Infrastructure for Machine Understanding of Video Observations in Skilled Care Facilities – Implications of Early Results from CareMedia Case Studies

Howard D. Wactlar, Michael Christel, Alexander Hauptmann,
Datong Chen, Jie Yang

Carnegie Mellon University, School of Computer Science
Pittsburgh, Pennsylvania, USA
wactlar@cmu.edu
{christel, alex, datong, yang}@cs.cmu.edu

Abstract. CareMedia captures and analyzes a continuous audio and video record of behavior and activity in a skilled nursing facility. Through computer vision and machine learning we automatically identify individuals, classify activities, recognize behaviors, and extract relevant events. Two extensive field trials have been undertaken which produced meaningful but sometimes limited clinical results. Based on an analysis of this experience, combined with the development of new approaches and algorithms, we describe a radically improved audiovisual recording and computing infrastructure to be implemented, enabling longitudinal studies with comprehensive video and audio coverage provided with a mix of resolution, frame-rate, compression and storage requirements. Changes in the amount or rate of social interaction, eating, walking, gait, arm swing, and other attributes of motion from each patient's baseline behavior will also be made easy to flag for professional review and diagnosis through the recording and analysis infrastructure, as such changes are key indicators to the benefits and possibly detrimental side effects of pharmacological interventions.

1 Introduction

The CareMedia [1] [2] project at Carnegie Mellon University seeks to create a meaningful, manageable video information resource that enables more complete and accurate assessment, diagnosis, treatment, and evaluation of behavioral problems for the elderly. As part of this research, a continuous audio and video record is captured to monitor activity in a skilled nursing facility. Through work in information extraction and behavior analysis, this record will ultimately be transformed into an information asset whose efficient, secure presentation can empower geriatric care specialists with greater insights into problems, effectiveness of treatments, and determination of environmental and social influences on patient behavior. This work focuses on senile dementia patients in a specialized care facility, where close monitoring is an objective toward providing better care.

Data for diagnosis and clinical studies are now typically gathered by hand during a few brief periods per week. More detailed, exhaustive behavioral assessment scales have been developed, but have the drawback of requiring many observations and hence being too time-consuming for regular use by the clinical staff. The information technology developed in this work provides geriatric care specialists with a better window into the lives of senile dementia patients. Their behavior can then be more accurately measured and interpreted, enabling detection of gradual changes in baseline behavior and leading to treatment that optimally reduces agitation while allowing awareness and responsiveness.

2 Technical Challenges

The goal of CareMedia is to automatically transform the captured video and audio into a meaningful information resource. This implies accomplishing the following:

1. Tracking people in the captured video stream. To accumulate information about any person, we need to be able to continuously track moving people. Tracking is perhaps the most mature technology, with research going back for several decades. While in simple cases it is trivial to separate a moving person from the background, in practice this effort is complicated by occlusions, multiple, difficult to separate individuals criss-crossing the room, background and lighting changes, and inanimate objects deposited in new locations.
2. Identifying and labeling individuals. We also need to separate out the track of a single person over multiple days. This involves associating a particular 'track' with an individual of interest. The many hours of observations prohibit a strictly manual procedure here.
3. Analyzing activities of specific individuals. 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, keeping in mind that they need to be robustly detected in different real-world situations. This includes the range of gross motion (e.g., walking), small motor (e.g., eating) and fine motion (e.g., tremors, twitching) behavior.
4. Detecting subtle and gradual changes in human performance and recognizing low frequency events and patterns. The infrastructure must provide for the capture and analysis of data over multiple weeks for all individuals. Clinically this is essential for detecting progressive degeneration and assessing whether pharmacological, environmental and behavioral interventions are having their intended effects. Logistically, this implies data generation of terabytes per day and petabytes per year.

3 Prior and Existing Recording Infrastructure

Under National Science Foundation support, we have successfully deployed two recording infrastructures and collected over 200 hours of video in a skilled nursing home environment from the first. This installation was in an Alzheimer's unit that cares for about 20 patients and consists of a hallway connecting patient rooms, a dining/activity room, and a television room. Each public hallway and room had two color NTSC cameras with wide angle lenses connected to a computer performing MPEG-2 video and audio compression and data storage

In addition, we are in the midst of a second field experiment with improved infrastructure over the first in terms of coverage and resolution. This latest installation is a 55-bed dementia unit in which all non-private spaces of the dementia units (hallways, nursing station, and recreation/dining rooms) are equipped with 640x480 pixels, 30 frames per second, NTSC, progressive scan digital video cameras. Each 100-foot long hallway has eight cameras, the entrance hallway has four cameras, the dining/recreation room has six, and the central core/nursing station has six. The output from each camera is compressed in real-time to high quality but still lossy MPEG-2 format and transmitted via Ethernet to a server, with the exception of one section of a common recreation/meeting room that is not instrumented in order to minimize infringement on resident/family privacy (as recommended by the state department of health). Each server receives data from eight cameras and is equipped with two terabytes of hard disk storage enabling continuous recording for approximately four days. During data collection the hard drives are replaced every four days. Each camera contains two microphones affording limited audio capture (due largely to distance constraints) as well as video capture. In the high-use activity rooms, a phased-array microphone system consisting of eight microphones is used to reconstruct and isolate sound sources. Our first large scale data collection will begin in the autumn of 2004. The recordings will take place 24 hours per day, generating a total of 75 terabytes of data over four weeks. Note that administrators from the nursing home report that incidents of critical nighttime agitation would be missed if we were to limit the recordings to just the daylight hours, necessitating such complete coverage.

