

# 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.*

Our society is increasingly aging. The US Census reports that the population over age 85 will triple from 4.5 million in 2003 to 14.2 million by 2040. The ratio of retired adults to working adults will grow over 50 percent by 2040. Five percent of Americans over 65 currently reside in nursing homes, with 20 to 50 percent of those over 85 expecting to be placed in a nursing home at some point in their lives.<sup>1,2</sup>

At Carnegie Mellon University, the CareMedia project is using information technology to assist this growing segment of the US population and their caregivers.<sup>3</sup> Currently, in skilled-nursing facilities physicians might see a patient for only a few minutes once a week. The assessment of a patient's progress is based on staff reports that, owing to time and personnel constraints, might have resulted from few actual observations of the patient.

A critical element in long-term patient care is an accurate account of the patient's physical, behavioral, and psychosocial functioning.<sup>4</sup> As direct behavioral observation becomes indispensable to data gathering and treatment planning, we're developing technologies that can automatically capture and analyze all that they hear and see,<sup>5</sup> with the potential for significantly affecting clinical care. For example, early recognition of gait

instability might help reduce the tremendous morbidity and excessive economic burden associated with unwitnessed falls. More accurate, complete behavioral logs would also facilitate better use of psychotropic medications.

## The core technology problems

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 pro-

Alexander G. Hauptmann, Jiang Gao, Rong Yan, Yanjun Qi, Jie Yang, and Howard D. Wactlar  
Carnegie Mellon University

hibits 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 can 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. So, the caregiver can still review a comprehensive record of important patient activities in a short time period.

### Experimental setup

For an initial feasibility investigation, we mounted four cameras and microphones in the dementia unit of a Pittsburgh-area nursing home, where residents' mean age was 89. Recording went on for eight hours a day over one week from four viewpoints: one in the dining room, one in the hallway, and two in the television room. The medical staff prohibited any wearable sensors as well as devices that couldn't be concealed from

patients or that required patient cooperation. We took special care to protect the privacy of the participating residents so that names would remain confidential and faces unidentifiable in publicized material.

Manual analysis of a portion of the data, which we accomplished by reviewing a 10x speeded-up version of the video wherever the system detected any motion, took several months. Patients spent at least 13.6 percent and at most 24.6 percent of their time in the recorded spaces. Interpersonal interactions covered less than 20 percent of that time. Surprisingly, meals accounted for over 75 percent of interpersonal interactions. Additionally, unwitnessed by the staff, two subjects successfully escaped from their locked unit on six of 11 attempts behind unsuspecting staff or visitors.<sup>6</sup>

### Tracking people

We began our observational analysis by tracking people on the recorded video. After studying the three main approaches to tracking—that is, temporal differencing, optical flow, and background subtraction—we adopted background subtraction<sup>7</sup> because the nursing home's background was stable and the cameras were fixed.

We first classify each pixel as either foreground or background given a background pixel  $B$  and an absolute difference threshold  $D$ . Pixel  $P$  is background if  $|P - B| < D$ ; otherwise, it's foreground. Two kinds of errors still exist: confusions due to similar-looking foreground and background pixels and fragmented regions due to background noise. We use a noise removal and region-growing technique<sup>7</sup> to remove these inconsistencies. Then, a low-band pass filter smoothes the object boundary and filters background noise.

Tracking provides the spatial boundaries for moving objects, while event detection encodes the history of previ-

ous tracks. Our framework combines them to make the tracking algorithm more robust to transient visual noise and occlusions. We define an event  $e$  as a video sequence of a tracked object beginning when the system starts tracking the object and ending when it can no longer track the object. Typically, an entry-to-exit sequence for a person constitutes an event. However, we may also identify as one event a sequence from a person's entry up to his or her merging with a group.

Consider a video frame at time  $t$ . To find the correspondence between  $N$  extracted blobs  $B_i$  and  $M$  potential tracking paths  $E_j$ , we must first evaluate a matching matrix, defined as  $M^t(i, j) = f(B_i, E_j)$ , where the matching function  $f(B, E)$  is 1 if  $B$  corresponds to event  $E$ , or 0 otherwise. Effectively, the matching function is a distance metric between the blob features

$$f(B_i, E_j) = \delta(\text{Dis}(B_i, E_j) - \text{threshold}),$$

where  $\delta$  is the Kronecker delta function. Our system applies a first-order motion model to estimate the object's location in the next frame. The current feature set that characterizes a blob consists merely of the bounding-box position and the tracking region's size.

