Multimodal Assisted Living Environment

Barış Evrim Demiröz 1, İsmail Ari 1, Alexander Ronzhin 2, Aysun Çoban 3, Hülya Yalçın 4, Alexey Karpov 2, and Lale Akarun 1

1 Dept. of Computer Engineering, Boğaziçi University, 34342 Bebek, İstanbul, Turkey
{baris.demiroz,ismailar,akarun}@boun.edu.tr

2 Laboratory of Speech and Multimodal Interfaces
Saint-Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), Russia
{ronzhinal,karpov}@iias.spb.su

3 Dept. of Electrical and Electronics Engineering, Boğaziçi University
34342 Bebek, İstanbul, Turkey
aysun.coban@boun.edu.tr

4 Mechatronics Education & Research Center, İstanbul Technical University, 34469 Maslak, İstanbul, Turkey
hulyayalcin@itu.edu.tr

Abstract—The global aim of this project is to establish a multimodal environment for an assisted living smart space. We have designed the multimodal assisted living environment, where multiple people are tracked by omnidirectional video cameras installed in the environment, as well as speech and audio signals are processed by microphone arrays with an automatic system for speech recognition and acoustic event detection. The multiple objects tracking precision of the proposed video monitoring system was 0.78 and 0.73 for single person and three people scenario, respectively. The precision is measured as 62.81% and 72.31%, correspondingly. The proposed speech and audio event recognition system is multilingual one and it is able to recognize a set of voice commands both in English and Russian, as well as some non-speech acoustic events. Recognition accuracies for speech commands and non-speech acoustic events were 96.5% and 93.8%, correspondingly. The multimodal interfaces make possibility for users to access the assisted living information system by speech and gestures, combining several modalities in order to interpret requirements and behavior of the user.

Index Terms—surveillance systems, elderly healthcare, foreground segmentation, keypoint tracking, audio signal processing, speech recognition, acoustic event detection

I. INTRODUCTION

With the advances in real time monitoring systems, intelligent environments equipped with video and audio sensors offer promising solutions for home care and independent living applications and improve the quality of the care provided to the users, especially for the rapidly growing population of elderly. There are numerous techniques involving different modalities such as special designs of wristwatches equipped with sensors to get information from the subject’s body, special beds sensing the person laying on it, infrared or colour cameras mounted to the walls, and audio sensors to detect any emergency call [1, 2, 3]. Among these, video cameras and microphones installed in the smart environment are pervasively used, and these modalities are taken into account in this study for monitoring the elderly people.

Visual data is mainly used for person tracking and automatic activity recognition in an assisted environment. In action recognition, videos are readily segmented in time. A video is expected to have only one action. Additionally, an activity is defined to be an ordered set of actions. For example, cooking is an activity whereas stirring is an action. Human action understanding has many application areas concerning security, surveillance, assisted living, and even entertainment. In activity detection, first, the activity is localized over time and space and the label assignment follows. With an activity recognition unit, it is possible to detect some set of actions involving emergency states such as falling and alert the users of the system about the extraordinary situation as studied in [4]. State-of-the-art human activity recognition modules consist of tracking of people, describing their motion in means of visual clues such as keypoints and classifying the performed action [5, 6]. The detection and tracking of people are based on lower level background / foreground segmentation techniques such as Gaussian mixture models [7, 8] and codebook model [9]. The blobs found after foreground segmentation are used for extracting features which describe the related person. [10] gives a detailed survey on the widely-used interest point detection and feature extraction methods in literature. Current
works in the literature show that the interest point descriptors are mainly used in cameras with low distortion. Interest point descriptor usage based on omnidirectional (fish-eye) cameras appears to take attention of the researchers with the increase of surveillance systems equipped with them.

Reasoning based on only visual information is vulnerable to be misleading since the illumination might change severely, occlusions may occur and so the action might be misclassified. Using audio helps to make the system more robust and maintainable. Audio consists of human speech and environmental sounds such as knocking of the door and water coming out of pipe. Human’s speech refers to the processes associated with the production and perception of sounds used in spoken language, and automatic speech recognition (ASR) is a process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a software or hardware module. Several kinds of speech are identified: spelled speech (with pauses between phonemes), isolated speech (with pauses between words), continuous speech (when a speaker does not make any pauses between words) and spontaneous natural speech. Recent automatic speech recognizers exploit mathematical techniques such as Hidden Markov Models (HMMs), Artificial Neural Networks (ANN), Bayesian Networks or Dynamic Time Warping (dynamic programming) methods. The most popular ASR models apply speaker-independent speech recognition though in some cases (for instance, personalized systems that have to recognize owner only) speaker-dependent systems are more adequate.