This system design was developed considering the clinical need for long-term, longitudinal study and image processing need for quality and resolution, while constrained by budget. Clinically it was necessary to ensure a high probability that study participants would be in the field of view of at least one camera at all times in all instrumented spaces. Further, image quality must be sufficient for a clinician to make meaningful observations. Computationally, it was necessary to field sufficient cameras to ensure a reasonable probability that the residents can be automatically tracked and identified. Fine-grained behaviors/movements such as tremors, eye gaze, and facial expressions cannot be detected automatically with the cameras deployed here. However, automatic patient identification and tracking are requirements of the current CareMedia effort.

For automatic patient tracking, experiments in a fixed campus test setting showed that with proper camera placement, reliable tracking of multiple individuals is possible if all points in the space are in simultaneous view of at least three cameras (barring highly unusual occlusions such as one person surrounded by people and objects such that no camera could view that person). Resolution for this tracking task is not as critical as for patient identification.

For automatic patient identification, from a processing perspective it would be ideal if we could affix a unique tag to each participant in the study. Multiple color and varied shape tags or electronically identifying bracelets would enable highly reliable identification. Unfortunately nursing home staff is adamant that this is not feasible in the day-to-day clinical setting. Frequent changing of clothes during the day and even changing mobility appliances (wheelchairs, walkers and canes) would increase the staff burden too greatly should we require a unique tag to be associated with and visible on each patient. Electronically sensed identifying bracelets, armbands, necklaces and any form of badge are also objectionable as patients in these facilities are not required to wear any identifying items. Given the constraint that we may not add any visually discriminating features to study participants, we chose to utilize automatic face identification. Reliable machine face identification minimally requires a full frontal, 64-pixel square image [3] [4]. We placed two cameras at both the entrance and exit of each patient hallway. Each hallway is approximately 10 feet wide. The average face is approximately 6 inches wide. Allowing for field of view overlap between cameras, and that faces will seldom be found next to the walls, this configuration constitutes the absolute minimum with a chance of face recognition (640 pixels/camera x 2 cameras divided by 10 feet times .5 feet/face = 64 pixels/face horizontally). Unfortunately, patients are seldom positioned ideally. More typically their head position as seen by ceiling mounted cameras is off axis to their view, thereby greatly degrading reliable machine face recognition.

4 Analyzing the Data with Computer Vision and Machine Learning

In order to recognize and classify objects, actions, activities and events, we apply machine learning techniques for extracting low level features and inferring high level content. Classifiers such as face detectors and recognizers can be trained using manually labeled data as input to machine learning algorithms such as SVM, GMM, AdaBoosting, and ECOC, as well as hierarchical and ensemble models of clusters of classifiers built on lower level features. In TREC video retrieval evaluations [5] [6], we successfully leveraged the multimodal integration of audio and video features to obtain more accurate classification analysis for broadcast video materials. The infrastructure will support similar experimentation with field data in the nursing home environment. Some features like “anger” will have a visual component that is difficult to automatically distinguish but which may often be marked with vocal yelling. Other features like walking will have visible motion but may be silent. Still other features like identifying talking patients have both aural and visual cues.

We have demonstrated modest success to date through a combination of new and established techniques in the areas of (1) activity analysis, using articulated motion modeling and motion segmentation combined with tracking [7] [8]; (2) analyzing social interaction patterns [9]; (3) dining activity analysis, using hidden Markov models [10]; (4) detecting unusual events automatically [11]; and (5) providing privacy protection through automated person elimination [12], for those not consenting to be participants.

5 Improved Infrastructure for Future Research

To overcome the limitations discovered through these initial deployments, we propose to instrument a participating nursing home with specialized high resolution cameras of sufficient type and numbers to achieve reliable patient tracking and identification as well as to permit experimentation in fine-grained behavior analysis.

Beyond more and better cameras, the improved infrastructure will implement lossless compression of video data. Lossless compression is needed because we cannot anticipate exactly which features may be useful to future image understanding techniques. Any lossy compression scheme will necessarily alter and potentially obscure some image features. Lossless compression will increase storage requirements by almost two orders of magnitude over what is currently available in the CareMedia infrastructure. We will also move from uniform consumer grade cameras to four different specialized camera configurations located to achieve four objectives, as shown in Figure 1:

1. Reliable automatic patient identification
2. Reliable patient tracking
3. High quality, broad area video capture, and
4. High quality, task-specific behavior capture.

Reliable patient identification. As discussed in the previous section, both camera resolution and the orientation of the face towards the camera will improve automatic face identification. To increase the number of pixels captured per face as well as to increase the probability of a full frontal view, we propose to deploy ultra high resolution (1600 by 1200 pixel resolution) color IEEE 1394 video cameras (e.g. Point Grey Research's Scorpion, capturing at 14 frames per second) facing hallway and doorway entrances. Each unit represented in Figure 1 will be deployed as a two camera stack, one camera at approximately eye level for a standing patient and one at eye level for a patient confined to a wheelchair. Eight cameras will be focused on the same plane yielding at least 128 pixels per face and a more optimized camera/face orientation.