When two people stand very close together, the background subtraction algorithm segments them as one single region. Merging is undesirable because it will interrupt the tracking history. Frequently, if the overlap is incomplete, resplitting the merged people into separate regions is still possible using a mean-shift tracking algorithm<sup>8</sup> based on color or texture similarity, which is resilient to partial occlusion. However, a drawback of this algorithm is that a region to be tracked has to be separately initiated—that is, manually identified. In our integrated solution, when the mean-shift algorithm splits a merged region, we can

Figure 1. An example of using a mean-shift algorithm to split merged people: (a) the original video, (b) merged tracking, and (c) tracking a split with mean shift.



use the regions tracked before merging to automatically initialize the new regions. The mean-shift iterations find the most similar candidate pre- and post-merger. With the aid of this event detection algorithm, our tracking approach is robust in defining events even when the objects are temporarily occluded or split into smaller pieces. Figure 1 shows an example of such a merged-object split.

The only major tracking errors derive from merged regions. To measure how often the tracker will confuse multiple persons as one region, we manually tracked events in two hours of a single-view video from the nursing home at  $320 \times 240$  pixel resolution and 30 frames per second. For this manually annotated portion, we found that someone was visible for 1,575.3 seconds, or 21.8 percent of the time. Of that time, the tracker pursued the correct number of tracks for 1,374.9 seconds, roughly 87.3 percent of the time. The rest of the time tracks were either split or merged inappropriately.

While substantial, this error rate doesn't invalidate the overall analysis of nursing home observations. The system never actually fails to track an individual's movement but only gets confused about whether this track belongs to the same individual as before or constitutes a new event, merged or split from previous tracks. So, failure at this analysis stage merely results in more distinct events that must be labeled and associated with the particular individual in the next phase.

Figure 2. Identification accuracy as a function of training examples. We identified 1,012 tracked events from 10 individuals as test data. Initially, the system randomly selected 21 labels until it obtained at least one event from every individual. It then added manually labeled training examples in increments of 5.

### Identifying and labeling individuals

Next, we automatically identify the tracked events as particular individuals. Manually labeling all the data is impossible because even for only one camera, we would need to manually label 2,592,000 frames recorded in one day.

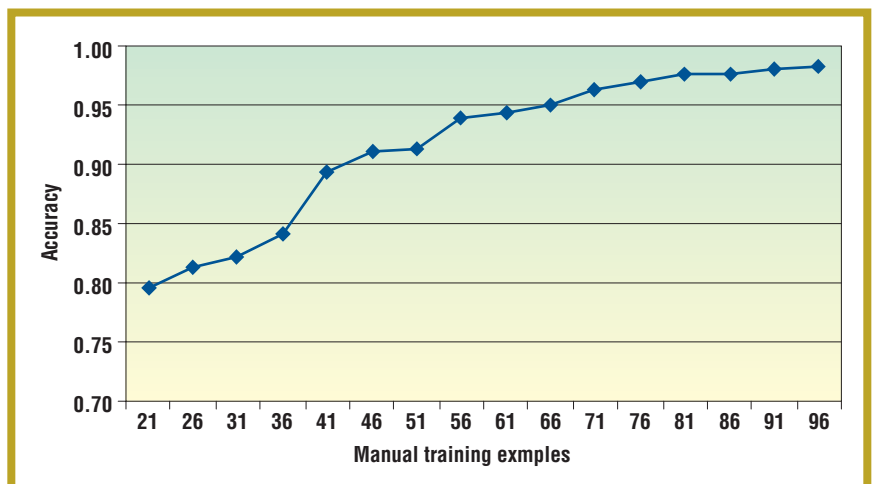
The system can't identify a person without some prior knowledge of the person's appearance. So, we try to semi-automatically identify all the events involving a single person with just a few manually labeled frames for that person. This motivated us to adapt an incremental, or active, learning framework with interaction and supervision from a human. We can formulate it as a multi-class classification problem, where each example (event) is associated with one of a given set of classes (persons). Let  $x$  denote the domain of possible examples,  $y$  be a finite set of classes, and  $k$  be the size of  $y$ . Formally, the learning algorithm takes a set of training examples

$(x_1, y_1) \dots (x_m, y_m)$  as input, where  $y_i$  is the label assigned to example  $x_i$ . Assuming that people don't change clothes during a video sequence, the color histogram is one of the most robust image features in this environment. So, we represent tracked people using a color histogram.

