

## A System of Video Information Capture, Indexing and Retrieval for Interpreting Human Activity

Howard D. Wactlar, Michael Christel, Alexander Hauptmann, Scott Stevens  
*School of Computer Science, Carnegie Mellon University*  
*(wactlar, christel, alex, sms)@cs.cmu.edu*  
Ashok Bharucha, M.D.  
*Department of Psychiatry, University of Pittsburgh Medical Center*  
*bharuchaaj@msx.upmc.edu*

### Abstract

*This system creates a manageable information resource that enables more complete and accurate interpretation, assessment and diagnosis of human behavior in constrained physical spaces. Through activity and environmental monitoring, a continuous, voluminous audio and video record is captured. Through work in information extraction, behavior analysis and synthesis, this record is transformed into an information asset whose efficient, secure presentation empowers specialists with greater insights into problems, effectiveness of treatments, and determination of environmental and social influences. Application environments range from nursery schools to nursing homes. The foundation for this work, the Informedia Digital Video Library [1], has demonstrated the successful integration of speech, image, and natural language processing in automatically creating an indexed, searchable multimedia information resource for broadcast-quality video, upon which this system builds.*

### 1. Problem Setting

This research will create a continuously recorded, digital history of human activities in a well-defined environment by capturing all that is heard, seen and experienced in that space. We choose the public areas of a skilled-care nursing home unit as the venue for our observation and analysis. Our research challenge is to transform a voluminous amount of captured video, audio and sensor data into a meaningful information resource that enables more complete and accurate assessment, diagnosis, treatment, and evaluation of behavioral problems for the elderly. The combined aspects of this work in (i) information extraction and retrieval, (ii) behavior recognition, analysis and summarization, and (iii) secure, efficient visual information access, will give geriatric care specialists greater insight into problems, effectiveness of treatments, and environmental and social influences on individual and group patient behavior. Prototype systems for activity monitoring and behavior analysis

are being deployed in wards at local area nursing homes to ultimately be utilized by medical professionals in trials conducted by our project partners from the University of Pittsburgh Medical Center (UPMC), Western Psychiatric Institute and Clinic (WPIC).

A critical element of patient care in the long-term care setting is the ability to gather an accurate account of the patient's physical, behavioral and psychosocial functioning. Given the impaired ability of these patients to report their experiences or to comply with psychometric testing, direct behavioral observation becomes an indispensable aspect of data gathering and treatment planning [2]. Currently in skilled nursing facilities, physicians may see a patient only briefly once a week. Assessment of a patient's progress is often based on inadequate staff reports that due to time and personnel constraints may have resulted from few actual observations of the patient. Manual approaches alone simply cannot keep track of all the residents all the time. This captive, analysis, index and retrieval system, dubbed *CareMedia*, will let physicians query and explore an automatically created profile of patient activity. A continuous audio-visual record, complete with automated techniques to collapse redundancy and highlight key behaviors, will enable better health monitoring and improved care. For example, identification of environmental and other contextual factors that contribute to a behavioral disturbance would offer an opportunity for a non-pharmacological intervention that may reduce the need for psychotropic drugs with their associated toxicities and sedation.

Along with functional and cognitive decline, behavioral disturbances are part of a triad of disability seen in dementia [3], where "agitation" is used to describe all classes of behavioral disturbances. Disturbed behaviors likely require drug treatment. Disturbing behavior may simply be a form of communication. The technology applied here does not endeavor to differentiate between these categories, but rather will present occurrences of certain behaviors to the medical professionals for their interpretation. By analyzing the events leading up to a bout of agitation,

we plan to use information technology to better inform and assist the professionals in their assessments.

It is already feasible to capture complete audio, video and selected sensory data for extended periods of time. Technology is in place to record all of the events in a physical space, estimated to accumulate at only tens of gigabytes per day with terabytes of storage soon projected to cost only hundreds of dollars. However, state-of-the-art technology cannot identify arbitrary events and activities from these records. Problems in organizing and presenting such huge quantities of information efficiently also exist. Our solution is to limit the problem to tractable aspects of behavioral monitoring of nursing home residents in order to aid physicians with diagnosis and clinical research. Figure 1 illustrates the overall process of information monitoring, extraction and access in geriatric care. We emphasize automating visual and sensory information extraction, analysis and synthesis to better utilize the physician's time and to provide secure and efficient access into this data through behavioral assessment tools and reporting aids for the medical staff.

