CareMedia: Automated Video and Sensor Analysis for Geriatric Care

NSF Cooperative Agreement No. IIS-0205219
Annual Progress Report
March 2005

Carnegie Mellon University
School of Computer Science
Pittsburgh, PA 15213-3890

Principal Investigator: Howard Wactlar
School of Computer Science

Co-PI's: Alexander Hauptmann
Computer Science Dept.
Takeo Kanade
Robotics Institute
Scott Stevens
Computer Science Dept.
Ashok Bharucha
Western Psychiatric Institute

Copyright © 2005 Carnegie Mellon University
1 Project Participants

What people have worked on the project?

<table>
<thead>
<tr>
<th>Participant’s Name</th>
<th>Project Role</th>
<th>&gt;160 Hours?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Howard D. Wactlar</td>
<td>Principal Investigator</td>
<td>Yes</td>
</tr>
<tr>
<td>Takeo Kanade</td>
<td>Co-Principal Investigator</td>
<td>Yes</td>
</tr>
<tr>
<td>Ashok Bharucha</td>
<td>Co-Principal Investigator</td>
<td>Yes</td>
</tr>
<tr>
<td>Scott M. Stevens</td>
<td>Co-Principal Investigator</td>
<td>Yes</td>
</tr>
<tr>
<td>Alexander G. Hauptmann</td>
<td>Co-Principal Investigator</td>
<td>Yes</td>
</tr>
<tr>
<td>Christopher Atkeson</td>
<td>Project Scientist</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Are any other persons involved with the project?
None.

What other organizations have been involved as partners?
University of Pittsburgh Medical Center (UPMC).
Western Psychiatric Institute and Clinic (WPIC).

Have you had other collaborators or contacts?
Longwood at Oakmont Retirement Resort, Dementia Care Facility
Asbury Heights Dementia Care Facility

2 Project Activities and Findings

Overview
This research will create a continuously recorded, digital history of patients’ activities in a nursing home by capturing all that is heard, seen and experienced. 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 proposed work in (i) information extraction and retrieval, (ii) behavior recognition, analysis and summarization, and (iii) secure, efficient visual information access will let geriatric care specialists obtain greater insights into problems, effectiveness of treatments, and environmental and social influences on patient behavior. Prototype systems for activity monitoring and behavior analysis will be deployed at local area nursing homes and dementia wards to 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).
2.1 Activities

We work through the privacy, consent, and data protection concerns associated with placing surveillance audio and video equipment in the nursing facilities on a home-by-home basis. We have received Carnegie Mellon University Institutional Review Board approval to conduct such research investigations with human subjects, as well as approval from the State Board of Health. The experiments at the nursing facilities themselves are conducted under the supervision of Dr. Ashok Bharucha, M.D., a geriatric psychiatrist with the CareMedia project. Our Advisory Board for the project gauges ethical concerns and assesses broader societal implications for patient data collection and use in controlled settings.

We encountered unexpected difficulties securing appropriate patient/guardian consent at one anticipated deployment site, which cost us considerable man-hours in equipment installation and set-up. We are currently in the process of recruiting new study sites, with much improved understanding of the politics involved with securing patient, family, and staff participation.

We have continued to experiment on preliminary “surveillance-type” video, gathered at a nursing care facility, Longwood at Oakmont’s Dementia Care Facility. We exploit this set of data extensively, and have designed and prototyped an interface into the massive video record, allowing clinicians and others ready access to items of interest. Additionally, we have simulated a nursing home environment at Carnegie Mellon multiple times, duplicating lighting conditions and occlusion problems we anticipate in our future large-scale deployment. Using this captured video data, we developed a technique to track multiple individuals across multiple camera views. We also prototyped a technique for obscuring the faces of those individuals who choose not to participate in our studies, directly addressing very real privacy issues our research project faces.

2.1.1 Automated Analysis of Nursing Home Observations

Pervasive activity monitoring in a skilled-nursing facility helps capture a continuous audio and video record. The CareMedia project analyzes this video information by automatically tracking people, helping to efficiently label individuals, and characterizing selected activities and actions.

