Chapter 12

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

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 [29, 40] solution for mobile surveillance by humans.

12.1 INTRUDUCTION: MOBILE HUMAN SURVEILLANCE

There are a number of applications today in which information is collected in a mobile and pervasive manner by numerous people in the field going about their jobs, businesses, lives, and activities. For instance, law enforcement personnel patrol certain areas, look for interesting/suspicious activity and report and take action on such activities for security, surveillance, and intelligence gathering purposes. In this kind of an application there is a lot of value in being able to access, review and analyze the information gathered by people in the field both by the individual gathering the information as well as their peers and higher officers who would like to compare notes, integrate information, form a more complete picture of what is happening, and discover new patterns and insights. Similarly emergency personnel such as fire departments and medical emergency teams can record their activities and review operations for effectiveness, failure points, people involved, operational insights etc. Another example is the army in which soldiers are involved in battlefield, peace-keeping, and anti-insurgency operations.

System Goals: Our goal is to develop an end-to-end electronic chronicling system for mobile workers to much more effectively capture, relive, analyze, report, and re-use their experiences from field operations, enabling:

Auto-diary Creation: As a worker operates on a mission, her wearable system should generate a richly annotated multimedia diary.

Chronicle Navigation and Reporting Tools: The worker should be able to produce effective after-action mission/intelligence reports by using a unified chronicle management and navigation system that enables them to find events and content of interest, drill down to the desired level of detail, and find interesting correlations.

Theater Level Navigation and Search: The system should enable appropriate people (such as a higher officer) to combine and navigate the chronicles from multiple workers to obtain the bigger picture. The system should also allow individual workers to subscribe to events of interest and receive automatic notifications.

To this end we are developing a system and architecture (Figure 12.1) that supports multi-modal and multi-sensory capture, provides a variety of sensor and data analytics to detect and extract events of interest, and provides multi-dimensional navigation and search capabilities.

Technical challenges: There are several challenges in developing such an electronic chronicling system.

Wearable capture system: This should be easily carried by a mobile worker, integrate appropriate sensors to create a rich record of the user's activities, ensure synchronized real-time capture of all data, support real-time annotation (both automatic and manual), and work uninterrupted to last the length of the worker's mission.

Analytics for event detection: These should analyze the significant volume of multi-sensory chronicle data to extract events of interest, enabling the users to access the data by such events. Clearly, the challenges here span the gamut of individual sensory processing technologies such as image analysis, computer vision, speech recognition, movement analysis, and location understanding. The challenge is also to combine these analytics effectively into an integrated system that presents the users with the ability to review episodes and derive situational understanding.

Event and data management: The system should have appropriate representations and assimilation mechanisms to link raw data, events derived from sensory processing, and domain-level events of interest to end users.

Experiential navigation techniques: The utility of such a chronicling system is heavily dependent on the kind of mechanisms provided for browsing, filtering, and searching. End users should be able to relive their experience or the experiences of others, look for events and episodes of interest, or perform deeper analysis to derive valuable insights.

Clearly, our goals are ambitious and demand an ongoing research effort, both on individual technologies and also a platform that effectively unifies these technologies into an end-to-end electronic chronicling solution. In this paper, we provide an overview of our approach and present a version of such an integrated eChronicling system.

12.2 RELATED WORK

The most popular related research theme to eChronicling has been personal information management and retrieval of personal information [4,12,13,16,17,18,38]. The advent of wearable devices and wearable computing [50] has had significant influence on the recent emergence of such efforts [e.g. 48] around personal chronicles that enable rich capture of an individual's life and retrieval based on context [24]. Other efforts [11,15,42,47] have addressed the difficulties in organizing the regular information stored on computers. The area of group and enterprise level chronicles [21,27,31] remains largely unexploited, especially relative to the ongoing efforts on personal chronicles. A promising area for research is mining of captured data [36] and generation of associated alerts and notifications. Another related area is distributed event-based systems [7,10,14,28,41], although these have not yet addressed event extraction from multi-sensory captured data as discussed here.

Video Surveillance is the use of computer vision and pattern recognition technologies to analyze information from situated sensors [8, 20, 22, 34]. The key technologies are video-based detection and tracking, video-based person identification, and large-scale surveillance systems. A significant percentage of basic technologies for video-based detection and tracking were developed under a U.S. government-funded program called Video Surveillance and Monitoring (VSAM) [8]. This program looked at several fundamental issues in detection, tracking, auto-calibration, and multi-camera systems [23, 25, 51]. There has also been research on real-world surveillance systems in several leading universities and research labs [52]. The next generation of research in surveillance is addressing not only issues in detection and tracking but also issues of event detection and automatic system calibration [60]. The second key challenge of surveillance—namely, video-based person identification—has also been a subject of intense research. Face recognition has been a leading modality with both ongoing research and industrial systems [2, 39]. A recent U.S. government research program called Human ID at a Distance addressed the challenge of identifying humans at a distance using techniques like face at a distance and gait-based recognition [26]. One of the most advanced systems research efforts in large-scale surveillance systems is the ongoing U.S. government program titled Combat Zones That See [9]. This program explores rapidly deployable smart camera tracking systems that communicate over ad hoc wireless networks, transmitting track information to a central station for the purposes of activity monitoring and long-term movement pattern analysis. There are several technical challenges that need to be addressed to enable the widespread deployment of smart surveillance systems. The video surveillance systems which run 24/7 (24 hours a day and seven days a week) create a large amount of data including videos, extracted features, alerts, and etc. How to manage this data and make it easily accessible for query and search are other challenges.

Audio analytics, also known as audio scene analysis [3,6], provides the ability to label acoustic events in an audio recording. Some examples include performing speech transcription, speaker identification and identifying machine sounds. Audio surveillance has traditionally taken second position after video monitoring in terms of its widespread use. However, it is recognized that in many applications, audio analytics provide complementary information to video analytics. Research outcomes in audio-visual speech recognition [44] exemplify this. A significant benefit of audio analytics is the in-depth detail that is potentially available. One example includes a person being recorded while discussing a complex situation in their environment. In essence, the human in the environment analyzed the context of the situation and then provided a verbal description of the events that occurred. The audio recording may then be readily transcribed, indexed and searched. Past research in the audio scene analysis area has investigated many categories of audio events from speech recognition [46] to snore and sleep apnea detection to finding explosions in videos [32].

12.3 SYSTEM ARCHITECTURE AND OVERVIEW

Figure 12.1 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 (for more on this, refer to [21,31,33]). 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 [29,40], 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 12.1 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 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 12.1 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. This plug-in approach is based upon and inspired by [22].

Figure 12.2 Simplified view of the eChronicling system

For the rest of this paper, we will focus on an implementation of the part of this system that involves automatic mobile capture and back-end analysis and navigation of the captured data. Figure 12.2 shows a simplified view of the electronic chronicling architecture in Figure 12.1 to focus on this mobile capture system with attendant back-end processing. This system consists of a wearable capture component that can be used for automatic capture in the field. The system was created to support an open-ended assortment of data gathering devices: digital/video cameras, microphones, GPS trackers, skin conductance sensors etc.

Figure 12.3 shows examples of capture devices and two different examples of users sporting a wearable capture system.

Figure 12.3 (a) Example of Devices used for capture; (b) one example of a user wearing a capture system consisting of devices in (a) (c) another example of a user wearing an extended set of capture devices

There is also flexibility in on-board versus back-end processing. Some of this data may be processed on-board a wearable computer; some will be sent real-time to the back end-system for processing; some will have to wait for a more opportune time for ingestion to the back-end. The user can annotate this data during capture (if/when possible) and certainly after ingestion.

Once ingested the data is analyzed by analytics engines. The results of this analysis are "elemental" events detected at the sensory data level (e.g. detection of a face, a moving vehicle, somebody speaking a particular word etc); these results are ingested in the database (which could be different than the one containing the sensor data).

The ingested results of the analysis may be

a) fed to another analytics engine which will try, during a second pass, to "correlate" them (for example correlate the detected-by-the-image-analysis vehicle with the detected-by-the-audio-analysis motor sound, to enhance the confidence level that a vehicle was indeed present).

b) the results of b) may be similarly further analyzed to derive domain-specific events. 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.

Eventually some GUI client will retrieve what has been already analyzed (note that the analysis need not be completed). The client should summarize the information; a user, however, should be able to track down any detected "event" to the media captured. Another set of clients will be notified the moment some domain-specific event gets ingested.

12.4 EVENT MANAGEMENT

The eChronicling system fundamentally manages "Events" – data entities with an associated time stamp/time interval and often with a location stamp. The following types of events are included:

a) Elementary events, detected at the sensory level. There are specified either by the corresponding analytics engine or by a user when browsing the data.

b) Composite events. In two flavors:

b1) As in a) only surmised from more than one medium (e.g. "car in a location" may be inferred from two pictures and the audio of what-was-recognized-to-be an engine)

b2) Higher-level, domain-specific events

c) Semi-independent. Typically events are directly related to some media (e.g. "car in certain a location" will be inferred from the images). A soldier can annotate a part of a patrol as "chasing the suspect". During that time interval several images will be displaying the suspect. Others will not. The "event" is associated implicitly (by its duration) with a subset of the pictures/video/audio.