Rapid physical and cognitive changes are common in the elderly. An important part of their care is detecting and tracking these changes. Sometimes it requires substantial investigation to understand what underlying deficits cause a behavioral loss. Complete

“24/7” monitoring and logging of behavior provides an important asset for tracking behavioral changes.

Technologies capable of capturing and analyzing a video record of nursing home residents may significantly impact clinical care. For example, environmental and other situational factors that contribute to a behavioral disturbance could be identified. Early recognition of gait instability (either primary or medication-induced) may assist in reducing the tremendous morbidity, mortality and cost associated with unwitnessed falls. Accurate assessment and characterization of behavioral disturbances and circadian patterns of such problems would not only assist in individualizing patient treatment plans but would also improve the ability of clinician-researchers to minimize dosing and measure the impact of their interventions. The ability to create more accurate, complete behavioral logs would facilitate compliance with federal and state guidelines concerning the appropriate use of psychotropic medications.

The CareMedia system integrates and extends a number of research disciplines, computer vision, audio analysis, data mining, machine learning and information retrieval tailored to deal with errorful metadata. In subsequent sections we continue with a discussion of the techniques applied to the framework shown in Figure 1: activity information extraction,

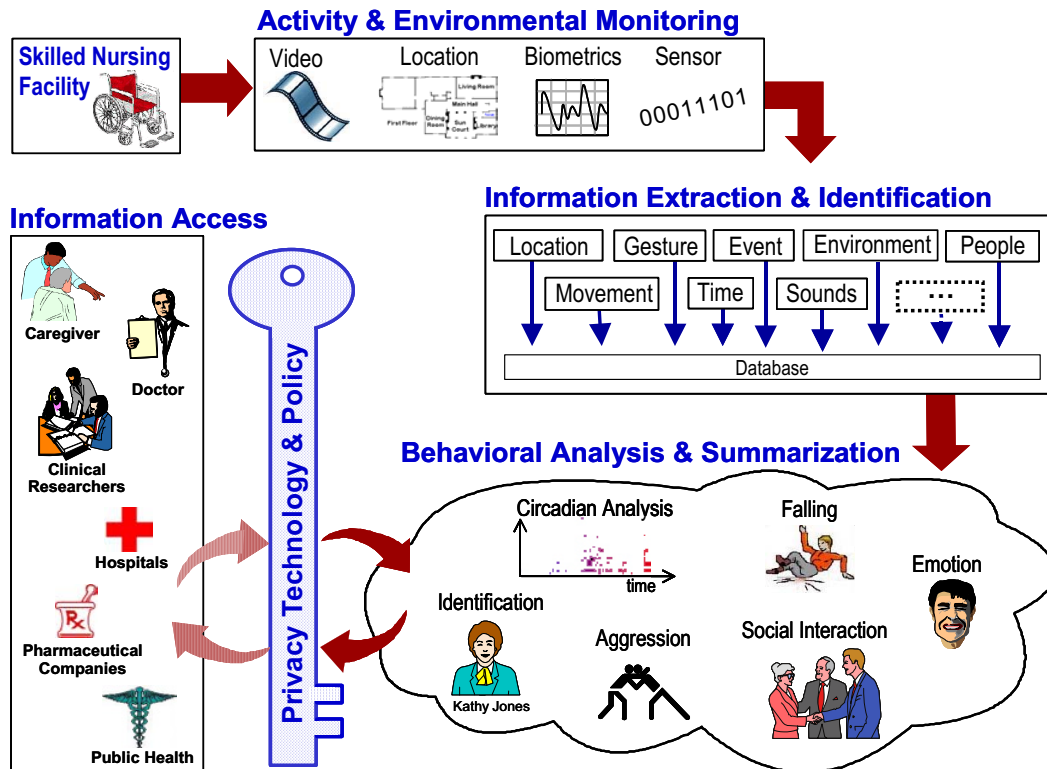


Figure 1. Conceptual overview of geriatric patient behavior monitoring and analysis

behavioral analysis and summarization, and access to a synthesized patient record.