In experiments with this nursing home data,<sup>9</sup> the best sample selection strategy achieved a more than 50 percent error reduction over random sample selection with learning. Figure 2 shows how the number of training examples (each requiring one human click to identify the person) that the active-learning system chose increases labeling accuracy. This demonstrates that an active learner with careful sample selection can achieve remarkably good performance with only a small percentage of the potential human labeling effort.

### Analyzing activities

Given that we can follow a specified individual over time, we want to charac-



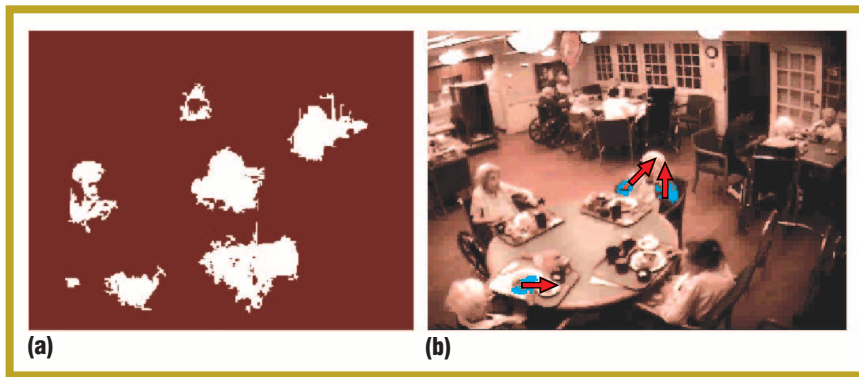


Figure 3. Examples of how the system detects eating motions: (a) regions where the system detected people’s motions at a dining room table over a period of 2 minutes. Each cluster of motions is associated with a person. (b) The motion vectors (directions) are estimated to be hand motions related to eating (in red), while the blue areas indicate the part of the body that is moving.

terize and ideally quantify what the individual is doing. We focus on two types of activities: eating in the dining room and personal hygiene in front of a mirror.

**Dining room activity**

Because the dining room was a focal point of observed interactions, we developed an automatic characterization of the main dining room activity, eating. Although tracking in the dining room was more difficult owing to occlusions and the visually cluttered background, labeling was easier because, once seated, patients didn’t move from their seats at the table, requiring only a single manual identification click.

Our algorithm for detecting “eating” has three steps:

1. Finding an individual’s detailed outline region.
2. Detecting and tracking the motion of body parts—that is, estimating motion vectors and associated moving regions.
3. Characterizing activity from the body part motion. This involves finding the person’s face and identifying dominant motions in the body area.

**Finding a person’s detailed outline region.**

The people-tracking algorithm just described gives a coarse idea of the region identified with one person. To get a finer distinction of the person’s outline, we examine the individual’s motion subspace and accumulate the moving pixels into clusters, which provide an accurate boundary of each person in the scene.

Let  $M(x, y, t)$  be a binary mask indicating all regions of motion in frame  $t$ ; that is,

$$M(x, y, t) = 1$$

indicates the pixel at  $(x, y)$  in frame  $t$  is in a moving region. Then, we obtain the individual’s regions by accumulating  $M(x, y, t)$  over time. Let

$$Cluster_{\tau}(x, y, t) = \bigcup_{i=0}^{\tau-1} M(x, y, t - i),$$

where  $\tau$  is the temporal duration. If  $\tau$  is large enough (typically around 2 minutes),  $Cluster_{\tau}(x, y, t)$  will approximate the individual person-regions at time  $t$ , as Figure 3a shows.

**Detecting and tracking body part motion.**

We developed a motion segmentation algorithm that uses RANSAC (*Random Sample Consensus*)<sup>10</sup> to find affine motion patterns of apparent motion of human body parts in video on the basis of optical flow.<sup>11</sup> This algorithm effectively detects and labels parametric motion patterns of natural human motion with normal clothing in natural environments.