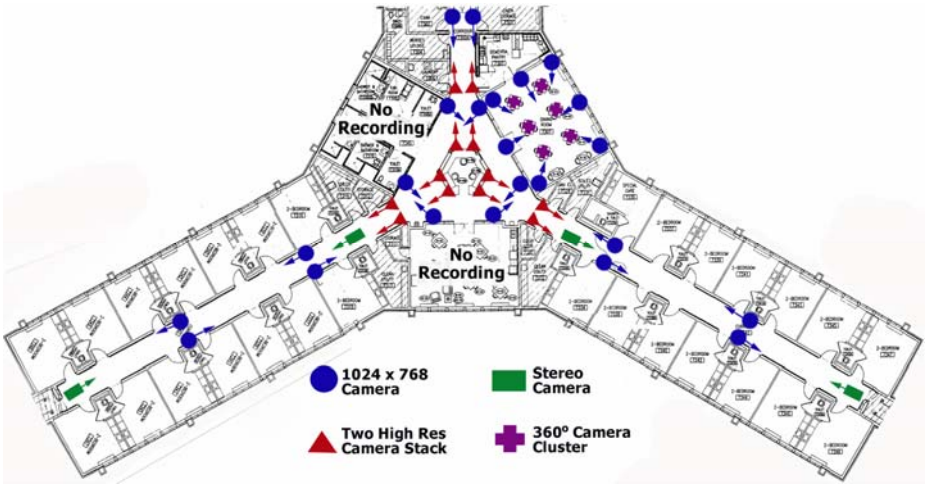

Fig 1. Proposed deployment of improved infrastructure

In future years of the project, we will expand the infrastructure to include other promising technologies for recording clinical behavior, including people identification in the thermal infrared spectrum. In comparison to the images in the visible band, which are formed primarily due to reflection, the choice of infrared makes the system less dependent on external light sources and more robust with respect to incident angle and light variation. Investigation has been done on face recognition [13] [14] [15] [16]. Up to 89% recognition rates were reported by using a commercial long-wave infrared camera. Furthermore, some sicknesses are shown to be easier to characterize in the thermal infrared spectrum domain than the visible spectrum domain. A longer term objective would be to identify not only the individual person but also his or her current health state.

Reliable patient tracking. Patient tracking will be achieved by a combination of two camera types, a 1024 by 768 (760P HDTV) 30 frames per second (fps) progressive scan camera (e.g., Point Grey Research's Flea) and a stereo camera consisting of two 1024 by 768 color 15 fps progressive CCD imagers, pre-calibrated for lens distortions and aligned within 0.05 pixel RMS error (e.g., Point Grey Research's Bumblebee). A single stereo camera provides tracking accuracy between 0.1 inch to 2 inches in the range of coverage in the hallway, as shown in Figure 1.

High quality broad area video capture. Broad area video capture at a quality suitable for both clinical observations and automatic analysis of large grain behavior analysis (e.g., walking and social interaction) will be achieved through distributed 1024 by 768, 30 fps, progressive scan cameras.

High quality, task-specific behavior capture. Task-specific behavior capture might focus on eating, social, and recreation activities occurring around tables in the dining/recreation room. To achieve the necessary 360 degree coverage and resolution, both temporal and spatial, for machine behavioral analysis for such targeted activity, the proposed infrastructure will employ a six-camera cluster, each

with a resolution of 1024 by 768 15 fps progressive scan (e.g., Point Grey Research's Ladybug, which implements a data transfer between the camera unit and PC over a proprietary 1.2 Gbps optical link).

Our proposed infrastructure will employ lossless compression to preserve image quality necessary for clinical use and machine understanding. Typical lossless video codecs achieve compression rates of between 2 and 5 depending on background complexity and object movement. For the purposes of data rate calculation we assume a mean compression of a factor of 3.

Data storage requirements will be extraordinary. Each camera has a data rate of approximately 283 Mbps with the exception of the cluster camera, which has a data range of 1.2 Gbps. In the representative configuration shown in Figure 2, there are 14 broad area cameras covering each hallway, 3 broad area cameras covering the entrance, 24 extremely high resolution cameras for face identification covering the hallways and entrance, 8 broad area cameras in the dining/recreation room, 6 broad area cameras in the core, and 6 six element cluster cameras in the dining/recreation room.

The hallway and entrance cameras will be processed in real-time to detect motion, only recording video when there is motion within the field-of-view. Our prior experience indicates that during approximately 20% of a 16-hour day (non-sleep hours) these 17 cameras will record video. In an ideal world, the 24 facial identification cameras would collect a single still image for each person passing through the threshold. Unfortunately, the orientation of peoples' faces will vary as they walk. To afford us the best chance of capturing a suitable image, we will capture video at a distance of 10 feet either side of an ideal image plane. The time people will be recorded will vary greatly from fast walking staff to extremely slow moving patients. Analysis of our prior data suggests these cameras will record approximately 5% of a 16-hour day.