In order to accurately represent events, we must ensure that the devices are properly synchronized so that temporal searches and correspondences are accurate. We take the GPS time to be the master clock and use the OQO clock to compute offsets between GPS readings. We estimated the camera delay by taking a series of automated captures of a graphical display of the OQO's clock and computing the time difference between issuing the capture command and the visible time in the photograph. For manually captured images, we read the capture time embedded in the EXIF data and offset this by a constant determined by photographing the OQO display as before. The second issue is that of labeling each voice annotation and image with a GPS coordinate and correlating images with audio clips. In both cases, we use temporal correspondence to guide the mapping. Currently, we search the GPS log for the nearest location and make the association if the difference is less than five seconds.

12.4.1 Storage

Each kind of event is stored in its own table; this provides for "cleaner" data. Alternatively we would need to encode extra information in the table (i.e. what kind of "event" a certain row represents). The downside is that the single-table approach is more extensible; new kinds of events can be specified with less database interference. The current thinking is that this more flexible path should be taken in the future.

User interfaces want to look at the events chronologically/by location/type etc. Some of this aggregation happens at the database level, some happens in the client. Since we anticipate an open-ended set of interfaces, what happens where will certainly vary: things that need time-consuming SQL joins (e.g. "retrieve all images of cars grouped by detected plate number, taken the past month, in those two towns") should be computed in the database (or, even, pre-computed); simpler data relationships can be managed/cached by the client.

Figure 12.4 Overview of types of data associated with events and their representation scheme

12.4.2 Representation

There are three categories of data associated with events as seen in Figure 12.4.

1. Sensor data. These are immutable

2. Annotations. These are either output of the analytics or comments manually/vocally entered by humans.

3. Correlations and Histories of navigation of the data by a user using a GUI client

The constraint here is the ability to do efficient SQL-type joins for data retrieval. In the current incarnation of the system, all entries are stored in the same database (more details in section 5). Sensor data may be stored anywhere - only the URLs are stored in the tables. Also, every datum has a UUID associated with it. This was created to be used as a foreign key to correlate the data in the tables as well as a means to maintain sanity once the sensor data starts getting replicated in order to be cached/distributed/performance-scaled.

12.4.3 Retrieval

During the development of the system it became clear that the retrieval requirements for the data defied prediction. At a minimum, an SQL query would be needed for each such access as well as a "hosting" script (which could be in the form of a servlet, a CGI script, a DLL loaded by the client, or anything else). The decision was made to create a very generic http-accessible script and let the client submit the full-blown SQL query (this can be done in a semi-transparent way, with the database maintainer making the query available for download by the clients). This way the set of queries is open-ended; the server does not need to be touched as new ones are developed.

12.5 MULTIMODAL ANALYTICS

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. Face detection aids in automatically retrieving those images in which there were human faces. 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.

12.5.1 Image Classification

Numerous images or long hours of video data could accumulate very quickly in the context of electronic chronicling applications. Sifting through these visual data for when a "building" was seen or a "vehicle" was spotted becomes a challenging task if one only relies on time and space-based navigation tools. Efficient means for accessing these data via the *semantic concepts* they portray becomes indispensable. Our goal is to equip the electronic chronicling system with automatic concept tagging capability for visual data to provide such means of interaction.

Automatic semantic concept tagging in image-based data requires bridging the semantic gap for the domain of application of the chronicling system. As evidenced by the top performing approaches reported in video tagging benchmarking exercises, such as TRECVID, machine learning based methodologies for tagging are becoming the de facto standard. In our chronicling system, we also train discriminative classifiers using both global and local image features extracted from a collected set of positive and negative example images for each concept of interest.

Figure 12.5 Overview of image tagging scheme.

Figure 12.5 gives an overview of our approach for image classification. As shown in the figure, for each distinct data set, where distinction is due to the nature of the images and the image collection process, a *Support Vector Machine* (SVM) [58] classifier is learned for different types of image features in that data set. Different classifiers for the same data set are fused to provide one semantic concept classifier for each individual semantic concept tag which the user is interested in and annotation has been provided for the training data set. Classifiers obtained for any given semantic concept from various data sets are then fused to obtain a single classifier for that concept.

The reason for employing various data sets is to cover the *multi-view* manifestations of the same concept for different applications and context. For example, concept vehicle could have a different visual manifestation in consumer photos than its visual manifestation in military-related applications. In other words the multiple data sets and therefore models for a given concept are employed to address the issue of visual polysemy for a given concept.

Figure 12.6 Illustration of image tagging.

The result of applying the array of semantic concept taggers to a test image is a set of confidence values associated to the image by each of the concept models. Only those concept tags are assigned to the image, which have a confidence value greater than a given threshold. This threshold is obtained such that it best matches the satisfaction of a human user. We obtain this threshold from a set of training data before hand using a utility maximization scheme, referred to as T10U [55].

The confidence values, which pass the threshold, are then converted to posterior probabilities using the sigmoid function according to Platt [43], in order to be used for further fusion with the outputs of other modules of the electronic chronicling system. Figure 12.6 illustrates the process of assigning semantic concept tags to an image.

12.5.2 Face Detection and License Plate Recognition from Images

12.5.2.1 Face Detection

Face detection is a challenging problem because of variations in pose, lighting,

Face detection is a challenging problem because of variations in pose, lighting, expression, and occlusions (beards, mustaches, glasses, hair, hat etc.) There are a number of techniques that can successfully detect frontal faces, such as Neural Networks [49], statistics of parts [53,56], and Adaboost learning methods [30,35,57]. In our system, we implemented multi-view face detectors based on Haar and optimized wavelet features by using a cascade of boosted classifiers.

Figure 12.7 Example Harr-like features for face detection

Figure 12.8 Example optimized wavelet features for face detection. (a) Original image; (b) – (e) Wavelet Representation.

Haar features are very useful for training extremely fast detectors. However, when learning from small datasets, these features are limited to produce robust detectors. In fact, the choice of good features is very important for projecting boosted classifiers when few examples are provided. We observe that discriminative features in general tend to match the structure of the object. For example, the first selected Haar features encode the fact that eyes are darker than nose, etc. This suggests us to use an optimization technique – wavelet networks – that creates features (Gabor, Haar, etc.) whose scale, orientation and position aligns with the object local structure. So, our approach is to learn optimized features for each specific object instance in the training set and then bring together these features into a "bag of features". Then we project a cascade classifier that selects the most general features from the bag/pool of features for discrimination.

Figure 12.9 Cascade of classifiers for face detection

First, the input image is scanned across location and scales. At each location, an MxM (20x20 in our system) image sub-window is considered. Following [57], a boosted collection of five types of Harr-like features (Figure 12.7) and a bag of optimized wavelet features (Figure 12.8) are used to classify image windows. Then, cascade classifiers are employed to evaluate the input sub-window (Figure 12.9). If the first classifier of the cascade returns false then computation on the sub-window stops and the sub-window is non-face. If the classifier returns true, then the sub-window is evaluated by the next classifier in the cascade. Face detection is declared only if a sub-window passes all the classifiers with all returning true. Since most sub-windows in an image are non-faces, they are rejected quickly. This process greatly improves computational efficiency as well as reduces the false positive rate [30,57]. Figure 12.10 shows some example results.

Figure 12.10 Examples of face detection result

12.5.2.2 License Plate Detection

We also integrated the license plate recognition into our system. The license plate recognition technology is licensed from Hi-Tech Solutions [54]. This technology could be deployed at the entrance to a facility where it catalogs the license plate of each of the arriving and departing vehicles. First, the license plate is located in the image. Then the optical character recognition (OCR) solutions are used to recognize the number of the license plate. Some examples are shown in Figure 12.11.

Figure 12.11 Examples of License plate recognition result

12.5.3 Audio and Speech Analytics

The audio and speech analysis system is an invaluable resource for labeling relevant and interesting audio events. Such technology is useful for applications where rapid search and indexing capabilities are paramount for the early identification of consistent routines, new trends or isolated events [6]. For businesses, the rapid location of such audio events can present new opportunities and trends while exposing potential threats that would otherwise remain concealed.

Figure 12.12 Structure of the acoustic event extraction system

The audio and speech analysis system exploits multiple independent acoustic event detectors. The advantage of using independent detectors is that a change made (or a critical failure) for a detector will not influence other operational audio event detectors. Figure 12.12 presents the basic structure of the acoustic processing system tailored for analyzing audio from military scenarios. In this system the audio is preprocessed with a bulk audio processing component which down-samples the audio and removes uninteresting audio such as silence. This has the added benefit of reducing the processing load on the following processing components. This processed audio is passed to the impulse, vehicle and speech analytics elements accordingly. The speech analytics block includes both speech recognition and language identification. Other technologies that could be included are speaker recognition, emotion detection, dialect detection and background noise profiling. Each of the technologies used in this system will be briefly described.

