query the database for new data via XML, appropriately process the raw data, and ingest the processed results back into the database. Finally, the navigation tools retrieve data and present it to the user as appropriate. This architecture allows the system to be distributed, easily changed to add or remove sensors, add or replace a particular analytic component, or modify the navigation tools. Thus, for example, the same base architecture is able to support synchronized automated capture of audio, video, and location in one domain while allowing manual capture of data in another domain.

3. Multimodal Event Analytics and Example Results

We focus here on three types of analytics – image classification, face detection, and speech/audio analysis. Image classification helps in searching through the numerous images captured by the user based on concepts associated with the images(*Outdoors, Indoors, Vegetation, Vehicle_Civil (Car, Truck, Bus), Vehicle_Military, Person (Soldier, Locals), Weapon, Building)*. Face detection aids in automatically retrieving those images in which there were human faces (frontal and profile views). Speech/audio analytics help in a) transcribing and extracting keywords from the annotations made by the user when on their mission; and b) analyzing environmental sounds such as other people talking, sounds of vehicles, explosions etc. Together, these analytics aim to enhance the user's ability to identify and retrieve interesting events that occurred during missions. The interface and some example results are shown in Figure 2, 3, and 4.

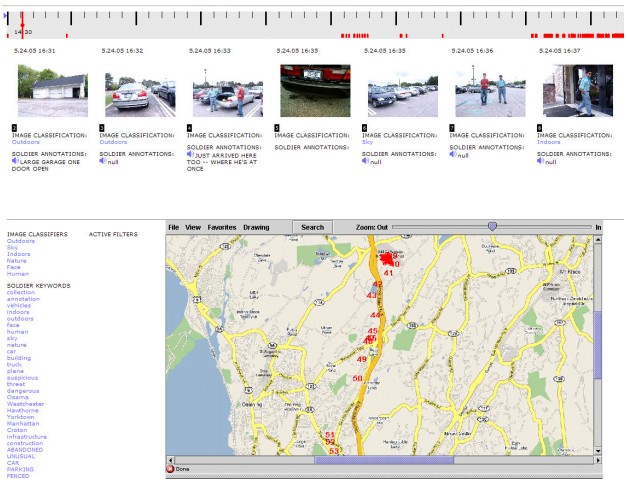

Figure 2 Example result: User views all data from a trip without filtering. Notice the ability to view the data by space, time, concepts, and keywords.

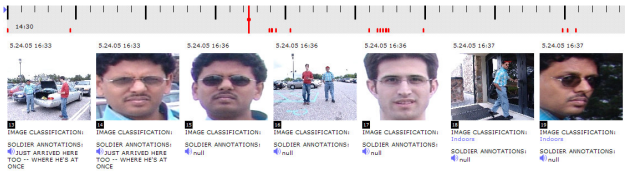

Figure 3 Example result: User filters data in Fig. 12 to view only images with "faces". This view also shows the sub-images of detected faces in each original image.

Figure 4 Example result: User further filters the images with "faces" to view only those labeled as "indoors"

ACKNOWLEDGMENTS

This work was partially supported by DARPA under the ASSIST program under contract NBCHC050097. The authors would like to acknowledge the input and influence of the rest of the people in the EC-ASSIST project team at IBM, Georgia Tech, MIT, and UC Irvine. The authors specially thank Milind Naphade for providing the models trained for LSCOMLite ontology.