Activity is almost continuous in the core/nursing station and the dining/recreation room for 16 hours per day. We have observed times when there are a small number of patients in the room and are all sleeping in their chairs. This is clinically important data that would be lost if we did not continuously record in this area. However, during the evening these cameras record video only when motion is detected. The data requirement for the cluster camera is so large that we will use them only during meal times, 3 hours per day. Assuming a factor of three lossless compression and the recording assumptions above, the 61 camera systems we propose will generate approximately 22 terabytes of data per day.

We will instrument each hallway and the core/nursing station with a PC based audio server that will digitize and store the data from independent microphones in each zone. Eight microphones will cover each zone. The dining/recreation room will be served by two PC audio servers, each digitizing and storing the data from eight coupled microphones configured as two orthogonal phased arrays. Data will be collected only when audio is detected. To improve source location reconstruction,

audio data will be captured at 96,000 24-bit samples per second from each microphone. Our prior data indicates that each microphone will generate approximately 16.6 GB of uncompressed data per day or .664 TB per day for all forty microphones. Since this is small by comparison to the video and more importantly, phase information is necessary for source analysis, no compression will be used.

6 Facilitating Research and Clinical Analysis with Advanced Infrastructure

The proposed improved infrastructure is designed for reusable deployment into natural environments with performance assessments on meaningful tasks. In order to make the difficult clinical research problems tractable, we designed the infrastructure to support the automatic assessment of the following activities which are acknowledged as quality of life indicators for the nursing home patient [17] [18]: residents' mobility (how much they walked, and also pace, gait, arm swing, other attributes of motion); eating behavior; social interactions; indicators of physical impairments; and expressions of positive and negative affect. Detecting small perturbations and charting gradual behavior differences over time are of particular interest in effectiveness studies focusing on interventions in the nursing home.

6.1 Person Tracking

The goal of person tracking is to accurately follow a moving human shape while it is visible, despite changes in lighting, shadows, occasional occlusions or transitions through the fields of view of different cameras [19] as illustrated in Figure 2. One of the primary limitations in the current CareMedia infrastructure was the use of too few cameras, such that each patient was often in the field of view of only a single camera. We were able to automatically track individual mobility using background subtraction and mean-shift tracking algorithms within the single camera, and could tell when and where a motion (usually human activity) took place as long as the lighting condition did not change dramatically [12]. Since the cameras were uncalibrated and each individual camera covered an isolated area, the performance of the people tracking system was not satisfactory, and the system failed in the case of occlusions, which were prevalent when multiple people were present, e.g., in the dining hall.

Fig. 2. An example of tracking and shadow removal, (a) original video, (b) background subtraction and noise removal, (c) shadow removal, (d) final tracking result

While this is typical for tracking a few people in an uncluttered environment, multiple high-resolution cameras [20] [21], preferably with overlapping fields of view, are necessary to achieve accurate tracking in real-world environments, where the algorithms cannot be tuned to the current background or the number of people in the scene, the changing lighting conditions, or the color distinctiveness of the clothes that people happen to wear. Using the improved infrastructure, we will employ multiple calibrated cameras to cover a larger, heavily used area with overlapping perspectives [22]. These high-resolution cameras will form a camera network that covers the area from a few different view angles. With this infrastructure, tracking failures caused by occlusions can be mostly recovered from our newly developed algorithm based on sequential Monte Carlo methodology [23].

6.2 Person Identification

Beyond tracking a human, it is critical for most applications that the person being tracked is identified, such that the same person can be recognized over time and across different situations. Given the current CareMedia camera resolutions and coverage, we are only able to achieve less than 30% accuracy of identifying an individual, given a sample image of that person. When a human supplied 20 additional images of the individual in varied poses, accuracy approached 80%, under the assumption that the person did not change clothing [24]. However, color appearance by itself is not sufficient to identify people since a person may have different color clothes, while two individuals may wear the same color at the same time. However, via person tracking, color information can be used to maintain an identification once it is obtained by face and gait identification. The current image resolution is insufficient for face recognition, and even face detection – a much simpler task – is very errorful due to the distance of the camera from the person and the ceiling-mount viewing angle. Better face detection and alignment is expected from high-resolution video in the new infrastructure, which is a crucial factor of face recognition [25]. Data from camera networks can be used to more accurately identify people through multiple pose face models and from video sequences [28]. Spatial constraints among cameras physical locations can be integrated into both the face modeling process and the identification process.

People can also be identified visually through their gait [27]. Spatial constraints among cameras' physical locations can be integrated into both the face modeling process and the identification process [28]. The infrastructure will support research into utilizing multimodal cues (speaker identification, speech localization, gait, posture, motion characteristics, facial appearance, and other temporal behavior cues) to make the identification more robust. Ultimately we would target 95% person recognition accuracy in this type of environment, which has been shown to be within reach for high quality data from multiple views [23].

6.3 Gross Motor Behavior

Clinical observational research builds from being able to measure and detect changes in behavior. In our initial deployment, we could not reliably measure walking distance and speed of a patient because we could not generate three-dimensional information based on multiple camera views. We can capture individuals' activities in a larger area using the proposed camera network infrastructure. Important parameters of mobility, such as walking distance and speed, are extracted in real 3D space rather than 2D estimations [23].