Impulse Type Event Detection: The task here is to detect acoustic events that occur as a result of a projectile or bomb explosion. These types of acoustic events are impulsive in nature and generally have high signal energy when in close proximity. A two phase approach was implemented to detect such events. The first phase used Hidden Markov Models (HMMs) trained on cepstral and energy based features in a similar manner to [32] to detect the candidate events. These candidates were then submitted to a second phase test whereby events that were below a specified power level were disregarded.

Acoustic based Vehicle Presence Detection: Acoustic analysis may be used to determine the presence of different types of utility vehicles; both land based and airborne. The system established for this evaluation was designed to detect land based vehicles. A car, for example, generates sound from the combustion engine, moving parts on the vehicle and the tires. A HMM structure trained on cepstral based features was established to detect the presence of a vehicle. At the time of the experiment, it is important to note that no noise or channel compensation was applied to these features. Preliminary studies examined the misclassification of vehicles (or when the system was making the classification errors). It was observed that some types of diesel engine vehicles were being confused with other diesel engine vehicle types. Similarly, non-diesel (car fuel) engines were misclassified as other vehicles of the non-diesel type.

Speech Analytics: The **speech recognition** block is comprised of two main parts; namely a speech segmentation block followed by a state-of-the-art speech recognition engine. The speech recognition engine is based on a HMM framework. Some of the significant developments for improved speech recognition performance include Subspace Precision and Mean (SPAM) models [19] and fMPE [45]. For this particular evaluation, the speech recognition system was evaluated within the framework of keyword spotting. (A keyword spotting system would have a considerable advantage.) The **language identification** component utilized speech segmentation boundaries identified by the speech recognition preprocessing engine. The identified segment was phonetically transcribed and later classified according to a binary tree classifier structure [5, 37]. The language identification system was trained on seven different languages.

12.5.4 Multimodal Integration

After the processes of image classification, object detection and speech analysis, the processes results are integrated to achieve more accurate results. To integrate all the different features and results, several approaches such as multi-layer HMMs method and rule-based combination of a hierarchy of classifiers can be used. For example, without multimodal integration, the license plate recognition is running on all the input images. By combining with the image classification results, the license plate recognition is skipped for indoor images. This process reduces the computation cost and decrease the number of false positives.

12.6 INTERFACE: ANALYSIS AND AUTHORING/ REPORTING

We have implemented several navigation interfaces that enable end users to search, retrieve, and filter events and data of interest to them, either from their own missions/experiences or those of other people. These interfaces are essentially multi-dimensional and allow navigation based on space, time, or events of interest. Our aim has been not only to retrieve events of interest but to allow users to re-live relevant experiences, especially those shared by others.

Figure 12.13 Screenshot of a multi-dimensional event navigation and filtering tool

Figure 12.13 shows a screen shot of the multi-dimensional navigation tool for retrieving, searching, and filtering the captured and analyzed data in the electronic chronicle. The interface consists of a document/data summary view (the left column in Figure 12.13), a set of browsing controls (top right of Figure 12.13), and a map (bottom right of Figure 12.13). The document summary shows a thumbnail view of the document/data, if available, and metadata about the document/data. A document (captured from the user's PC) and several images (captured in the field) are seen in Figure 12.14. All these have time stamps associated with them and images also have GPS locations associated with them. The pull down on a document gives further information associated with the document. For example, the image numbered three in Figure 12.14 shows further information including the author, date of creation etc. Notice that the concept associated with this image is "outdoors", obtained by the analytics discussed in Section 12.5.1. Also notice that the image has associated "annotations". In this case there is an audio file, which can be played by clicking on the speaker icon. The transcription of the speech in the audio is also seen ("large garage one door open"). This is obtained by analyzing the recorded audio as discussed in Section 12.5.2.

The browsing controls on the right enable the user to filter the vast quantities of captured data based on their interest. The user can navigate the data by space, time, and a variety of meta-data including image “concepts”, presence/number of faces, type of document, keywords from speech, audio events, author/creator of the data etc. The tool allows multiple filtering criteria to be combined. When location information is available the data is also shown on a map. In Figure 12.13, the geographic location of documents is shown on the map with the number associated with the particular document. If the user is interested in browsing data/events that occurred only in a particular geographic area, they can simply draw a box to indicate their area of interest on the map, which results in displaying only data/events from that area.

The interface in Figure 12.13 is implemented within a web browser running on the user’s machine. The web page requests XML data via an HTTP call to the database server. The interface is generated from the XML via XSLT and Javascript. The XSLT stylesheet processes the XML to produce the HTML for the visual aspects of the interface. It also generates embedded Javascript calls to pass information gleaned from the XSLT parser to the client-side Javascript logic. This allows us to avoid having to parse the XML data ourselves. This interface approach has been inspired by [59].

We have started running data collection and navigation experiments with our system, initially with users collecting data within a 20 mile radius around our research labs. For illustration purposes, we will use data collected in one such trip where a user made a round trip from our lab location in Hawthorne, New York, and our other lab located in Yorktown, New York. The user collected images, GPS, and audio all along the way (the car was driven by a colleague). The user and colleagues surveyed the Yorktown facility, both outdoors and indoors, had lunch there and returned to Hawthorne. In addition to the data collected during the trip, the user also input some trip planning documents and presentations into the database. The following examples show some of the different ways of browsing and retrieving data of interest from this trip.

Figure 12.14 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 12.14 shows one view of the data without any filters applied. In this view, the user sees a timeline and events marked along the timeline to indicate when images were taken. The corresponding locations are also shown in the map (marked with numbers overlaid on the map – along the highway in the middle of Figure 12.14). On the left are a variety of image concepts and keywords detected from the user's speech annotations as well as from his documents and presentations. All the images captured are seen in this view – the user simply has to move his cursor along the timeline.

Figure 12.15 Example result: User filters data in Figure 12.12 to view only images with "faces". This view also shows the sub-images of detected faces in each original image.

Figure 12.16 Example result: User further filters the images with "faces" to view only those labeled as "indoors"

Figure 12.15 shows the result of a "face" filter applied to this data – i.e. only the user chooses to see only those images which had faces in them. Figure 12.15 shows a subset of the images in sequence. Notice how only the face images from Figure 12.14 appear in Figure 12.15 and how the images of cars etc. are eliminated. The view in Figure 12.15 shows the original captured image as well as sub-images for each face detected in that image. Figure 12.16 shows a sequential sample of the result when the user combines two image concepts – "faces" and "indoors". "Indoors" indicates that the image classifier categorized the image as one taken indoors, and "faces" indicates that the image had at least one face detected in it.

Figure 12.17 Example result: A second user selects the area of the map that is of interest to him, by drawing a rectangle. The spatial filtering results in only the data obtained in that area being highlighted along the time line.

Figure 12.18 Example result: Noticing the presence of the word "garage" in the speech annotation for the image labeled 2 in Figure 12.15, the user searches for other images with the word "garage" in the annotation. The user notices from the images and the annotations that there was a

significant change in the garage in the 80 minutes between the two images and decides to investigate further

Figures 12.17 and 12.18 illustrate the case where a second user combines spatial and speech annotations for navigating the data captured by a first user. The new user first marks out an area in the map corresponding to a portion of the Yorktown facility, which is of interest to this user. Only the data corresponding to this area is highlighted with a rectangle on the map in Figure 12.18. Similarly, only this data appears normally on the time line while the rest is grayed out.

From Figure 12.17, the user notices that there is a picture of a garage in the area with a corresponding audio annotation. The user listens to this annotation, recorded at 16.31 on 05-24-2005, which talks about seeing a large garage with one door open. To see if there were any other interesting pictures of this garage, the user further searches by the keyword “garage” in the speech annotation. This leads to the picture shown in Figure 12.18 from the same area of the map taken at 17.54 on the same day with a corresponding audio annotation that talks about how two doors are open and there is a red car parked in front of the garage, but with no people in the car or in the garage. The user decides to investigate further.

Experiential Interface

This filtering of discrete events represents but one way to navigate the chronicled events. Figure 12.19 shows a very different eChronicle navigation interface. This is another multi-dimensional interface that enables the authorized user to review, analyze, and even re-live the mission experiences captured by themselves or more importantly, by other people who share their experiences with the user. This multi-dimensional interface consists of a map viewer, event-annotated timelines, media windows, an event browser, and a report authoring/review tool.

Figure 12.19 A second eChronicle navigation system that allows the viewer to browse, replay, retrieve, and analyze mission experiences.

The map viewer provides a map of the area relevant to events of interest and displays path overlays, and event annotations. Paths and annotations get updated on the map as a user browses the timeline. The browsable timeline shows selected detected events with associated confidence levels. Browsable media windows show video, audio, and high resolution images. The media windows get automatically updated as a user browses the timeline – enabling the user to rapidly scan a captured experience. Media can be played at any time, allowing the user to relive the experience. The event browser shows a list of events and concepts that are automatically detected from captured data. User can select events or event-combinations to view on the timeline and map, thus enabling them to rapidly zero-in on times and spaces of interest and review the experiences in the spatio-temporal zone of interest to them. Finally, the interface also enables users to author reports by selecting portions of media/events/timeline/map and associating an annotation/report with these selected portions. The author or other users can then rapidly view these reports through the multi-dimensional interface, again based on space, time, and events of interest.