The core technological challenge for CareMedia is transforming captured video and audio into a meaningful information resource. Observations in a nursing home provide a concrete setting for this challenge. In particular, success requires these components:

- **Tracking people** in the captured video stream. To accumulate information about any person, we must be able to continuously track moving people. Tracking is perhaps the most mature technology, with research going back several decades. In simple cases, separating a moving person from the background is trivial; in practice, this effort is complicated by occlusions, multiple, difficult-to-separate
individuals crisscrossing the room, background and lighting changes, and inanimate objects deposited in new locations.

• **Identifying and labeling individuals.** We also need to separate a single person’s track over multiple days. This involves associating a particular track with an individual of interest. The large volume of continuous observations prohibits a strictly manual procedure here.

• **Analyzing specific individuals’ activities.** Given that we can follow an individual over time, we want to characterize and quantify what the individual is doing. This analysis is open-ended— we would like to identify as many different activities as possible, remembering that they must be robustly detected in different real-world situations to be useful.

We will achieve these components through automated analysis of video and audio information collected in the nursing home. The video lets caregivers directly see and hear evidence of episodes as they review statistical summaries. Once the system has tracked and identified an individual and recognized his or her activities, we can create meaningful summaries of these activities and associated changes over multiple days or weeks. Realistically, the tracked people and their activities or gestures will be frequently misidentified. Even though automated analysis might have flaws, a caregiver can overcome many errors by directly linking back to the original source video. For example, the caregiver can dismiss a system-reported fall as a false alarm upon viewing the actual video record. The caregiver can still review a comprehensive record of important activities in a short time period.

### 2.1.2 Addressing Privacy Concerns

We have developed an algorithm to protect visual individual identity by “blocking” the face of specified individuals. This algorithm automatically tracks
Fig.1. The face blocking algorithm with automatic or interactive initialization.

face and obscures it using a pixilation effect. Figure 1 is a diagram of the new algorithm.

The person whose identity is to be protected can be either specified interactively by pointing out his/her face location in the video image with a GUI tool, or automatically by using face detection and face identification modules associated with the person’s name (when high resolution is available.) To avoid missing the face in the face detection process, we push the video frame that failed in the detection step into a stack. We backtrack through the stack when a face model is built. A face model is built for each specified person once its location is given in a video image.

We use color histogram to model the face. The color histogram model is kept updated throughout the video. Face tracking is performed by a color histogram based mean-shift algorithm which is very fast. We then block the face identity by using a pixilation effect.

Fig.2 illustrates some results enabled by our face-blocking algorithm. We can either protect individual or multiple persons at the same time.

2.1.3 Image Registration with Uncalibrated Cameras in Hybrid Vision Systems

We have investigated the problem of robust registering of images among perspective and omnidirectional cameras in a hybrid vision system (HVS). Nonlinearity in an HVS introduced by omnidirectional cameras poses challenges for computing pixel correspondences among images. In previous HVSs, cameras must be calibrated by performing registration. We developed a non-linear
approach for registering images in an HVS (fig. 3) without requiring calibration of cameras. We first analyzed the homographies between omnidirectional and perspective images under a local planar assumption. We then introduced a robust patch level registration algorithm by exploiting a constraint on large 3D spatial planes. Our approach enables an HVS for applications that require quick deployment or active cameras. Experimental results have demonstrated feasibility.

Fig.3. An illustration of an HVS in an indoor space.

2.1.4 Towards Automatic Analysis of Social Interaction Patterns in a Nursing Home Environment from Video

We developed an ontology-based approach for analyzing social interaction patterns in a nursing home from video. Social interaction patterns are broken into individual activities and behavior events using a multi-level context hierarchy ontology framework. To take advantage of an ontology in representing how social interactions evolve, we designed and refined the ontology based on knowledge gained from 80 hours of video recorded in the public spaces of a nursing home.
Activities of groups of people are firstly treated as interaction patterns between any pair of partners and are then further broken into individual activities and behavior events using a multi-level context hierarchy graph. The ontology is implemented using a dynamic Bayesian network to statistically model the multi-level concepts defined in the ontology. We developed a prototype system to illustrate the concept. Experimental results have demonstrated the feasibility of our approach.