Still, in order to compute residents' mobility, several challenging problems will have to be addressed, including how to compute reliable motion statistics from non-rigid motion [29], and how to identify body regions to localize the region of activity (e.g., hands or legs). Subspace flow techniques[30] could be explored to compute motion vectors from noisy video sequences. Subspace methods for flow computation have been recently introduced and proven to have better accuracy when computing flow estimates. In order to account for outliers, robust subspace techniques [31] could be used to provide more accurate motion estimates. The infrastructure will enable research for deriving statistics on residents' mobility, including time and instances walked, pace/rate of motion, and gait. Given multiple silhouette views, a challenging problem will be detecting human violence such as fighting and kicking [32] [33].

6.4 Small Motion Behavior

In our current environment, we use a head-hand model to analyze the eating behavior of an individual patient who sits at some acceptable location within the single surveillance camera's view [7]. Average motion directions and magnitudes in the model specified regions are estimated using optical flow, as shown in Figure 3.

Fig. 3. Estimated region motion vectors for each moving body region. The red arrows are computed motion vectors; the blue masks indicate the segmented moving regions.

The system works at limited data sets, but restricts the individual to sit at only certain locations; that is, the infrastructure placed undue restrictions on when eating behavior could be automatically assessed. Hence, eating behavior could only be analyzed for those individuals seated in front of the camera, while all other individuals remained unstudied [12].

With higher-resolution camera networks and panoramic cameras on dinner tables, we will be able to capture more detailed information on patients' eating, such as eating patterns and how much food is left on a plate anywhere in the dining room. This makes it possible to quantitatively study patients' eating habit and changes, tracked temporally to other data recorded in the environment. Research into quantifying the amount of food that the patient is moving toward his/her mouth through the rate and duration of the eating motions [8] and characterizing hand and upper body motions [34] can be conducted. Such research could explore the application of motion vectors and stochastic hand tracking techniques based on color and particle filtering algorithms [35] [36] to particular eating behaviors. Several initialization methods could be tested based on learning statistics of the skin color and hand shape from multiple views. The principles of this infrastructure could then extend to the analysis of small motions to help improve physical therapy and assessments of functional occupational movement.

6.5 Fine Motion Behavior

Much interesting human behavior manifests in minute movement, e.g., twitches, tremors, eye-blinking, wrinkled foreheads and frowns. These motions are crucial cues to detect early stage physical impairments. CareMedia could not address fine-grained attributes like tremors because of a lack of visual resolution, limiting its scope instead to detecting audio indicators of affect and focusing on identifying disruptive vocalizations as an indicator of negative affect. Visual indicators are not feasible in CareMedia because the resolution of the cameras is too low for even human inspection to identify facial attributes of affect like smiles and wrinkles in the forehead and around the eyes.

The face has been repeatedly validated as one indicator of emotion, even across cultures and race [37]. Unique patterns of facial musculature have been demonstrated for specific emotions such as joy, disgust, and anger [38]. Other cues to emotional states include voice quality [39], and body position and movement [40]. Our infrastructure will be capable of delivering the resolution necessary for research into detection and classification of facial and body cues.

High-resolution imaging allows researchers to model different subtle non-rigid motion patterns of the body parts in contrast with the primary motion of the body. For example, physical impairments can then be detected by high frequency of the local motions [41] or asymmetric spatial motions. With high-resolution imaging and audio techniques, we will be able to begin to exploit the characteristics of the positive/negative affects of human activities in their subtle behavior pattern and detail facial information. Thus, the report to the caregiver in the nursing home might note that a patient had increased his eyeblink rate to 50 blinks per minute, and had developed small tremors in his left hand at 8 hertz.

6.6 Social Interactions

Interaction with others is generally considered a positive and necessary part of our daily life. Naturally, the level of social interaction a person has can depend on a wide range of factors, such as his/her health, his/her personal preference, and aptitude for social interaction. We can identify social interactions by analyzing individuals' activities. This requires robustly tracking people and accurately detecting individual events. The current CareMedia system has little ability to detect and analyze complex social interactions. We are able to detect only simple events of general human activities using video and audio cues, because there is no three-dimensional information reconstruction possible from multiple camera views. Essentially, we are only able to detect the proximity of two tracked persons [12]. For current quality observational video data, we use a background subtraction technology and shape analysis to detect frames that contain human activities, together with acoustic events using features from the audio.

Individual activities and interactions are classified using tracking and color/shape analysis. These algorithms work well only when the individual person is easily separable from others and from the background. In principle, camera networks will allow us to analyze individual human activities and some typical social interaction patterns. We have tested these basic ideas on simulated data recorded in university laboratory settings. By using multiple-camera tracking technology and color/shape property analysis, we are able to understand some typical activities and interaction patterns and produce semantic descriptions [9] [42]. The system is able to detect more complex social interaction patterns with multiple higher resolution cameras. The system will be able to detect a large scale of social interaction patterns, from a waving hand at a distance to two people shaking hands. Given the segmented silhouette of several people derived from the new infrastructure's captured data, the identified faces and the detection of the mouth area, temporal classifiers can now be trained to recognize if the people are talking based only on the visual data. Such research is especially interesting in noisy environments, where microphones will not provide enough information to infer if a person speaking, and localize the source.