12.7 EXPERIMENTS AND SYSTEM EVALUATION

We conducted targeted evaluations of the eChronicling system in a specific domain – soldiers performing patrols of urban areas. While this by no means covers the gamut of application scenarios of the systems, it allowed us to evaluate the system in at least one specific situation. The evaluations were performed by an independent evaluation team and consisted of two types of evaluations: i) elemental and ii) vignette. The elemental evaluations involved evaluations of the specific elemental event detection capabilities. The vignette evaluations judged the system as a whole by having targeted users perform missions in simulated settings. Both people who performed missions and people who received mission reports (without having been on actual missions) evaluated the value such an eChronicling system provides.

The vignette evaluations were more qualitative in nature and will not be reported here in detail. In summary, the system and the interface illustrated in Figure 12.18 were given high ratings in the vignette evaluations and seen as considerably enhancing the mission reporting, recall, and intelligence analysis. The multi-dimensional browsing capability by space, time, and events was received very positively. The ability to easily relive a relevant experience was of special interest. This capability allowed people to get their own perspective on different aspects of a mission, independent of automatic event detection and already added great value. Another feature that was found to be of great value is the ability to select events of interest and display them with associated confidence values. This, according to users was of immense value compared to filtering based on a preset threshold. This allowed users to use their own judgment in deciding which events to explore at what levels of confidence and not be entirely dependent on machine-made decisions. Also, this approach gave visibility to clusters of events on the timeline or on the map and allowed users to find spatio-temporal zones of interesting events.

While, these qualitative results were the most important outcome of the evaluations, we also describe below the quantitative evaluations on specific elemental capabilities. These evaluations were certainly useful in estimating the state of event detection capabilities, but also had limitations in the way they were conducted. Firstly, the data set was very limited (for example, only a total of 25 images were used to evaluate the image tagging and object detection capabilities) and was not statistically rich enough to draw a conclusion on the event detection capabilities. Secondly, the evaluation required a binary answer on the presence or absence of an event/concept and hence was driven by hard thresholds used in the analytics. This did not account for the ability to show confidence values for events to humans in a visual event browsing scenario as described earlier. With this preamble, we go into the specifics of the elemental evaluation results.

12.7.1 Image Tagging Performance and Observations

The automatic semantic concept tagger was trained from three distinct data sets for the following semantic concepts: *Outdoors, Indoors, Vegetation, Vehicle_Civil*

(Car, Truck, Bus), Vehicle_Military, Person (Soldier, Locals), Weapon, Building. The three data sets had representative images in the following three context: (1) news videos data used in the TRECVID benchmark, (2) personal photo collection, (3) military data collected both on and off field (web). The depiction of the same concept in these different data sets could drastically be different. The reason why only the military related data set was not used, was due to the few images collected on field and the questionable quality of the content harvested off the web. The other two data sets were used to enhance the visual examples for the given concepts. Note that not all concepts were modeled using all three different sets. For example, concepts *Indoors* and *Outdoors*, were only modeled using the news video data set, whereas three different sets of models for the vehicle related concepts one from each of the three different data sets.

Each training and testing data were represented by the color correlogram and color moments features, and SVM classifier with RBF kernels were trained for each of the concepts using each of the representations. Fusion across feature models was done using simple yet good performing averaging mechanism, and fusion across data sets fro same concept was also done using averaging. Note that in order to address the issue of visual polysemy, in addition to fusing models of the same concept across data sets, we also fused the models of sub-categories of concepts. For example, for concept *Vehicle* the results of concepts *Truck, Car,* and *Bus* were fused together to provide the confidence value for *Vehicle*.

Figure 12.20. Tags associated with the left-hand side image: Car, Outdoors, Soldier, Vegetation, Vehicle_Civil; Tags associated with the middle image: Indoors, Soldier; Tags associated with the right-hand side image: Building, Outdoors, Soldier.

For testing 25 images were provided for automatic tagging. Figure 12.20 shows three such image and the tags automatically associated to them. The image on the left is being tagged without any mistake, neither false positive nor false negative. In the middle image the person is wrongly identified as soldier (although it is likely) and in the image on the right vehicles have been missed and soldier and building are wrongly tagged (again absence of building is questionable). Table 12.1 shows the results of the evaluation. According to the table, the most challenging category of concepts appeared to be *Weapon*, this is both due to the various forms of the appearance of the weapon and the context in which weapon occurs. Note that our approach does not locate the manifestation of the concept in the image, neither it counts the number of occurrences of the concept, but only predicts the degree of its presence or absence in the image. The results could also be attributed to the quality and quantity of the training data provided in the three different sets. The concept, however, appeared abundantly in the vignette evaluations. So, the threshold for the binary decision seemed to have been a major factor in the elemental test. For concepts such as *Outdoors*, which is a highly recurrent concept, there were plenty of positive examples provided in the news data set and therefore a good model was obtained, whereas for concept *Weapon*, there were not enough training data. Concepts such as *Vehicle*, which had fair amount of training data in both the news video set and the personal photo collection one, performed poorly due to the different depiction of such concept in the military context.

Table 12.1. Image tagging performance table. Note that results are not reported for all tags.

Concept	Outdoors	Indoors	Building	Vegetation	People	Weapon	Vehicle
% Pos ID over all instances of presence	95.2%	100%	86%	100%	29%	0%	50%
%Miss over all instances of presence	4.7%	0%	14%	0%	71%	100%	50%
% Pos ID over all ID'd	91%	80%	100%	67%	40%	0%	57%
% Neg ID over all ID'd	9%	20%	0%	33%	60%	0%	43%

In the image elemental test, the face detection and license plate detection were evaluated. The test images were captured at 25 viewpoints by inserting the following factors in images: distance, background clutter, occlusion and angle of view. The face detection and LPR results are presented in Table 12.2. The failures of the face detection are mainly caused by small face size, occlusions, and lighting changes. The failures of the license plate recognition are mainly caused by small license plate size, the bad viewpoints, and non-US license plates.

Table 12.2 Face detection and license plate recognition elemental results

	Face	**License Plate**
Ground Truth	43	19
Correctly Detected	10	2
False Positives	4	2
Detection rate	23.3%	10.5%

We now present the audio analytics results for the elemental evaluation data set. For this set we evaluated the performance of impulse detection, land vehicle detection, language identification, keyword spotting and speech detection. To simplify scoring across all audio domains, we proposed a single consistent metric that has a range from 0 to 100% performance. This figure of merit (FOM) is calculated as the number of correct hits divided by the total number of event labels (or the summation of the total hits, misses, false alarms and substitutions).

$$\text{FOM} = \frac{\#\,\text{Hits}}{\#\,(\text{Hits} + \text{Misses} + \text{False Alarms} + \text{Subst})} \times 100\%$$

In this evaluation, each detection system is evaluated in isolation and the results are given in Table 12.3.

Table 12.3 Audio analysis elemental results

Analytics Type Detected	Headset Microphone FOM
Impulse	37%
Vehicle	3%
Language	45%
Keyword	23%
Speech	68%

The results show promise considering the audio environment is relatively harsh. Of special mention is the impulse detection result. For the impulse detection, a two phase approach was used; a first phase event location pass followed by a second phase that checks if a minimum average energy threshold was reached for the audio event. This reduced the number of false alarms from 53 to one. (The single false alarm was the starting whistle for the session.) The minimum energy criterion improved the FOM from 12% to 37%. As a consequence of the second pass a number of low power impulse audio candidates identified from phase 1 were rejected in phase 2.

The vehicle detection system performed poorly. A possible explanation may be mismatch between the audio data used to create the models and the audio actually recorded in the elemental evaluation. The training audio was derived by recording cars on closed asphalt pavements while the evaluation audio was staged on gravel tracks. To perform better vehicle detection, it would be useful to examine more robust classifier features and to train the system on diverse scenarios.

The language identification system performed well under the recording conditions. The system was trained to differentiate between seven languages. The language identification setup classified the audio segments identified by the acoustic segmentation block produced as a bi-product of the speech recognition engine. It is interesting to note that for the vignette data, the language identification tool correctly located foreign language speech in background music on multiple occasions.

Keyword spotting performed reasonably considering that it is based on a large vocabulary continuous speech recognition (LVCSR) system. Although the hit rate was relatively low for the keyword spotting configuration, there was only one false alarm identified for nearly 320 spoken keywords. The benefit is that if the system identifies a hit, it is highly likely the word event occurred.

The speech detection component performed sufficiently. It is essential that the speech detection and segmentation element performs reliably because it is the first component in the line-up to parse the audio. Errors introduced here are propagated to the follow-on speech analytics. Various delayed decision techniques may be introduced to minimize this effect.

12.8 CONCLUSIONS AND FUTURE WORK

This paper presented an overview of our electronic chronicling system and architecture that enables mobile and pervasive data capture, assimilation, and retrieval. The main contribution of this work lies in developing an end-to-end architecture unifying a variety of sensors, logging software on PC's, analytic engines, data management, and navigation tools. By building such a system with components that are state-of-the-art, we are able to explore unprecedented capabilities that are greater than the sum of the parts. We view this still as the early stage for research on such pervasive electronic chronicling systems. The initial results presented here show promise that this approach could indeed impact a variety of applications involving field operations, mobile workforce management, and situation analysis.