2.1.5 Towards Robust Face Recognition From Multiple Views

We have introduced a novel approach to aid face recognition: Using multiple views of a face, we construct a 3D model instead of directly using the 2D images for recognition. Our framework is designed for videos (which contain many instances of a target face from a sequence of slightly differing views), as opposed to a single static picture of the face. Specifically, we reconstruct the 3D face shapes from two orthogonal views and select features based on pairwise distances between landmark points on the model using Fisher's Linear Discriminant.

While 3D face shape reconstruction is sensitive to the quality of the feature point localization, our experiments show that 3D reconstruction together with the regularized Fisher's Linear Discriminant can provide highly accurate face recognition from multiple facial views. Experiments on the CMU PIE (Pose, Illumination and Expressions) database of 68 faces, with at least 3 expressions under varying lighting conditions, demonstrate vastly improved performance.

2.1.6 Articulated Motion Modeling for Activity Analysis

Our new algorithm for articulated human motion segmentation estimates parametric motions of body parts and segments images into moving regions accordingly. Our approach combines robust optical flow estimation, RANSAC, and region segmentation using color and Gaussian shape priors. This combination
results in an algorithm that can robustly estimate and segment multiple motions, even for moving regions with small support and in low resolution images.

Fig. 6. (a) Activity analysis at a nursing home. Arrows indicate directions of movement. (b) Articulated motion segmentation of a dancer in our lab. Both sets of results are produced by our algorithm.

Based on the raw motion segmentation, consistent body motions are detected over time to characterize human activity. The effectiveness of this approach has been demonstrated, characterizing dining activities of patients at a nursing home.

2.1.7 Multiple Frame Motion Inference Using Belief Propagation
We have developed a new algorithm for automatic inference of human upper body motion. A graph model is used for inferring human motion, and motion inference is posed as a mapping problem between state nodes in the graph model and features in image patches. Belief propagation is utilized for Bayesian inference in this graph. A multiple-frame inference model/algorithm combines both structural and temporal constraints in human motion. We also developed a method for capturing constraints of human body configuration under different view angles. The algorithm has been applied in a prototype system that can automatically label upper body motion from videos, without manual initialization of body parts.
In our experiments, we found error in about 12% of the total frames of videos tested. This does not include the cases where the estimation is roughly correct but inaccurate. Errors occur mostly in occlusion situations or more subtle situations, such as when two hands are too close together.

In the future, we look to improve the algorithm by: 1) comparing the results of using detected candidate states and uniformly sampled candidate states; 2) utilizing more features or better methods of using these features, in order to deal with some difficult situations; 3) attempting to find a better solution to the view-specific problem.
2.2 Findings


Much of the motion capture data used in animations, commercials, and video games is carefully segmented into distinct motions either at the time of capture or by hand after the capture session. As we move toward collecting more and longer motion sequences, however, automatic segmentation techniques will become important for processing the results in a reasonable time frame. We have found that straightforward, easy to implement segmentation techniques can be very effective for segmenting motion sequences into distinct behaviors. In this paper, we present three approaches for automatic segmentation. The first two approaches are online, meaning that the algorithm traverses the motion from beginning to end, creating the segmentation as it proceeds. The first assigns a cut when the intrinsic dimensionality of a local model of the motion suddenly increases. The second places a cut when the distribution of poses is observed to change. The third approach is a batch process and segments the sequence where consecutive frames belong to different elements of a Gaussian mixture model. We assess these three methods on fourteen motion sequences and compare the performance of the automatic methods to that of transitions selected manually.


This paper addresses the problem of robust registering of images among perspective and omnidirectional cameras in a hybrid vision system (HVS). Nonlinearity in an HVS introduced by omnidirectional cameras poses challenges for computing pixel correspondences among images. In previous HVSs, cameras must be calibrated by performing registration. In this paper, we propose a non-linear approach for registering images in an HVS without requiring calibration of cameras. We first discuss the homographies between omnidirectional and perspective images under a local planar assumption. We then propose a robust patch level registration algorithm by exploiting a constraint on large 3D spatial planes. The proposed approach enables an HVS for applications that require quick deployment or active cameras. Experimental results have demonstrated feasibility of the proposed approach.