Another important aspect of social interaction will be to determine which patients exhibit frequent social interactions and with whom. To solve this problem, we anticipate that researchers will rely on detailed face and body recognition approaches and analyzing data across time. Determining the type of contact between people and the affect characteristics is a challenge, especially detecting violent situations, which is of great interest to caregivers in a nursing home. Doctors want to know that, on a new medication, a patient previously withdrawn and suffering from chronic pain is now smiling while talking to fellow patients, and frequently shaking hands.

6.7 Privacy Protection

In many surveillance situations, a large amount of video gets recorded, but depending on the local laws and the perspective of the camera view, as well as the nature of the

setting that is being recorded, researchers must secure the consent and permissions of the recorded persons to use that video data. Furthermore, some people who permit data capture choose not to have their recordings become part of an analyzed and archived collection.

Especially in the geriatric nursing home setting, our video observations record patients, nurses and health professionals as well as visitors. In our first study about 70% of the patients (or their legal representatives) gave permissions to use the video for research; the remaining patients did not consent. The staff also consented to being observed, but their behavior is not the focus of our study. However, while visitors were notified through signs that recording was taking place, their explicit consent for re-use of the video was not obtained.

After the person identification step described above, we can decide which individuals are consenting participants. Those who are not need to have their images rendered unidentifiable and their audio rendered unintelligible in the permanent record [43]. The privacy protection in video data may be implemented in several forms: hiding the face, blurring the person, making the person half-transparent and completely concealing the person. In our current implementation, we use the last type of protection by directly substituting the foreground regions with the background. One example of privacy protection is shown in Figure 4.

Fig 4. Privacy protection: (a) original video,
(b) blocking the person on the right to protect her privacy

6.8 Gradual Trends Over Time

A change in behavior may be one of the first signs of illness. For example, Alzheimer's disease causes the death of large numbers of brain cells over a period of time; this results in the mental deterioration, dementia, and eventual death of the afflicted individual [44]. Yet a great many changes have gone on in the brain of that person, corresponding to a gradual change in personality. An AD patient loses basic abilities in a similar order to that in which the abilities were learned: handling finances, choosing clothes, bathing oneself, control over bodily functions, feeding oneself, speech, the capacity to walk, and even the capacity to sit up [45]. Detection of gradual change in behavior requires a long term observation. Currently, observational period are not long enough, with too many gaps in observations to track

gradual changes over time. These gradual changes include effects of medication on behavior that can be correlated and documented. The storage capacity in the CareMedia infrastructure is not sufficient to provide the long-term observation necessary for clinical determination that a behavior has changed. In long-term care facilities, physicians modify the dosages of psychotropic drugs periodically. They rely on observational records to determine the optimal pharmacological interventions. Thus, we need to be able to record a window of at least one month of behavior to observe the effects of an altered medication. With the proposed infrastructure, we will be able to automatically record and detect gradual behavior pattern changes of each patient through a long-term observation. To this end, statistics based on different time periods will be performed in wavelet feature space.

Longitudinal visual data also enables the establishment of baseline or "normal" behavior patterns for each patient. The research focus can then be identifying differences from the baseline, and tagging that for inspection by a trained clinical rater, with a more ambitious goal being the automatic tagging of differences in baseline using medically accepted rating scales. Such research will likely center attention on the face, voice, body, eyes, and touch, as these are called out as the obvious locales for observing emotion in AD patients [46]. Hence, our infrastructure needs to capture high resolution, fine-grained data. Changes in the amount or rate of walking, gait, arm swing, and other attributes of motion from the baseline should also be made easy to flag for professional review and diagnosis through the recording infrastructure, as such changes are key indicators to the benefits and possibly detrimental side effects of pharmacological interventions.

6.9 Rare Events

Infrequent abnormal events may be unobserved and thus omitted by care providers in their reports. However, abnormal events are associated crucially to early disease detections, help requirements, and dangerous alarms. The current capacity in the CareMedia storage systems is not sufficient to provide the long-term, detailed observation necessary for research into detection of low-frequency events. Current observational periods are not long enough to capture infrequent behavior and there are too many locations not visible to cameras. In addition, the analysis cannot identify the small behaviors that are indicative of the events. A comprehensive record captured by the proposed infrastructure supports research with semi-supervised machine learning methods to detect and characterize both known and unknown rare events.

7 Conclusions

Current machine understanding of aural and visual data recorded in the clinical nursing home environment is limited by incomplete data with inadequate resolution. By implementing an advanced audiovisual recording environment with comprehensive-coverage, but not yet economically practicable in most commercial facilities, the capabilities of machine understanding of audiovisual data can be better

assessed and improved, allowing for categorizing and dynamically summarizing antecedents and consequences of behavior temporally aligned with various interventions. The requirements of the described infrastructure are based on prior field studies, thousands of hours of accumulated video data analysis and recent algorithm development. This infrastructure supports the generation of information that limits redundancy and highlights episodes useful for evaluating a person's quality of life while addressing privacy and policy restrictions.

8 Acknowledgements

This material is based on work supported by the National Science Foundation (NSF) under Grant No. IIS-0205219.