Multiple conclusions have been derived from the image tagging unit of the reported electronic chronicling system. First, those concepts that are related to specific objects could not be supported adequately using the reported approach to image tagging. We plan to extend generic approach for semantic concept modeling to a palette of approaches specifically designed for different classes of semantic concepts, for example a salient region based approach could be more suitable for modeling object related concepts.

In addition, one difficulty we faced in designing the concept model, was manually annotating the data training data sets prior to concept tagger training. It is a very tedious task to obtain the ground truth data for the concepts of interest. We are planning to devise methodologies to leverage socially obtained tags via the specific applications offered on the web.

Audio analytics provides substantial information toward an eChronicling system.

Follow-on work will involve finer grained classification of the audio events. Some examples include the identification of the type of gunshot or vehicle.

Our future work will continue to extend the individual analytic capabilities and corresponding search and navigation tools. Future work will focus on the long-term data association based on multimodal integration.

Finally, there is room for significant research on the underlying data and event representation models, and techniques for event and activity mining.

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.

References

[1] B. Adams, A. Amir, G. Iyengar, C.-Y. Lin, M. Naphade, C. Neti, J.R. Smith, “Semantic Indexing of Multimedia Content Using Visual, Audio and Text Cues”, EURASIP Journal of Applied Signal Processing, Special Issue, February 2003.

[2] Blanz and Vetter, “Face recognition based on fitting 3D morphable model,” IEEE PAMI, vol. 25, no. 9, pp. 1063–1074, Sept. 2003.

[3] A. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound, MIT Press: Cambridge, Massachusetts, USA, 1994.

[4] V. Bush As we may think. Atlantic Monthly, 176 (1), pp. 641 - 649, January 1945.

[5] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, P. Torres-Carrasquillo, “Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 Language Recognition Evaluation”, IEEE Odyssey Speaker and Language Recognition Workshop, 2006.

[6] S. Caskey, U. Chaudhari, E. Epstein, R. Florian, J. S. McCarley, M. Omar, G. Ramaswamy, S. Roukos, T. Ward "Infrastructure and Systems for Adaptive Speech and Text Analytics", 2005 International Conference on Intelligence Analysis, 2-6 May, 2005, McLean, VA.

[7] C. Collet, G. Vargas and H.G. Ribeiro, Towards a Semantic Event Service for Distributed Active Database Applications, Proc. of 9th International Conference on Database and Expert Systems Applications (DEXA'98), Vienna, Austria,Aug. 1998.

[8] R. Collins, et al. `A system for video surveillance and monitoring', VSAM Final Report, Technical Report, CMURI-TR-00-12, May 2000

[9] Combat Zones That See, U.S. Government DARPA Project.

[10] T. Coupaye, C. L. Roncancio, and C. Bruley. A Visualization Service for Event-Based Systems. Proceedings of 15emes Journees Bases de Donnees Avancees, Toronto, Canada, August 2003.

[11] S. Dumais, E. Cutrell, JJ Cadiz, G. Hancke, R. Sarin, D.C. Robbins. Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. Proceedings of ACM SIGIR '03, Toronto, Canada, August 2003.

[12] Eclipse Platform Technical Overview, Object Technology International, Inc. February 2003, http://www.eclipse.org/

[13] D. Ferrucci and A. Lally, Building an example application with the Unstructured Information Management Architecture, IBM Systems Journal 43, No. 3, 455-475 (2004).

[14] L. Fiege, G. Muhl, and F. C. Gartner. A modular approach to build structured event-based systems. In Proceedings of the ACM Symposium on Applied Computing (SAC'02), pages 385--392, Madrid, Spain, 2002.

[15] E. T. Freeman and S. J. Fertig. Lifestreams: Organizing your electronic life, AAAI Fall Symposium; AI Applications in Knowledge Navigation and Retrieval, Nov. 1995.

[16] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. MyLifeBits: Fulfilling the Memex Vision. ACM Multimedia '02, pp. 235 - 238.

[17] J. Gemmell and H. Sundaram. Proceedings of the First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE 2004), New York, October 2004.

[18] J. Gemmell. ACM SIGMM CARPE home page. http://www.sigmm.org/jgemmell/CARPE.

[19] V. Goel, S. Axelrod, R. Gopinath, P Olsen, and K. Visweswariah, “Discriminative Estimation of Subspace Precision and Mean (SPAM) Models”, in proc. of ISCA Eurospeech, pp. 2617-2620, 2003.

[20] M. Greiffenhagen, D. Comaniciu, H. Niemann, V. Ramesh, `Design,analysis and engineering of video monitoring systems: an approach and case study', The Proceedings of the IEEE, vol. 89, no. 10, pp. 1498-1517, October

[21] S. Guven, M. Podlaseck, and G. Pingali. “PICASSO: Pervasive Chronicling, Access, Search, and Sharing for Organizations,” IEEE International Conference on Pervasive Computing, Hawaii, March 2005.

[22] Hampapur, A.; Brown, L.; Connell, J.; Ekin, A.; Haas, N.; Lu, M.; Merkl, H.; Pankanti, S. “Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking,” IEEE Signal Processing, Vol. 22, Issue 2, March 2005, Page(s): 38- 51.

[23] Haritaoglu, “Harwood and Davis, W4: Real time surveillance of people and their activities,” IEEE Trans. Pattern Anal. Machine Intell., vol. 22, no. 8, pp. 809–830, Aug. 2000.

[24] Hori and K. Aizawa. Capturing Life Log and Retrieval based on Context. In Proc. IEEE ICME 2004, June 2004.

[25] T. Horprasert, D. Harwood, and L. Davis, “A statistical approach for real-time robust background subtraction and shadow detection,” in Proc. IEEE Frame-Rate Workshop, Kerkyra, Greece, 21st Sept. 1999.

[26] Human ID at a Distance, U.S Government, DARPA Project.

[27] D. Huynh, D. Karger, and D. Quan. “Haystack: A platform for creating, organizing and visualizing information using RDF”, The Twelfth International World Wide Web Conference**,** 20-24 May 2003, Budapest, HUNGARY

[28] International Workshop on Distributed Event-Based Systems (DEBS’05). Columbus, Ohio, June 10, 2005. http://www.cs.queensu.ca/~dingel/debs05/index.html

[29] R. Jain. Media Vision: Multimedia Electronic Chronicles. IEEE Multimedia, July 2003.

[30] M. Jones and P. Viola, “Fast Multi-view Face Detection,” CVPR, 2003

[31] P. Kim, M. Podlaseck, and G. Pingali. Personal Chronicling Tools for Enhancing Information Archival and Collaboration in Enterprises. ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE 2004), New York, October 2004.

[32] T. Kristjansson, B. Frey and T. Huang, “Event Coupled Hidden Markov Models”, in proc. of the IEEE International Conference on Multimedia and Expo (ICME), 2000.

[33] A. Levas, G. Pingali, M. Podlaseck, J. W. Murdock. Exploiting Pervasive Enterprise Chronicles using Unstructured Information Management. In IEEE International Conference on Pervasive Services (ICPS 2005), July 2005.

[34] Alan J. Lipton, Craig H. Heartwell, Dr Niels Haering, and Donald Madden, Critical Asset Protection, Perimeter Monitoring, and Threat Detection Using Automated Video Surveillance, white paper, ObjectVideo.

[35] S. Z. Li, L. Zhu, Z. Q. Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical Learning of Multi-view Face Detection," ECCV, 2002.

[36] C. Moore, "Diving into Data," InfoWorld (October 25, 2002), http://www.infoworld.com/article/02/10/25/ 021028feundata_1.html.

[37] J. Navratil, "Recent advances in phonotactic language recognition using binary-decision trees", in proc. of ISCA Interspeech, 2006.

[38] NSF Workshop on Personal Information Management. Jan 27-29, 2005. http://pim.ischool.washington.edu/

[39] J. Phillips, P. Grother, R. Micheals, D.M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002 P," in Proc. IEEE Int. Workshop Analysis and Modeling of Faces and Gestures (AMFG'03).

[40] G. Pingali and R. Jain. Electronic Chronicles: Empowering Individuals, Groups, and Organizations. In the Proceedings of the IEEE International Conference on Multimedia and Expo, Amsterdam, July 2005.

[41] G. Pingali, Y. Jean, A. Opalach, and I. Carlbom. LucentVision: Converting Real World Events into Multimedia Experiences. In the Proceedings of the IEEE International Conference on Multimedia and Expo, New York, July 2000

[42] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. Lifeline: Visualizing Personal Histories. Proceedings of ACM Conference on Human Computer Interaction (CHI'96), Vancouver, USA, ACM Press 221-227, 1996.

[43] John C. Platt, “Probabilities for Support Vector”, Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans, eds., pp. 61-74, MIT Press, (1999).

[44] G. Potamianos, et al, “Recent Advances in the Automatic Recognition of Audio-Visual Speech”, in proc. of the IEEE, Vol. 91, No. 9, pp 1-18, 2003.