Accuracy is a major problem for motion segmentation. The algorithm can partially missegment motions in various body parts; this is the main cause of false alarms. We improve the algorithm by tracking segmented regions after body part motion segmentation and filtering out temporally inconsistent motions.<sup>12</sup> The combination of motion segmentation

and tracking lets the algorithm be sensitive to subtle body motions yet remain resistant to tracking errors.

**Characterizing activity from body part motion.**

To classify specific activities, we identify two main motion components (corresponding to the left and right hands and arms) in addition to the head region. First, a face detector<sup>13</sup> finds each person’s face in individual regions. The face detection algorithm uses statistical modeling to capture the variation in facial appearances. It’s accuracy is 94.4 percent on a per-frame basis. Knowledge of the head region lets us normalize subsequent hand, arm, and body motions with respect to their distance from and orientation to the head.

We then identify the two main motion components (corresponding to the left and right hands and arms) outside the head-motion region. Because several segmented moving regions usually exist in each person’s subspace, simply using the detected motions’ magnitude to define which region is dominant doesn’t work. Our solution is to use a temporal-consistency constraint. We developed a weighted sequential-projection algorithm to detect temporally consistent motions.<sup>14</sup> The algorithm projects the current motion vector to the one in the next frame for the same region and adds inner products within several frames. The system retains only regions with a large sum of inner products, because this indicates that the region is moving consistently in a similar direction over consecutive frames. The algorithm filters out inconsistent or random motions, which are irrelevant to

our objective, and keeps only consistently moving regions.<sup>12</sup> We assign the two regions with the most consistent (dominant) motions as the left and right arm and hand motions as appropriate. If more than two regions exist in the candidate-consistent region list, we select regions that overlap most with other candidate regions.

Finally, we use the relative distances and movements between the head and the two hand and arm components to characterize the eating activity. Figure 4 shows the head-hand model. To recognize eating motions, we simply map the two hand and arm regions' motion vectors to the main axes between the head and the hands, as Figure 4 shows. We use the projected distance change on these axes to indicate a person's eating gestures.

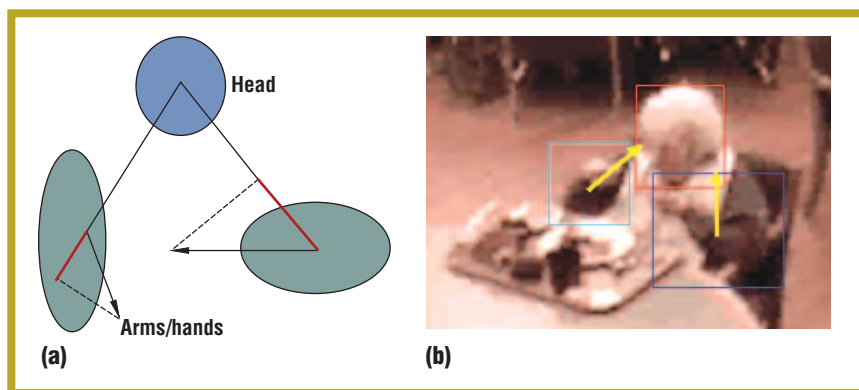
We evaluated our approach on 30 minutes of video recording 10 residents at the nursing home during lunch. Figures 3 and 4 show examples of both the segmented moving regions and their motion vectors.

We studied our approach's effectiveness by comparing the result from combined motion segmentation, tracking, and consistent motion filtering with the result using only motion segmentation (see Table 1). The combined approach appears more robust to random motions by giving fewer false alarms. Because tracking is more sensitive to subtle motions, which are hard to detect using only motion segmentation, the combined approach also improves the recall rate.

### Personal-hygiene activities

As we mentioned previously, we're also developing a system to observe what people are doing in the bathroom mirror during their personal-hygiene activities,<sup>15</sup> classify these observations into different activity types, and summarize their frequencies and durations.

Our system has learned to recognize specific activities including brushing



**Figure 4.** A head-hand model for individuals: (a) the abstract model, which only considers motions by the arm and hand either toward or away from the head, ignoring motions in directions other than the part along this head-hand axis. The red lines indicate normalized motion vectors that characterize eating motions. (b) The automatically analyzed motion from the recorded nursing-home video. The arrows are motion vectors of the two major arm and hand components. They are mapped to the head-hand axes to capture the person's eating gesture.

teeth, shaving, combing hair, face washing, and hand washing, as well as "miscellaneous activities." Owing to extreme privacy concerns regarding bathroom activities, we set up cameras to record this data privately at the homes of CMU researchers. The analyzed video contained 44 personal-hygiene activities from 11 people.