In this paper, we propose an ontology-based approach for analyzing social interaction patterns in a nursing home from video. Social interaction patterns are broken into individual activities and behavior events using a multi-level context hierarchy ontology framework. To take advantage of an ontology in representing how social interactions evolve, we design and refine the ontology based on knowledge gained from 80 hours of video recorded in the public spaces of a nursing home. The ontology is implemented using a dynamic Bayesian network to statistically model the multi-level concepts defined in the ontology. We have developed a prototype system to illustrate the proposed concept. Experiment results have demonstrated feasibility of the proposed approach. The objective of this research is to automatically create concise and comprehensive reports of activities and behaviors of patients to support physicians and caregivers in a nursing facility.


In this paper, we propose a multimodal system for detecting human activity and interaction patterns in a nursing home. Activities of group people are firstly treated as interaction patterns between any pair of the two partners and are then further broken into individual activities and behavior events using a multi-level context hierarchy graph. The graph is implemented using a dynamic Bayesian network to statistically model the multi-level concepts. We have developed a coarse-to-fine prototype system to illustrate the proposed concept. Experiment results have demonstrated feasibility of the proposed approaches. The objective of this research is to automatically create concise and comprehensive reports of activities and behaviors of patients to support physicians and caregivers in a nursing facility.


This paper presents a novel approach to aid face recognition: Using multiple views of a face, we construct a 3D model instead of directly using the 2D images for recognition. Our framework is designed for videos, which contain many instances of a target face
from a sequence of slightly differing views, as opposed to a single
static picture of the face. Specifically, we reconstruct the 3D face
shapes from two orthogonal views and select features based on
pair-wise distances between landmark points on the model using
Fisher's Linear Discriminant. While 3D face shape reconstruction
is sensitive to the quality of the feature point localization, our
experiments show that 3D reconstruction together with the
regularized Fisher's Linear Discriminant can provide highly
accurate face recognition from multiple facial views. Experiments
on the Carnegie Mellon PIE (Pose, Illumination and Expressions)
database containing 68 people’s faces with at least 3 expressions
under varying lighting conditions demonstrate vastly improved
performance.

Christel, M., Huang, C., Moraveji, N., “Evaluating Content-Based Filters for
Image and Video Retrieval.” 27th Annual International ACM SIGIR Conference,

This paper investigates the level of metadata accuracy required for
image filters to be valuable to users. Access to large digital image
and video collections is hampered by ambiguous and incomplete
metadata attributed to imagery. Though improvements are
constantly made in the automatic derivation of semantic feature
concepts such as indoor, outdoor, face, and cityscape, it is unclear
how good these improvements should be and under what
circumstances they are effective. This paper explores the
relationship between metadata accuracy and effectiveness of
retrieval using an amateur photo collection, documentary video,
and news video. The accuracy of the feature classification is varied
from performance typical of automated classifications today to
ideal performance taken from manually generated truth data.
Results establish an accuracy threshold at which semantic features
can be useful, and empirically quantify the collection size when
filtering first shows its effectiveness.

Christel, M., Moraveji, N., “Finding the Right Shots: Assessing Usability and
Performance of a Digital Video Library Interface.” Proceedings of ACM

The authors developed a system in which visually dense displays
of thumbnail imagery in storyboard views are used for shot-based
video retrieval. The views allow for effective retrieval, as
evidenced by the success achieved by expert users with the system
in interactive query for NIST TRECVID 2002 and 2003. This
paper demonstrates that novice users also achieve comparatively
high retrieval performance with these views using the TRECVID
2003 benchmarks. Through an analysis of the user interaction logs,
heuristic evaluation, and think-aloud protocol, the usability of the

CareMedia: Automated Video and Sensor Analysis for Geriatric Care
March 2005 Annual Progress Report
video information retrieval system is appraised with respect to shot-based retrieval. Design implications are presented based on these TRECVID usability evaluations regarding efficient, effective information retrieval interfaces to locate visual information from video corpora.