[45] D. Povey, et al, “fMPE: Discriminatively Trained Features for Speech Recognition”, in proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, vol 1, pp. 961-964, 2005.

[46] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, Prentice-Hall: Upper Saddle River, NJ, USA, 1993.

[47] J. Rekimoto. Timescape: A time machine for the desktop environment.

[48] B. Rhodes and T. Starner. Remembrance Agent: A continuously running automated information retrieval system. Proceedings of the First International Conference on the Practical Application of Intelligent Systems and Multi Agent Technology. London, April 1996.

[49] H. Rowley, S. Baluja, and T. Kanade, “Neural Network-based Face Detection,” PAMI, Vol. 20, p22-38, 1998.

[50] Thad Starner. Wearable Computing and Context Awareness. Ph. D. Thesis, MIT Media Lab, April 30, 1999.

[51] G. Stauffer, “Learning patterns of activity using real-time tracking,” IEEE Trans. Pattern Anal. Machine Intell., vol. 22, no. 8, pp. 747–757, Aug. 2000.

[52] Remagnino, Jones, Paragios, and Regazzoni, Video Based Surveillance Systems Computer Vision and Distributed Processing. Norwell, MA: Kluwer , 2002.

[53] H. Schneiderman and T. kanade, “A Statistical Method for 3D Object Detection Applied to Faces and Cars,” ICCV, 2000.

[54] SeeCar License Plate Recognition, Hi-Tech Solutions. http://www.htsol.com/.

[55] James G. Shanahan and Norbert Roma, "Boosting support vector machines for text classification through parameter-free threshold relaxation". CIKM,pp. 247-254, 2003.

[56] K. Sung and T. Poggio, “Example-based Learning for View-based Face Detection,” PAMI, Vol. 20, p39-51, 1998.

[57] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” CVPR, 2001.

[58] Vladimir N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.

[59] SIMILE project's Longwell (http://simile.mit.edu/longwell/index.html).

[60] VACE: Video Analysis and Content Exploitation [YLOnline]. Available: http://www.ic-arda.org/InfoExploit/vace/

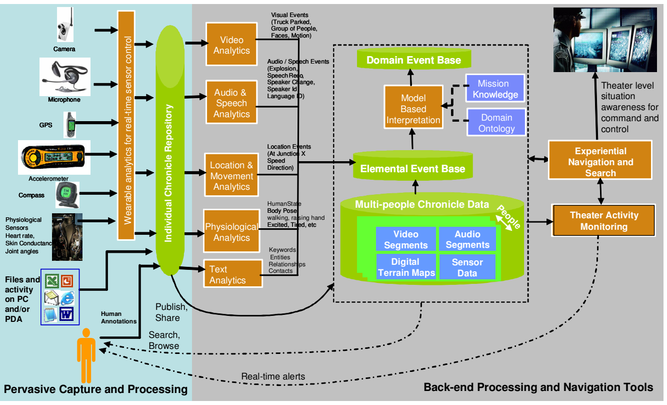

Camera
Microphone
GPS
Accelerometer
Compass
Physiological Sensors Heart rate, Skin Conductanc Joint angles
Files and activity on PC and/or PDA
Human Annotations
Wearable analytics for real-time sensor control
Individual Chronicle Repository
Publish, Share
Search, Browse
Pervasive Capture and Processing
Video Analytics
Audio & Speech Analytics
Location & Movement Analytics
Physiological Analytics
Text Analytics
Visual Events (Truck Parked, Group of People, Faces, Motion)
Audio / Speech Events (Explosion, Speech Reco, Speaker Change, Speaker Id Language ID)
Location Events (At Junction X Speed Direction)
HumanState Body Pose walking, raising hand Excited, Tired, etc
Keywords Entities Relationships Contacts
Domain Event Base
Model Based Interpretation
Mission Knowledge
Domain Ontology
Elemental Event Base
Multi-people Chronicle Data
People
Video Segments
Audio Segments
Digital Terrain Maps
Sensor Data
Theater level situation awareness for command and control
Experiential Navigation and Search
Theater Activity Monitoring
Real-time alerts
Back-end Processing and Navigation Tools

Figure 12.1 Overview of the end-to-end electronic chronicling system

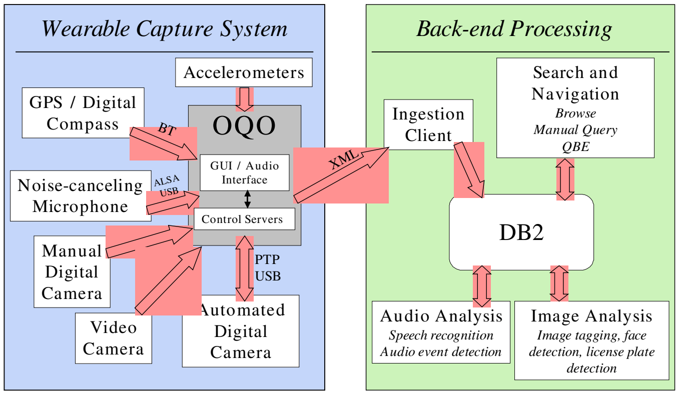


Figure 12.2 Simplified view of the eChronicling system

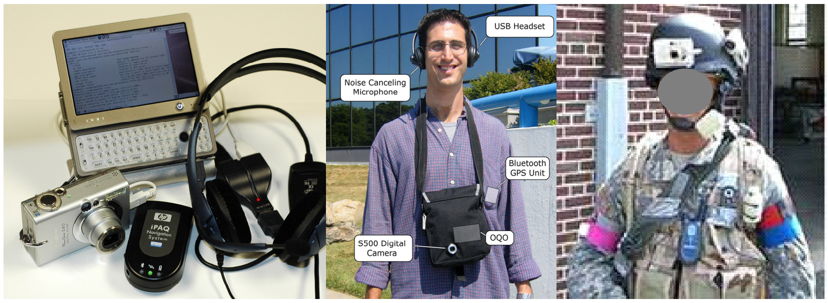


Figure 12.3 (a) Example of Devices used for capture; (b) one example of a user wearing a capture system consisting of devices in (a) (c) another example of a user wearing an extended set of capture devices

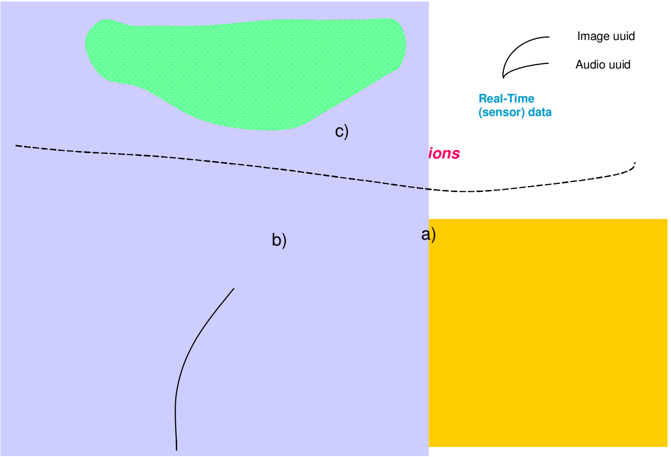


Figure 12.4 Overview of types of data associated with events and their representation scheme

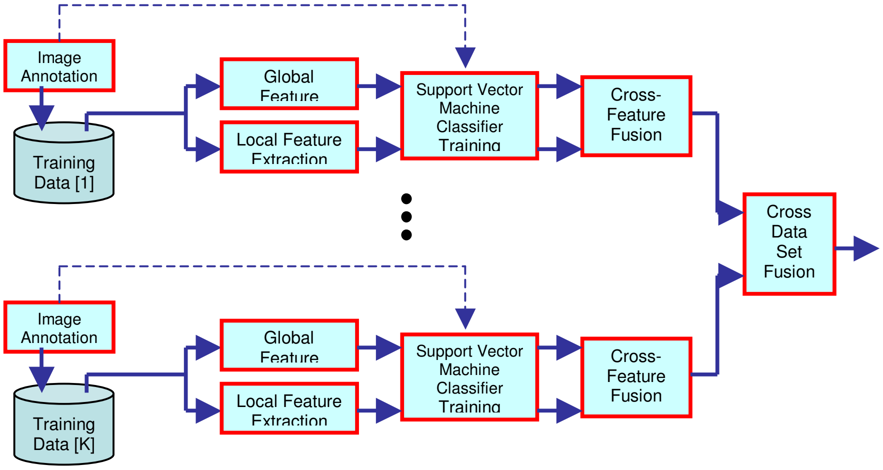


Figure 12.5 Overview of image tagging scheme.

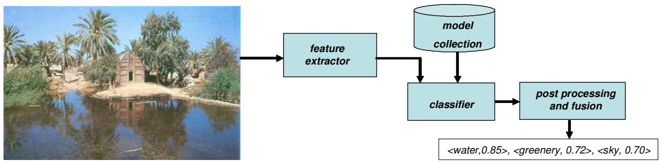

feature extractor
model collection
classifier
post processing and fusion
<water,0.85>, <greenery, 0.72>, <sky, 0.70>

Figure 12.6 Illustration of image tagging.