References

1. Wactlar, H., Bharucha, A., Stevens, S., Hauptmann, A., Christel, M. "A System of Video Information Capture, Indexing and Retrieval for Interpreting Human Activity," Third International Symposium on Image and Signal Processing and Analysis (ISPA'03), Invited paper for the Special Session on System Perspectives in Information Retrieval, Rome, Italy, September 18-20, 2003.
2. Allin, A.J., Atkeson, C.G., Wactlar, H., Stevenson, S., Robertson, M.J., Wilson, D., Zimmerman, J., and Bharucha, A. "Toward the Automatic Assessment of Behavioral Disturbances of Dementia," *Fifth International Conference on Ubiquitous Computing (UbiComp'03), 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications*, Seattle, WA, October 12-15, 2003. http://www.healthcare.pervasive.dk/ubicomp2003
3. Wang, J., Zhang, C., Shum, H. "Face Image Resolution versus Face Recognition Performance Based on Two Global Methods," Proceedings of Asia Conference on Computer Vision (ACCV'04), 2004.
4. Chen, M.-Y., Hauptmann, A. "Toward Robust Face Recognition from Multiple Views," International Conference on Multimedia and Expo (ICME'04), Taipei, Taiwan, June 27-30, 2004. http://www.informedia.cs.cmu.edu/documents/ChenICME04Face.pdf
5. Hauptmann, A., Baron, R., Chen, M.-Y., Christel, M., Duygulu, P., Huang, C., Jin, R., Lin, W.-H., Ng, D., Moraveji, N., Papernick, N., Snoek, C.G.M., Tzanetakis, G., Yang, J., Yan, R., Jin, R., and Wactlar, H. "Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video," Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference), Gaithersburg, MD, November 17-21, 2003.
6. "NIST TREC Video Retrieval Evaluation," TRECVID, Twelfth Text Retrieval Conference, Gaithersburg, MD, November 17-21, 2003. http://www-nlpir.nist.gov/projects/trecvid
7. Gao, J., Hauptmann, A., and Wactlar, H. "Combine Motion Segmentation with Tracking for Activity Analysis," The 6th International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May 17 - 19, 2004, IEEE.
8. Gao, J., Wactlar, H., Hauptmann, A., Collins, R.T. "Articulated Motion Modeling for Activity Analysis," Third International Conference on Image and Video Retrieval (CIVR'04), Workshop on Articulated and Nonrigid Motion (ANM'04), Dublin City, Ireland, July 21-23, 2004. http://www.informedia.cs.cmu.edu/documents/cvpr04_anm_33.pdf

9. Chen, D., Wactlar, H., Yang, J. “Towards Automatic Analysis of Social Interaction Patterns in a Nursing Home Environment from Video,” 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'04), New York, New York (held in conjunction with ACM Multimedia 2004), October 15-16, 2004.
10. Gao, J., Wactlar, H., Bharucha, A., Hauptmann, A. “Dining Activity Analysis Using a Hidden Markov Model,” 17th International Conference on Pattern Recognition (ICPR'04), Cambridge, United Kingdom, August 23-26, 2004.
11. Zhonh, H., Shi, J., Visontai, M. “Detecting Unusual Activity in Video,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), Washington, DC, June27-July 2, 2004.
12. Hauptmann, A., Yang, J., Qi, Y., Yan, R., Gao, J., Wactlar, H. “Automated Analysis of Nursing Home Observations,” IEEE Pervasive Computing, Special Issue on Pervasive Computing for Successful Aging,, 3(2): pp.15-21, April-June. http://www.informedia.cs.cmu.edu/documents/IEEE-PC04_AutoAnalysisofNHO.pdf
13. Prokoski, F., Reidel, E. “Infrared Identification of Faces and Body Parts,” BIOMETRICS: Personal Identification in Networked Society. Kluwer Academic Publishers, 1998.
14. Prokoski, F. “Current, Status and Future of Infrared Identification,” Proceedings of IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications, Hilton Head Island, South Carolina, USA, 2000, pp.5-14.
15. Srivastava, A., Liu, X., Thomasson, B., Hesher, C. “Spectral Probability Models for IR Images with Applications for IR Face Recognition,” *Proceedings of IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, Kauai, Hawaii, USA, 2001.
16. Buddharaju, P., Pavlidis, I., and Kakadiaris, I. “Face Recognition in the Thermal Infrared Spectrum, CVPR 2004,” *Computer Vision Pattern Recognition Conference (CVPR'04)*, Washington, DC, June 27-July 2, 2004.
17. *Assessing Quality of Life in Alzheimer’s Disease*. Ed(s) Albert, S.M. and Logsdon, R.G. Springer, New York, 2000.
18. Volicer, L. and Bloom-Charette, L., ed(s). “Enhancing the Quality of Life in Advanced Dementia.” Taylor & Francis, Philadelphia, 1999.
19. Hu, W.T., T.; Wang, L.; Maybank, S. “Survey on Visual Surveillance of Object Motion and Behaviors,” IEEE Transactions on Systems, Man and Cybernetics, 34(3): pp.334-352, August, 2004.
20. Zhao, T.a.N., R. “Tracking Multiple Humans in Complex Situations,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9): pp.1208-1221, 2004.
21. Patil, R., Rybski, P., Kanade, T., Veloso, M. “People Detection and Tracking in High Resolution Panoramic Video Mosaics,” *In Proceedings of IROS'04*, Sendai, Japan, September, 2004.
22. Khan, S., Javed, O., Rasheed, Z., Shah, M. “Human Tracking in Multiple Cameras,” ICCV01(I: 331-336).
23. Chen, M.-Y., Hauptmann, A. “Toward Robust Face Recognition from Multiple Views,” *International Conference on Multimedia and Expo (ICME'04)*, Taipei, Taiwan, June 27-30, 2004. http://www.informedia.cs.cmu.edu/documents/ChenICME04Face.pdf
24. Yan, R., Yang, J., and Hauptmann, A. “Automatically Labeling Video Data Using Multi-class Active Learning,” International Conference on Computer Vision 2003 (ICCV'03), 2003, pp. 516 - 523. http://www.informedia.cs.cmu.edu/documents/ICCV03activeLabeling.pdf
25. Chen, X., Yang, J., Waibel, A. “Calibration of a Hybrid Camera Network,” Proc. International Conference on Computer Vision (ICCV 2003), pp. 150-155.
26. Everingham, M.a.Z., A. “Automated Person Identification in Video,” Proceedings of the Challenge of Image and Video Retrieval (2004), Dublin City, Ireland, July 21-23, 2004.