Aural and visual cues can be automatically extracted from video and used to index its contents. This paper explores the relative merits of the cues extracted from the different modalities for locating relevant shots in video, specifically reporting on the indexing and interface strategies used to retrieve information from the Video TREC 2002 and 2003 data sets, and the evaluation of the interactive search runs. For the documentary and news material in these sets, automated speech recognition produces rich textual descriptions derived from the narrative, with visual descriptions and depictions offering additional browsing functionality. Through speech and visual processing, storyboard interfaces with query-based filtering provide an effective interactive retrieval interface. Examples drawn from the Video TREC 2002 and 2003 search topics and results using these topics illustrate the utility of multiple-document storyboards and other interfaces incorporating the results of multimodal processing. http://www.icassp2004.com/


We present an algorithm for automatic inference of human upper body motion. A graph model is proposed for inferring human motion, and motion inference is posed as a mapping problem between state nodes in the graph model and features in image patches. Belief propagation is utilized for Bayesian inference in this graph. A multiple-frame inference model/algorithm is proposed to combine both structural and temporal constraints in human motion. We also present a method for capturing constraints of human body configuration under different view angles. The algorithm is applied in a prototype system that can automatically label upper body motion from videos, without manual initialization of body parts.
We propose an algorithm for articulated human motion segmentation that estimates parametric motions of body parts and segments images into moving regions accordingly. Our approach combines robust optical flow estimation, RANSAC, and region segmentation using color and Gaussian shape priors. This combination results in an algorithm that can robustly estimate and segment multiple motions, even for moving regions with small support and in low resolution images. Based on the raw motion segmentation, consistent body motions are detected over time to characterize human activity. The effectiveness of this approach is demonstrated in a real scenario: characterizing dining activities of patients at a nursing home.


Pervasive activity monitoring in a skilled-nursing facility helps capture a continuous audio and video record. The CareMedia project analyzes this video information by automatically tracking people, helping to efficiently label individuals, and characterizing selected activities and actions. The project emphasizes detecting eating activity in the dining hall and personal hygiene. Through this work, the video record becomes an information asset that can provide geriatric-care specialists with greater insights to the elderly and evaluation of their behavioral problems. Evaluations of the effectiveness of analyzing such a large video record illustrate our approach's feasibility.


Combining retrieval results from multiple modalities plays a crucial role for video retrieval systems, especially for automatic video retrieval systems without any user feedback and query expansion. However, most of current systems only utilize query independent combination or rely on explicit user weighting. In this work, we propose using query-class dependent weights with in a hierarchical mixture-of-expert framework to combine multiple retrieval results. We first classify each user query into one of the four predefined categories and then aggregate the retrieval results.
with query-class associated weights, which can be learned from the development data efficiently and generalized to the unseen queries easily. Our experimental results demonstrate that the performance with query-class dependent weights can considerably surpass that with the query independent weights.

3 Training and Development
Graduate student support has enabled research into object and people tracking, and activity and behavior classification in video.

Clinical psychiatrist Dr. Ashok Bharucha has been introduced to computation techniques for capturing behavior in a clinical setting.

Undergraduate students have been exposed to techniques for video analysis during the analysis phase of our pilot study.

4 Outreach Activities
Our work at a nursing home during the pilot study has exposed staff and patient families to the potential clinical value of computational technology.

5 Publications and Products


6 Contributions

6.1 Contributions within discipline

A robust technique for tracking multiple people across a camera network
Techniques for auto-elimination of specified human subjects from video content.
Technique for masking the ID of specific individuals in video.
Semi-automated system for rapid (10% of real-time) viewing, annotating and
truthing of continuously captured video.

6.2 Contributions to other disciplines

Initial field studies were exploratory, but the subsequent planned CareMedia/nursing home collaboration will result in the world's most comprehensive study of the ecology of behaviors in a dementia ward to date.

6.3 Contributions to human resource development
Nursing home field studies have educated computer scientists with respect to issues of privacy and sensitivity in automated analysis of observational data.

6.4 Contributions to resources for research and education

Nothing significant yet.

6.5 Contributions beyond science and engineering

Nothing significant yet.

7 Special Requirements

7.1 Objectives and Scope
A brief summary of the work to be performed during the next year of support if changed from the original proposal.
No change.

7.2 Special Reporting Requirements
Do special terms and conditions of your award require you to report any specific information that you have not yet reported?
No.

7.3 Unobligated Funds
Do you anticipate that more than twenty percent of the funds under your NSF award will remain unobligated at the end of the period for which NSF currently is providing support?
No.

7.4 Animals, Human Subjects, Biohazards
Has there been any significant change in animal care and use, use of human subjects, or biohazards, from what has previously been approved?
No.