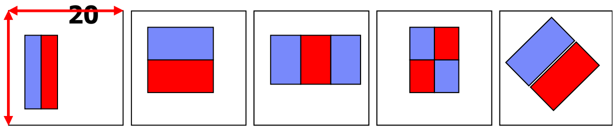
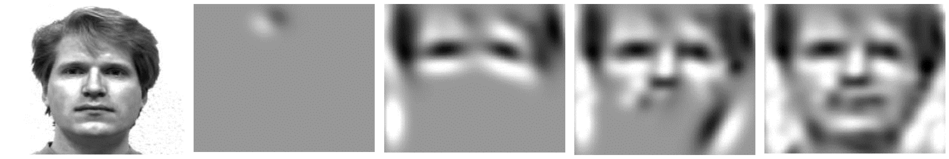


Figure 12.7 Example Harr-like features for face detection



(a) (b) (c) (d) (e)

Figure 12.8 Example optimized wavelet features for face detection. (a) Original image; (b) – (e) Wavelet Representation.

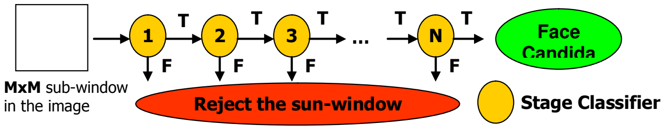


Figure 12.9 Cascade of classifiers for face detection

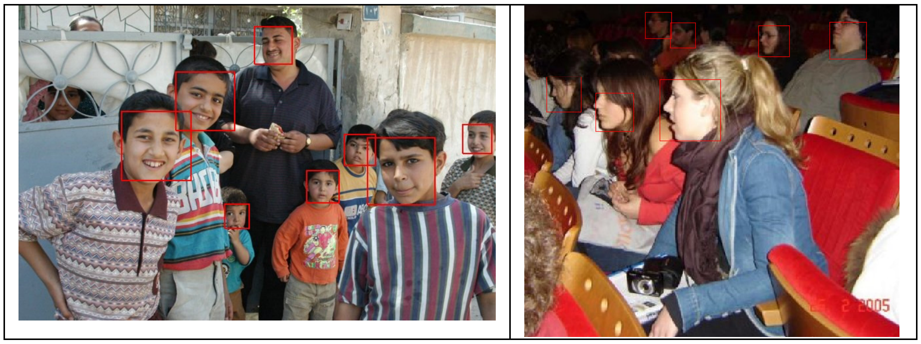


Figure 12.10 Examples of face detection result

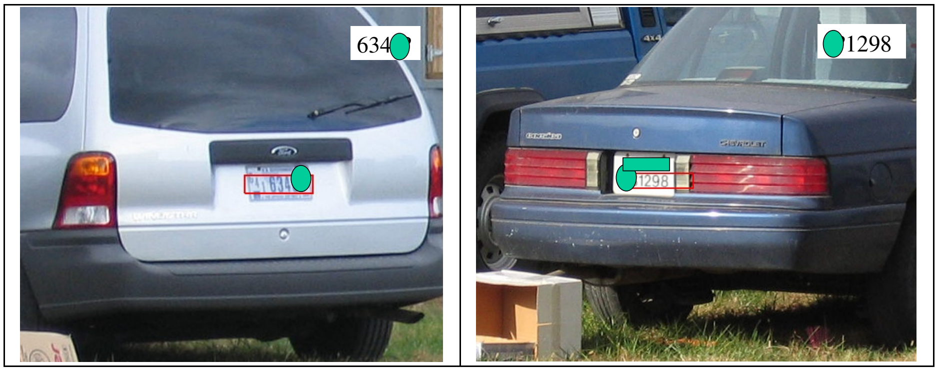


Figure 12.11 Examples of License plate recognition result (part of license plate has been occluded for privacy purpose)

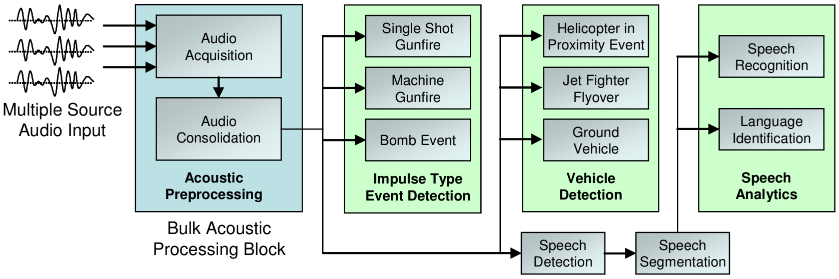
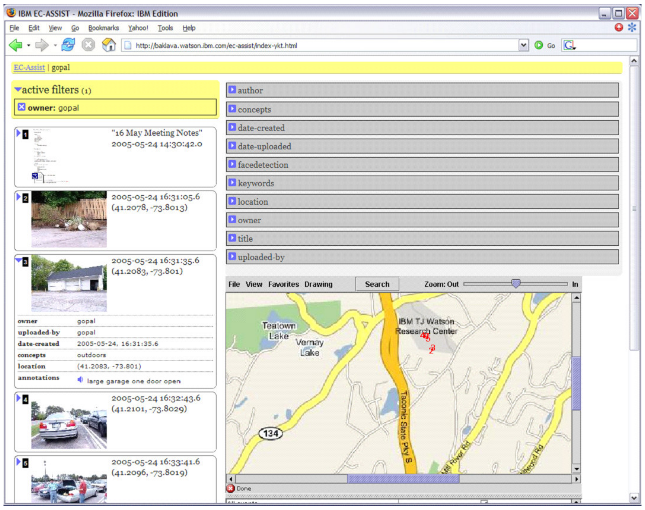


Figure 12.12 Structure of the acoustic event extraction system


IBM EC-ASSIST - Mozilla Firefox: IBM Edition
File Edit View Go Bookmarks Yahoo! Tools Help
http://baklava.watson.ibm.com/ec-assist/index-ykt.html
Go
EC-Assist | gopal
author
concepts
date-created
date-uploaded
facedetection
keywords
location
owner
title
uploaded-by
File View Favorites Drawing Search Zoom: Out In
Teatown Lake
Vernay Lake
IBM TJ Watson Research Center
134
Taconic State Pkwy S
Mill River Rd
Done

active filters (1)

owner: gopal

1 "16 May Meeting Notes" 2005-05-24 14:30:42.0

2 2005-05-24 16:31:05.6 (41.2078, -73.8013)

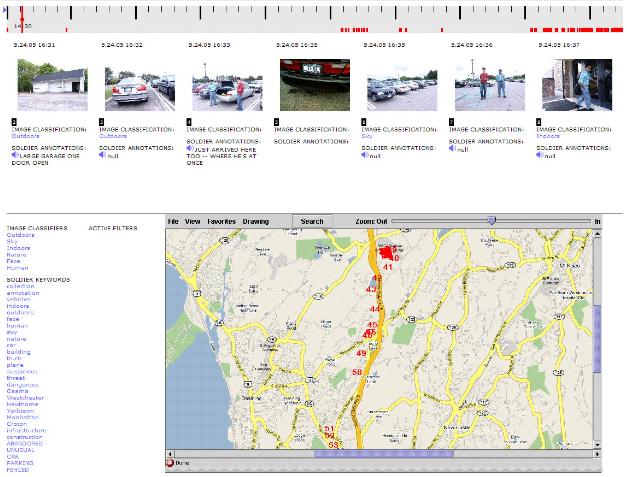
3 2005-05-24 16:31:35.6 (41.2083, -73.801)

owner	gopal
uploaded-by	gopal
date-created	2005-05-24, 16:31:35.6
concepts	outdoors
location	(41.2083, -73.801)
annotations	large garage one door open

4 2005-05-24 16:32:43.6 (41.2101, -73.8029)

5 2005-05-24 16:33:41.6 (41.2096, -73.8019)

Figure 12.13 Screenshot of a multi-dimensional event navigation and filtering tool



5.24.05 16:31 — 2 IMAGE CLASSIFICATION: Outdoors SOLDIER ANNOTATIONS: LARGE GARAGE ONE DOOR OPEN

5.24.05 16:32 — 3 IMAGE CLASSIFICATION: Outdoors SOLDIER ANNOTATIONS: null

5.24.05 16:33 — 4 IMAGE CLASSIFICATION: SOLDIER ANNOTATIONS: JUST ARRIVED HERE TOO -- WHERE HE'S AT ONCE

5.24.05 16:35 — 5 IMAGE CLASSIFICATION: SOLDIER ANNOTATIONS:

5.24.05 16:35 — 6 IMAGE CLASSIFICATION: Sky SOLDIER ANNOTATIONS: null

5.24.05 16:36 — 7 IMAGE CLASSIFICATION: SOLDIER ANNOTATIONS: null

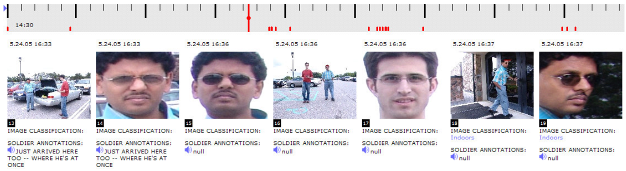
5.24.05 16:37 — 8 IMAGE CLASSIFICATION: Indoors SOLDIER ANNOTATIONS: null

IMAGE CLASSIFIERS
Outdoors
Sky
Indoors
Nature
Face
Human

SOLDIER KEYWORDS
collection
annotation
vehicles
indoors
outdoors
face
human
sky
nature
car
building
truck
plane
suspicious
threat
dangerous
Osama
Westchester
Hawthorne
Yorktown
Manhattan
Croton
infrastructure
construction
ABANDONED
UNUSUAL
CAR
PARKING
FENCED

ACTIVE FILTERS

Figure 12.14 Example result: User views all data from a trip without filtering. Notice the ability to view the data by space, time, concepts, and keywords.