## 2. Information Extraction

The goals for our systems are to locate, track, and identify humans and classes of human activity. The basic instrumentation will consist of video cameras for tracking location and identifying activities, and phased array microphones to localize and detect vocalization and noise. The constrained situational setting limits the number and classes of observables. We start with the fundamental requirement to identify who the people are as they move across cameras in a single event, and as they accumulate events over month-long periods. What they commonly do in the public spaces is wandering about, working on tasks, looking for things, speaking with others, eating and sleeping in public. The initial goal of the system is to (1) routinely *measure* normal patient behavior (e.g. time walking and speed, frequency and duration of social interactions) and to (2) *detect* clinically significant anomalous behavior (e.g. falling, agitated yelling) in this setting. In similar studies, wearable sensors that unobtrusively and continuously monitor heart rate, temperature, galvanic skin response and other physiological factors may be added. Technical problems include the synchronization of multiple data sources, dealing with highly redundant data, and compensating for missing data as residents move into and out of areas under visual and aural cover.

A longer term but more complex goal for the system is to develop automatic techniques to detect, assess and characterize aggressive behavior in these settings. e.g., did the patient shout at, push or strike another, how often, and under what circumstances.

Information extraction in the nursing home setting builds from a suite of technologies already in place within our laboratories:

- Locate and track moving individuals in real-time via visual tracking algorithms [4].
- Identify who is talking to whom, and what they are looking at, using head orientation tracking and gaze direction estimation [5].
- Recognize faces in video material [6].
- Extract speech utterances and other sounds under noisy conditions [7]

To assess disruptive vocalizations, audio analysis will need to identify loudness, frequency, and speaker, and to distinguish from normal speech and sounds. In previous research on recorded audio, we have successfully segmented audio recordings into continuous, related portions [8] using the signal-to-noise ratio of the waveform and clustering it into homogeneous regions, and will apply these techniques here. Furthermore, specific acoustic Hidden Markov

Models will be trained to recognize such events.. These models assume the vocalization can be modeled as a noise “word”, similar to our models of laughter [9]. Recent improvements to the CMU Sphinx-III speech recognition system have shown that it can perform well on noisy speech data [7] for large vocabulary, speaker independent speech transcription. Using the specialized acoustic models in the robust Sphinx-III system, combined with location-specific acoustic models, we can expect to identify interesting vocalizations and utterances with reasonable reliability.

A broad role of machine image understanding (IU) is to detect and recognize objects, track and interpret changes, and reconstruct and interpret events in the video collected from the environment. As with speech understanding, achieving such a total IU functionality is beyond the reach of current know-how. Yet by limiting the analysis to the specific domain and known physical environment, we can realize IU capabilities critical to the task of extracting and identifying individuals and their classes of actions.

Prior work in the Informedia systems have demonstrated that face detection [10], based on neural networks, facilitates object-content based video retrieval and video summarization [11], as opposed to conventional image-based techniques, such as color histograms. However, unlike Informedia, where the data source was edited, broadcast-quality video, this effort processes continuously captured video from albeit high resolution, surveillance-like cameras. For this we use probabilistic modeling of image properties [12], image segmentation [13], and tracking of individuals.

Figure 2 illustrates a system for real-time separation and tracking of individuals moving across realistic backgrounds [4]. This type of tracking serves as a basis for other more detailed sensors tailored for use here. Figure 3 illustrates the localization and tracking of faces within a larger image [5] (in this case taken from a panoramic camera). This in turn supports extraction of detailed facial features as illustrated in Figure 4. These features are used for head pose tracking and gaze direction estimation in real-time [14].