To analyze this, we first extract audio and visual features from the recorded video. The system converts the audio into mel-frequency cepstral coefficients and audio pitch.<sup>16</sup>

Because subjects are close to the camera attached to the mirror, the video has high-enough resolution to detect skin color pixels, which allows simple hand-motion tracking (see Figure 5). Our skin color model<sup>7</sup> is a 2D Gaussian model in a red, green, blue color space initialized by the detected face color. After applying

the model to each pixel, we look for connected skin areas.<sup>11</sup> Because the human shape is roughly symmetrical along its vertical axis, we use axis projection histograms to represent the outline shape at any moment. We extract other visual features as we described earlier.

Our 110 visual features include the human body shape, its relative size, its width and length, projection histograms, the magnitude and angle of the principal motion vectors, detected-face parameters, and the largest skin color objects.

After extracting audio and visual features, we classify each audio frame into one of six audio classes: Silence, Human Sound, Water, Brushing, Shaving, and Other. Meanwhile, we classify each video frame into one of six visual classes: Blank, Standing, Washing Face, Combing or Brushing Hair, Brushing Teeth, and Close\_to\_Mirror/Camera. To clas-

**TABLE 1**  
Eating activity analysis results.

	Correct detections	Misdetections	False alarms
Motion segmentation	43	13	39
Motion segmentation, tracking, and temporal consistency	50	6	9

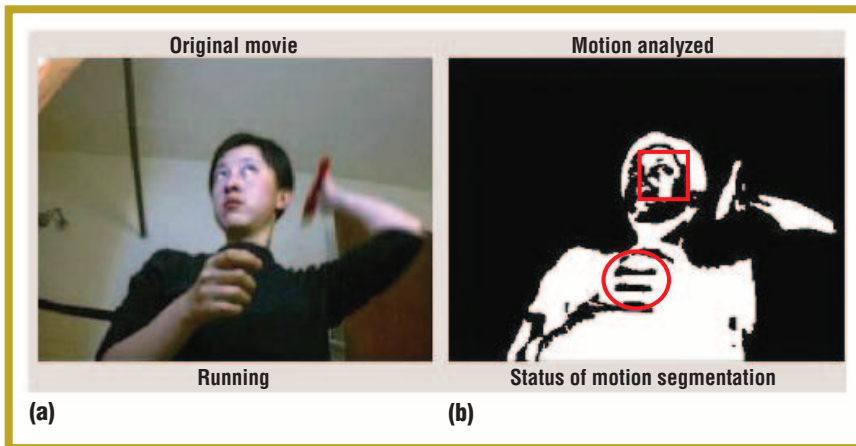


Figure 5. Visual feature extraction from the personal-hygiene video: (a) a combing activity and (b) the corresponding extracted segments. Red circles on the right show the detected skin color regions, and the red square shows the detected face.

sify an activity, we use a support-vector-machine classifier<sup>17</sup> separately for audio and imagery.

After separate visual and audio pattern classification, a secondary-level SVM meta-classifier combines the results and categorizes the current frame into one of the six hygiene activity classes mentioned earlier. At the frame unit, the system achieves an average of 55 percent precision and 53 percent recall in a ten-fold cross-validation over all six classes of hygiene activities. Performance varies widely for different activities, each recorded under dramatically different conditions in each subject's home. For example, shaving averages 70 percent precision and 73 percent recall, while brushing teeth has around 30 percent precision and 27 percent recall. These results are preliminary and we expect better performance from sequential models<sup>18</sup> that exploit such activities' temporal continuity.

The Informedia Digital Video project<sup>19</sup> demonstrated that extracted data doesn't have to be perfect to be useful. Specifically, we found that transcription error rates of 35 percent will result in only minimal information retrieval degradation compared to perfect transcriptions.<sup>20</sup> The end goal isn't perfect video analysis. Our medical experts have suggested that capturing *trend information* over time is critical for patient assess-

ments and diagnoses. So, it's not the exact number of bites the patient ate that's interesting, but rather that he or she took 50 percent fewer bites than last week. Similarly, trends in walking and personal care are interesting for seeing if a new drug has made the patient less mobile or less able to perform normal bathroom rituals. Despite analysis errors, we can capture long-term trends because independently distributed errors cancel out over many observations and long time periods.