27. Tolliver, D., Collins, R. “Gait Shape Estimation for Identification,” Fourth International Conference on Audio and Video-Based Biometric Person Authentication, University of Surrey, Guildford, UK, June 9-11, 2003.
28. Yan, R., Hauptmann, A., Yang, J., Zhang, J. “A Discriminative Learning Framework with Pairwise Constraints for Video Object Classification,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, June 27-July2, 2004. http://www.informedia.cs.cmu.edu/documents/CVPR04_yan.pdf
29. Bobick, A., Tanawongsuwan. “Performance Analysis of Time-distance Gait Parameters Under Different Speeds,” 4th International Conference on AVBPA, June, 2003.
30. Irani, M. “Multi-Frame Optical Flow Estimation Using Subspace Constraints,” IEEE International Conference on Computer Vision (ICCV), Corfu, September, 1999.
31. De la Torre, F. and Black, M. J. “A Framework for Robust Subspace Learning,” *International Journal of Computer Vision*, 54(1-3), August-October, 2003, pp. 117-142.
32. Nam, J., Alghoniemy, M., Tewfik, A.H. “Audio-Visual Content-Based Violent Scene Characterization,” IEEE International Conference on Image Processing (ICIP'98), volume 1, 1998, pp.353-357.
33. Datta, A, Shah, M., and Da Vitoria, N. “Person-on-Person Violence Detection in Video Data,” *IEEE International Conference on Image Processing*, Rochester, NY, 2002.
34. Gao, J., and Shi, J. “Inferring Human Upper Body Motion,” IEEE Computer Vision and
35. Isard, M. and Blake, A. “Contour Tracking by Stochastic Propagation of Conditional Density,” Proceedings of the European Conference on Computer Vision, volume 1, Cambridge UK, 1996, pp.343-356.
36. Doucet, A., de Freitas, N., and Gordon, N. “Sequential Monte Carlo Methods in Practice.” Springer-Verlag, 2001, ISBN 0-387-95146-6.
37. Ekman, P., Friesen, W.V., O’Sullivan, M. and Chan, A. “Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion,” Journal of Personality and Social Psychology, 53, 1987, pp.712-717.
38. Ekman, P., and Friesen, W.V. “Facial Action Coding System”. Consulting Psychologists Press, Palo Alto, CA, 1978.
39. Caporael, L.R., Lukaszewski, M.P., and Culbertson, G.H. “Secondary Baby Talk: Judgments by Institutionalized Elderly and Their Caregivers,” Journal of Personality and Social Psychology, 44, 1983, pp.746-754.
40. DeLong, A.J. “The Microspatial Structure of the Older Person,” in *Spatial Behavior of Older People*, ed(s) L.A. Pastalan and D.H. Carson. University of Michigan Institute of Gerontology, Ann Arbor, MI, 1970, pp. 68-87.
41. Wenzelburger, R., Raethjen, J., Loffler, K., Stolze, H., Illert, M., Deuschl, G. "Kinetic Tremor in a Reach-to-grasp Movement in Parkinson's Disease," Movement Disorders, 15: pp.1084-1094, November, 2000.
42. Chen, D., Malkin, R., Yang, J. "Multimodal Detection of Human Interaction Events in a Nursing Home Environment," Sixth International Conference on Multimodal Interfaces (ICMI'04), State College, PA, October 14-15, 2004.
43. Xiao Y, Hu P F-M, Seagull JF, Mackenzie CF. Distributed planning and monitoring in a dynamic environment: trade-offs of information access and privacy. 2003 Proceedings of IEEE International Conference on Systems, Man, and Cybernetics; 2003: 4141-46
44. Family Caregiver Alliance, National Center on Caregiving, “Fact Sheet: Alzheimer's Disease,” http://www.caregiver.org/caregiver/jsp/home.jsp, 2004.
45. Dementia (Alzheimer's): Mental Health of the Elderly, http://www.mentalhealth.com/p20-grp.html, 2004.
46. Lawton, M.P., Van Haitsma, K., and Perkinson, M. “Emotion in People with Dementia: A Way of Comprehending Their Preferences and Aversions,” chapter 5 in Interventions in Dementia Care Toward Improving Quality of Life, ed(s) M.P. Lawton, PhD, and R.L. Rubinstein. Springer, New York, 2000.