Activity recognition in this setting requires the design of algorithms capable of segmenting out “anomalous activities” in the context of a given space. The most applied techniques make use of models that rely on prior knowledge of the kinematics and dynamics that typify given activity [15, 16]. Such representations are powerful, yet the models are complicated, difficult to initialize automatically and unstable, especially when scenes are cluttered and reflect the activity of groups. Instead, we have pursued alternative activity models that do not rely on such detailed statistical knowledge, and are relatively robust

to real-world imaging situations, where resolution may be low or corrupted by noise [17, 18]. One such example is the work of Shi [19] who segmented 'related activities' in the video from our first clinical study by computing simple, global statistics over short sub-sequences. Shi's statistics were based on color and motion, and used to populate a graphical model. The model was subsequently segmented according to the normalized cuts framework [13], in an effort to group together portions of the video sharing similar color or motion characteristics. Preliminary results indicate that such low-resolution color and motion statistics can effectively characterize atypical 'activities' such as the occasional dispensing of medication from a cart.

Since we are using known, stationary and calibrated cameras, microphones and sensors, we more easily isolate background imagery and background noise to obtain more robust and valid feature information. Since the population is limited and identified by face, gait and voice, tracking individuals and re-synchronizing their identity is simplified. Information extracted for subsequent interpretation and summarization includes time, location, people, environmental factors, motion, gestures, speech, and sound, with potential links to a patient's medication schedule in future system versions.

### 3. Behavior Analysis and Summarization

In the previous section, we discussed how certain gestures, motions and verbalizations may be identified and classified. In this section, we will explain how to analyze the identified information, together with time and sensor data, and synthesize it into a comprehensive record of patient behavior.

The purpose is to provide caregivers better access to and management of information. Merely collecting and presenting complete 24/7 audiovisual records would overwhelm nursing staff and physicians rather than serving as a useful reporting aid and monitor. Hence,

extracted patient data needs to be synthesized into nuggets of information relevant and useful to the caregiver. The information technology applied automates the task of filtering down volumes of extracted aural, visual, and sensory data to key events and behaviors of interest. First phase goals are to at least identify material with unique characteristics requiring further inspection by the physician or the nursing home professional filling out behavioral logs. The longer-term goal is to utilize data and video mining techniques to uncover obscured trends and relationships of event occurrences with disruptive behavior.

To be realistic, the extracted gestures, motions and words are frequently misidentified. Therefore, our behavior analysis and summaries must take a possibly large error rate into account, while still providing useful data to the caregiver. Even though the analysis may be unreliable, the error may be adequately overcome by allowing a direct link back to the original source audio/video/sensor data, which has proved to be a successful strategy with Informedia summarization interfaces. Thus, a fall reported by the system can be verified to be an actual fall or a false alarm when looking at the source video record. The system also allows manual annotations to mark events that need to be brought to the attention of physicians and caregivers. The time required of the caregiver is reduced to quickly skimming through the portions of the complete record for the target behaviors suggested by the system at different levels of confidence. Despite errors, the caregiver can still review a comprehensive record of patient activities and behaviors in a short period of time.

The special events and event characterizations we intend to identify and track include:

- Social interaction networks – how often does this patient talk with someone and for how long.
- Physical activity and exertion – is the person



Figure 2. Real-time segmentation and tracking of individuals.



Figure 3. Locating and tracking human faces in a panoramic view.



Figure 4. Real-time facial feature extraction

lethargic or in a heightened state of activity, as evidenced by visible body motions or sensors.

- Physically aggressive behavior – does this person engage in striking or pushing others.
- Verbal behaviors associated with specific requests, e.g., loud talk being used to communicate the need to be fed or put into bed [20].
- Verbal behaviors with general, undefined needs, e.g., mumbling and disruptive talk during hallucinations [20].
- Verbal behaviors associated with self-stimulation, such as loud singing [20].
- Circadian activity patterns – are there times of day when the patient shows unusual activity or behaviors.