Through CareMedia, we're demonstrating that pervasive computing with long-term observation of the elderly in nursing homes can

- Effectively track people over long periods of time
- Identify individuals in tracked events (between entry and exit) with minimal human help
- Characterize detailed human activity such as eating or personal hygiene motions

We're working with nurses and physicians at the nursing home to develop an interface that presents the results of our analyses in a format suitable to caregivers' needs. Medical caregivers, especially, feel strongly that this type of observation's potential benefits far outweigh the inherent intrusions into privacy.

This research is limited by the types of activities that can be observed and detected automatically and by the extent

that patients deem such monitoring acceptable when weighing privacy concerns against care benefits. Even this early work has made it clear that observing patients in a nursing home is feasible and can provide a meaningful information resource that supports more complete and accurate assessment and evaluation of behavioral problems for the elderly. ■

## ACKNOWLEDGMENTS

The Advanced Research and Development Activity (ARDA) partially supported this research under Contract MDA908-00-C-0037, and the US National Science Foundation partially supported it under Agreements IIS-0105219 and IIS-0121641.

## REFERENCES

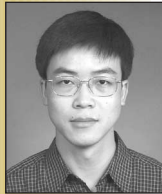
1. P.S. German et al., "The Role of Mental Morbidity in the Nursing Home Experience," *Gerontologist*, vol. 32, no. 2, 1992, pp. 152–158.
2. W.E. Reichman et al., "Psychiatric Consultation in the Nursing Home," *Am. J. Geriatric Psychiatry*, vol. 6, no. 4, 1998, pp. 320–327.
3. A. Bharucha, S. Allin, and S. Stevens, "CareMedia: Towards Automated Behavior Analysis in the Nursing Home Setting," *Int'l Psychogeriatrics*, vol. 15, supplement 2, 2003, pp. 47–48; [www.ipa-online.net/pdfs/29766\\_IPA\\_7x10.pdf](http://www.ipa-online.net/pdfs/29766_IPA_7x10.pdf).
4. C. Steele et al., "Psychiatric Symptoms and Nursing Home Placement in Alzheimer's Disease," *Am. J. Psychiatry*, vol. 147, no. 8, 1990, pp. 1049–1051.
5. H. Wactlar et al., "Informedia Experience-on-Demand: Capturing, Integrating and Communicating Experiences across People, Time and Space," *ACM Computing Surveys*, vol. 31, no. 2, article no. 9 (electronic publication).
6. A.J. Allin et al., "Toward the Automatic Assessment of Behavioral Disturbances of Dementia," *Proc. 5th Int'l Conf. Ubiquitous*

*Computing (UbiComp 03), 2nd Int'l Workshop Ubiquitous Computing for Pervasive Healthcare Applications*, 2003; [www.healthcare.pervasive.dk/ubicomp2003](http://www.healthcare.pervasive.dk/ubicomp2003).

7. J. Yang et al., "Multimodal People ID for a Multimedia Meeting Browser," *Proc. 7th ACM Int'l Conf. Multimedia (Multimedia 99)*, ACM Press, 1999, pp. 159–168.
8. D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects Using Mean Shift," *Proc. IEEE Computer Vision and Pattern Recognition (CVPR 00)*, IEEE CS Press, 2000, pp. 2142–2145.
9. R. Yan, J. Yang, and A. Hauptmann, "Automatically Labeling Video Data Using Multiclass Active Learning," *Proc. Int'l Conf. Computer Vision (ICCV 03)*, IEEE CS Press, 2003, pp. 516–523.
10. M. Fischler and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography," *Comm. ACM*, vol. 24, no. 6, 1981, pp. 381–395.
11. J.L. Barron, D.J. Fleet, and S.S. Beauchemin, "Performance of Optical Flow Techniques," *Int'l J. Computer Vision*, vol. 12, no. 1, 1994, pp. 43–77.
12. L. Wixson, "Detecting Salient Motion by Accumulating Directionally Consistent Flow," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, 2000, pp. 774–780.
13. H. Schneiderman and T. Kanade, "A Statistical Method for 3D Object Detection Applied to Faces and Cars," *Proc. IEEE Computer Vision and Pattern Recognition (CVPR 00)*, IEEE CS Press, 2000, pp. 746–751.
14. J. Gao, A. Hauptmann, and H. Wactlar, "Combining Motion Segmentation with Tracking for Activity Analysis," *Proc. 6th Int'l Conf. Automatic Face and Gesture Recognition*, IEEE CS Press, 2004.
15. A. Mihailidis et al., "The Use of Computer Vision in an Intelligent Environment to Support Aging-in-Place, Safety, and Independence in the Home," *Proc. 5th Int'l Conf. Ubiquitous Computing (UbiComp 03)*, LNCS 2864, Springer-Verlag, 2003.
16. G. Tzanetakis and P. Cook, "A Framework for Audio Analysis," *Organized Sound*, vol. 4, no. 3, 2000, pp. 169–175.
17. O. Chapelle, P. Haffner, and V. Vapnik, "SVMs for Histogram-Based Image Classi-