5.24.05 16:33
13
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:
JUST ARRIVED HERE TOO -- WHERE HE'S AT ONCE

5.24.05 16:33
14
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:
JUST ARRIVED HERE TOO -- WHERE HE'S AT ONCE

5.24.05 16:36
15
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:
null

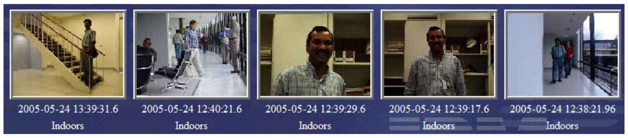
5.24.05 16:36
16
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:
null

5.24.05 16:36
17
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:
null

5.24.05 16:37
18
IMAGE CLASSIFICATION:
Indoors
SOLDIER ANNOTATIONS:
null

5.24.05 16:37
19
IMAGE CLASSIFICATION:
Indoors
SOLDIER ANNOTATIONS:
null

Figure 12.15 Example result: User filters data in Figure 12.12 to view only images with "faces". This view also shows the sub-images of detected faces in each original image.



2005-05-24 13:39:31.6
Indoors

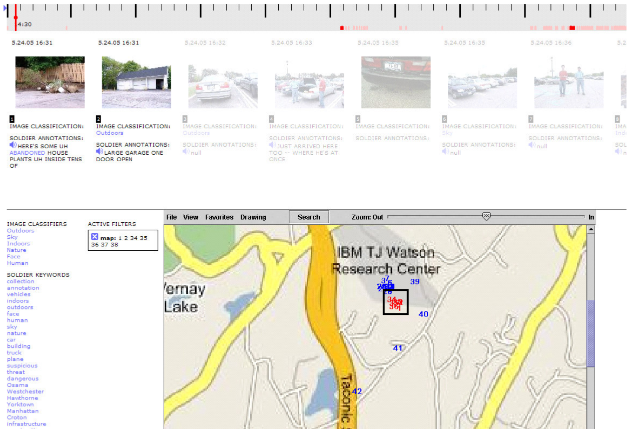
2005-05-24 12:40:21.6
Indoors

2005-05-24 12:39:29.6
Indoors

2005-05-24 12:39:17.6
Indoors

2005-05-24 12:38:21.96
Indoors

Figure 12.16 Example result: User further filters the images with "faces" to view only those labeled as "indoors"



5.24.05 16:31
1
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:
HERE'S SOME UH ABANDONED HOUSE PLANTS UH INSIDE TENS OF

5.24.05 16:31
2
IMAGE CLASSIFICATION:
Outdoors
SOLDIER ANNOTATIONS:
LARGE GARAGE ONE DOOR OPEN

5.24.05 16:32
3
IMAGE CLASSIFICATION:
Outdoors
SOLDIER ANNOTATIONS:
null

5.24.05 16:33
4
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:
JUST ARRIVED HERE TOO -- WHERE HE'S AT ONCE

5.24.05 16:35
5
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:

5.24.05 16:35
6
IMAGE CLASSIFICATION:
Sky
SOLDIER ANNOTATIONS:
null

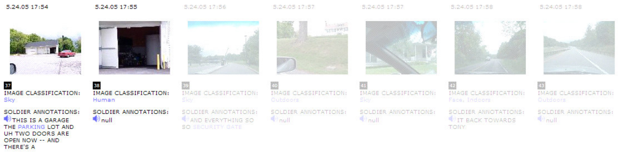
5.24.05 16:36
7
IMAGE CLASSIFICATION:
SOLDIER ANNOTATIONS:
null

IMAGE CLASSIFIERS
Outdoors
Sky
Indoors
Nature
Face
Human

SOLDIER KEYWORDS
collection
annotation
vehicles
indoors
outdoors
face
human
sky
nature
car
building
truck
plane
suspicious
threat
dangerous
Osama
Westchester
Hawthorne
Yorktown
Manhattan
Croton
infrastructure

ACTIVE FILTERS
☒ **map:** 1 2 34 35 36 37 38

Figure 12.17 Example result: A second user selects the area of the map that is of interest to him, by drawing a rectangle. The spatial filtering results in only the data obtained in that area being highlighted along the time line.



5.24.05 17:54

37
IMAGE CLASSIFICATION:
Sky

SOLDIER ANNOTATIONS:
THIS IS A GARAGE THE PARKING LOT AND UH TWO DOORS ARE OPEN NOW -- AND THERE'S A

5.24.05 17:55

38
IMAGE CLASSIFICATION:
Human

SOLDIER ANNOTATIONS:
null

5.24.05 17:56

39
IMAGE CLASSIFICATION:
Sky

SOLDIER ANNOTATIONS:
AND EVERYTHING SO SO SECURITY GATE

5.24.05 17:57

40
IMAGE CLASSIFICATION:
Outdoors

SOLDIER ANNOTATIONS:
null

5.24.05 17:57

41
IMAGE CLASSIFICATION:
Sky

SOLDIER ANNOTATIONS:
null

5.24.05 17:58

42
IMAGE CLASSIFICATION:
Face, Indoors

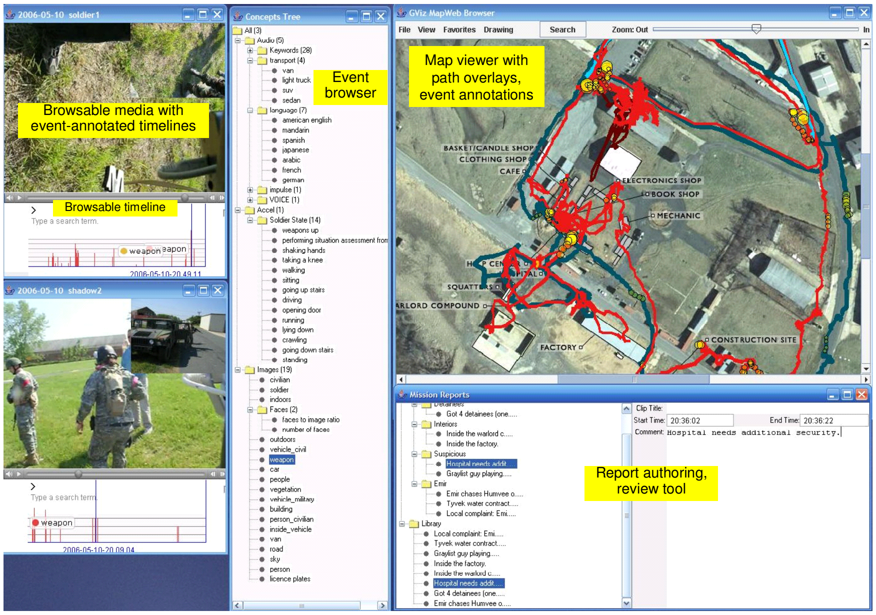
SOLDIER ANNOTATIONS:
IT BACK TOWARDS TONY

5.24.05 17:58

43
IMAGE CLASSIFICATION:
Outdoors

SOLDIER ANNOTATIONS:
null

Figure 12.18 Example result: Noticing the presence of the word "garage" in the speech annotation for the image labeled 2 in Figure 12.15, the user searches for other images with the word "garage" in the annotation. The user notices from the images and the annotations that there was a significant change in the garage in the 80 minutes between the two images and decides to investigate further



Clip Title:
Start Time: 20:36:02 End Time: 20:36:22
Comment: Hospital needs additional security.

Figure 12.19 A second eChronicle navigation system that allows the viewer to browse, replay, retrieve, and analyze mission experiences.

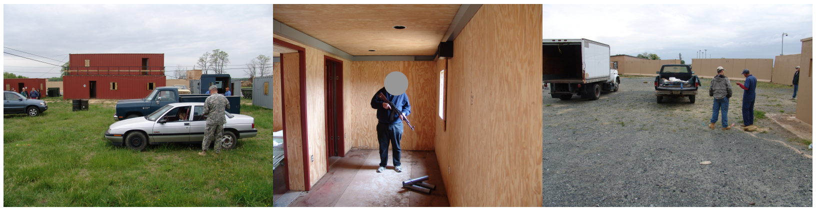


Figure 12.20. Tags associated with the left-hand side image: Car, Outdoors, Soldier, Vegetation, Vehicle_Civil; Tags associated with the middle image: Indoors, Soldier; Tags associated with the right-hand side image: Building, Outdoors, Soldier.