Disruptive vocalization (DV) is a very common, distressing problem that impacts the quality of life of residents and consumes considerable staff time. Event-alerting techniques will be developed to identify, characterize and log the amplitude, frequency and duration of the episodes of DV. Identifying environmental and/or psychosocial contextual factors captured by the digital log preceding the episodes of DV will not only add to our understanding of this phenomenon, but may suggest potential avenues of individualized interventions.

Visual and acoustic information of movement and activity can be collected and interpreted to assess a resident's status and needs. By classifying typical sounds and movements, we can determine normal acoustic and visual scenes, and/or determine dangerous or hazardous situations (e.g., person falling, shattering glass, etc.) that require attention and help. The acoustic and visual signal may also be useful for very specific signs of distress, e.g., waving, crying, and calling. Care must be taken in generalizing results, as agitation can be a very individualized form of expression.

#### 4. Information Access

It is known that most elderly people, nursing home residents, and people with dementia who exhibit agitated behaviors do so at very low frequencies [2]. A continuous audio-visual record will have volumes of redundant or trivial data with sparse occurrences of key episodes showing agitation. Interface work will need to filter out the irrelevant material and highlight relevant episodes in response to a query. Retrieval issues include how to deal with errorful metadata introduced by imperfect data analysis, e.g., errors in identification, speech recognition, and image classification. Other issues include the tradeoff between precision and recall: for recalling all events of high impact agitation behavior like biting, it is more important to recall all the events, even if additional non-biting events are also retrieved. For examining patterns of data it may be more important that all answers fit the query precisely.

For queries with high recall, the interface will be instrumented to enable a skilled user to quickly assess whether the flagged event requires treatment, consultation, or is irrelevant.

At the core of information access is a prototype system that presents an interface to the underlying raw data as well as the extracted and analyzed behavioral metadata. This system focuses on the geriatric care specialist as informed user interested in monitoring behavior of their patients. Our particular focus allows user queries to be better interpreted in the context of a geriatric care system. This focus lets us map physicians' and nurses' queries into the extracted patient data. With experience from caregivers, we will subsequently develop interactive visualizations and relevance feedback to support intelligent browsing and filtering by this user community. As a result, we can create tailored interfaces appropriate for filling out standardized behavioral logs.

Variable style summaries can collapse the redundancy while preserving portions of interest to the medical professional. Thus, location summaries can provide an overview of what behaviors were noted in different areas or generate a time-synchronized map displaying a patient's wanderings, and time summaries can show the distribution of behaviors over the selected time periods. All of these can be used together to form multimodal queries. The ability to interact directly with the visualization enables natural iteration during query refinement and relevance feedback.

Since the source data is always available and linked to the summary, a user can go back to the original captured record to verify the analysis. The caregivers will also be able to "bookmark" the source data for events that require further attention or review. Supporting this review of the source data will be intelligent video skim technology [21], which can provide a quick overview over time periods deemed interesting by the behavioral analysis.

These visualizations can also be used to do aggregate analysis, which is more difficult to extract from observation or event logs. Our prior projects have recorded global position information for multiple people over multiple days. Extracting and superimposing paths can help identify patterns more abstract than instantaneous location alone. We will build interfaces to help find periodic and sequential behavior, revealing that a problem tends to occur only on weekends, or that one patient's behavior tends to track another's.

We will also extend existing approaches to pool data without compromising an individual's privacy [22]. We will enable the patient and his or her family control of data that is provided to doctors, clinical researchers,

public health agencies, and other caregivers. We will develop algorithms to exchange and pool data, while maintaining privacy and confidentiality, that will provide the benefits of multi-site statistical analyses. Our plan is to create a tiered set of access levels, allowing for example a public health agency access only to aggregated data, with a different level of access for the attending physician, while the floor nurse might have access to yet another portion of the record of a specific individual. In the far-term, when the video data can be processed and accurately characterized contemporaneously, then it may be immediately discarded, retaining only the derived information, thus maintaining a higher level of privacy. Our system will provide a range of mechanisms for security and privacy of the data, their application subject to administratively and individually determined policy.

## 5. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0205219 and the Advanced Research and Development Activity, a U.S. Government entity which sponsors and promotes research of import to the Intelligence Community. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government.