**Alexander G. Hauptmann** is a senior systems scientist in Carnegie Mellon University's Department of Computer Science. His research interests include language, speech, and video analysis and retrieval. He received his PhD in computer science from CMU. Contact him at the Dept. of Computer Science, Carnegie Mellon Univ., Pittsburgh, PA 15213; [alex@cs.cmu.edu](mailto:alex@cs.cmu.edu).



**Jiang Gao** is a postdoctoral researcher with Carnegie Mellon University's Robotics Institute. His research interests include computer vision, pattern recognition, and human-computer interaction. He obtained his PhD in electrical engineering from Northwestern Polytechnic University. Contact him at the Robotics Institute, Carnegie Mellon Univ., Pittsburgh, PA 15213; [jgao@cs.cmu.edu](mailto:jgao@cs.cmu.edu).



**Rong Yan** is a doctoral candidate at the Language Technologies Institute in Carnegie Mellon University's School of Computer Science. His research interests include multimedia retrieval, video content analysis, and machine learning. He obtained his BS in computer science from Tsinghua University. Contact him at the Language Technologies Inst., Carnegie Mellon Univ., Pittsburgh, PA 15213; [yanrong@cs.cmu.edu](mailto:yanrong@cs.cmu.edu).



**Yanjun Qi** is a doctoral candidate at the Language Technologies Institute in Carnegie Mellon University's School of Computer Science. Her research interests include data mining and machine learning. She obtained her MS in language technologies from CMU. Contact her at the Language Technologies Inst., Carnegie Mellon Univ., Pittsburgh, PA 15213; [qyj@cs.cmu.edu](mailto:qyj@cs.cmu.edu).



**Jie Yang** is a senior systems scientist at Carnegie Mellon University's Human-Computer Interaction Institute. He has been involved in developing multimodal systems, such as real-time face tracking, gaze-based interfaces, and multimodal people ID. He received his PhD in electrical engineering from the University of Akron. Contact him at the Human-Computer Interaction Inst., Carnegie Mellon Univ., Pittsburgh, PA 15213; [yang+@cs.cmu.edu](mailto:yang+@cs.cmu.edu).



**Howard D. Wactlar** is the vice provost for research computing, is the associate dean, and holds the Alumni Research Professor of Computer Science chair in Carnegie Mellon University's School of Computer Science. Contact him at the Dept. of Computer Science, Carnegie Mellon Univ., Pittsburgh, PA 15213; [hdw@cs.cmu.edu](mailto:hdw@cs.cmu.edu).

fication," *IEEE Trans. Neural Networks*, vol. 10, no. 5, 1999, pp. 1075–1090.

18. A. Wilson and A.F. Bobick, "Parametric Hidden Markov Models for Gesture Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, 1999, pp. 884–900.
19. M. Christel, A. Hauptmann, and H. Wactlar, "Improving Access to Digital Video Archives through Informedia Technology,"

*J. Audio Eng. Soc.*, vol. 49, nos. 7–8, 2001; [www.informedia.cs.cmu.edu](http://www.informedia.cs.cmu.edu).

20. A.G. Hauptmann and H.D. Wactlar, "Indexing and Search of Multimodal Information," *Proc. Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP 97)*, vol. 1, IEEE Press, 1997, pp. 195–198.

For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/publications/dlib](http://www.computer.org/publications/dlib).