## 6. References

- [1] H. D. Wactlar, M. G. Christel, Y. Gong, and A. G. Hauptmann, "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library," *IEEE Computer*, vol. 32, pp. 66-73, 1999.
- [2] J. Cohen-Mansfield, "Assessment of Disruptive Behavior/Agitation in the Elderly: Function, Methods and Difficulties," *Journal of Geriatric Psychiatry and Neurology*, vol. 8, pp. 52-60, 1995.
- [3] D. Gilley, R. Wilson, D. Bennet, and et.al., "Predictors of Behavioral Disturbance in Alzheimer's Disease," *Journal of Gerontology*, vol. 46, pp. 362-371, 1991.
- [4] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel, "Multimodal People ID for a Multimedia Meeting Browser," presented at ACM Multimedia '99, Orlando, FL, 1999.
- [5] R. Stiefelhagen and J. Yang, "Gaze Tracking for Multimodal Human-Computer Interaction," presented at International Conference on Acoustics, Speech and Signal Processing (ICASSP'97), Munich, Germany, 1997.
- [6] H. Schneiderman and T. Kanade, "A Statistical Method for 3D Object Detection Applied to Faces and Cars," presented at IEEE Computer Vision and Pattern Recognition (CVPR), Hilton Head, SC, 2000.
- [7] R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern, "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination," presented at IEEE Conference on Acoustics, Speech and Signal Processing, Salt Lake City, UT, 2001.
- [8] M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," presented at DARPA Proceedings of the Ninth Spoken Language Systems Technology Workshop, Harriman, NY, 1997.
- [9] P. E. Kennedy and A. G. Hauptmann, "Laughter Extracted from Television Closed Captions as Speech Recognizer Training Data," presented at 6th European Conference on Speech Communication and Technology, Budapest, Hungary, 1999.
- [10] H. Rowley, S. Baluja, and T. Kanade, "Human Face Detection in Visual Scenes," Carnegie Mellon University, Pittsburgh, PA, School of Computer Science Technical Report CMU-CS-95-158 1995.
- [11] M. Smith and T. Kanade, "Video Skimming and Characterization Through the Combination of Image and Language Understanding Techniques," presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR97), San Juan, Puerto Rico, 1997.
- [12] H. Schneiderman and T. Kanade, "Probabilistic Modeling of Local Appearance and Spatial Relationships of Object Recognition," presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Santa Barbara, CA, 2000.
- [13] J. Shi and J. Malik, "Motion Segmentation and Tracking Using Normalized Cuts," presented at International Conference on Computer Vision (ICCV), Bombay, India, 1998.
- [14] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling Focus of Attention for Meeting Indexing," presented at ACM Multimedia '99, Orlando, FL, 1999.
- [15] C. Wren, Clarkson, B., Pentland, A., "Understanding Purposeful Human Motion," presented at IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000.
- [16] Y. Yacoob, Black, M.J., "Parameterized Modeling and Recognition of Activities," *Computer Vision and Image Understanding*, vol. 73, pp. 232-247, 1999.
- [17] R. Cutler, Davis, L., "Robust Real-Time Periodic Motion Detection, Analysis, and Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 781-796, 2000.
- [18] R. Polana, Nelson, R., "Non-Parametric Recognition of Non-rigid Motion," University of Rochester, Department of Computer Science, Rochester, NY TR 575, March 1995.
- [19] J. Shi, Zhong, H., "Finding (Un)Usual Events in Video," presented at The Learning Workshop, Snowbird, Utah, 2003.
- [20] J. Cohen-Mansfield and P. Werner, "Typology of Disruptive Vocalizations in Older Persons Suffering from Dementia," *International Journal of Geriatric Psychiatry*, vol. 12, pp. 1079-1097, 1997.
- [21] M. Smith, Kanade, T., "Video Skimming for Quick Browsing Based on Audio and Image Characterization," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.
- [22] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *Journal of Law, Medicine and Ethics*, vol. 25, pp. 98-110, 1997.