Although speech is the most informative acoustic event, but other types of events also convey useful information. Therefore detection and classification of acoustic events may be used for people monitoring and their activity detection [11]. For example, applause or laughs during dialog, chair movements, human cough or door slam, etc. There are many recent publications on automatic detection and recognition of individual acoustic events: cough [12, 13, 14], sound of human fall [15, 16, 17, 18], crying and screaming [19] by audio signal processing. Most of the automatic acoustic event recognition systems use standard methods for ASR like classifiers based on HMMs, Gaussian mixture models (GMMs), support vector machines (SVMs) or Neural Networks with some extra features: zero-crossing rate, audio signal energy, event duration, impulse form, etc. System reliability may also be increased by robust non-speech sounds detection in recorded signals.

Classification and/or detection of acoustic events are new field of acoustic scenes analysis [20], which are process an acoustic signals and convert it in symbolic description. This description is corresponding to listener perception of different acoustic events, which are represented in signals and their sources.

Acoustic events detection may be employed in different environments [21], such as hospital [22], a kitchen [23], or even bathroom [24]. In meeting environments such goal of acoustic events detection is rather new, but was already estimated in works of two international companies: in the project CLEAR 2006 with three participants and CLEAR 2007 with six participants. Most of presented systems use standard combination of cepstral coefficients and HMMs for audio event recognition.

In this work, a multimodal assisted living environment is designed where multiple people are tracked by omnidirectional cameras that are installed in the environment. Additionally, microphone arrays are used for achieving of audio signal processing to provide automatic speech recognition and acoustic event detection.

Organization of the paper is as follows: In the preceding chapter, the details of our approach are given in details. Then, the gathered database in the smart environment is given with the experiments and found results. Finally, the conclusion is given with further discussions and the future work.

II. METHODOLOGY

This project is composed of two modalities which are video and audio. We begin this section by giving the procedure used for multiple people tracking and continue with the automatic classification of speech and environmental sounds. People’s natural communication methods processing by multimodal user interfaces provides to organize comfortable and intuitive understandable interaction process between a user and smart environment [25].

A. Video-based Monitoring

1) Background/Foreground Segmentation

A camera is installed in the assisted living environment and it monitors the room in question. Loosely speaking, the information we need to gather for understanding the underlying positions and actions of the people is extracted from the non-stable regions in the scene which we call the foreground regions. The method of Gaussian mixture models (GMMs) is used in this study to segment the foreground objects and an additional step of shadow removal is applied to enhance the result by removing artefacts due to small changes in chromaticity and intensity [8]. In this approach each pixel is considered individually and modelled with a number of Gaussians. Each pixel is labelled as foreground or background using a statistical approach based on multiple Gaussian distributions. Let \( x(t) \) denote the value of the pixel in question at time \( t \). It may be multidimensional according to the number of channels in the video. We have a ratio found by using the posteriors of the pixel being a background (BG) or a foreground (FG) as:

\[
R = \frac{p(BG|x(t))}{p(FG|x(t))}
\]

which can be rewritten using the Bayes’ formula as:

\[
R = \frac{p(x(t)|BG)p(BG)}{p(x(t)|FG)p(FG)}
\]

The latter part of the formula can be ignored by assuming \( p(BG) = p(FG) \) since we don’t know too much about the pixel. Moreover, we may assume the foreground model given by \( p(x(t)|FG) \) is uniform since the foreground objects might have any colour and therefore, this can also be removed from the equation. Finally, we have the background model, \( p(x(t)|BG) \). In practice the underlying real model cannot be known but an approximate model is trained from what we have seen so far.
We choose a reasonable amount of time $T$ and form the training set as $X_T = \{x^{(t)} , x^{(t-1)} , ..., x^{(t-T)} \}$. Then, we can estimate the background model using our training set, update $X_T$ for each sample and re-estimate $p(x^{(t)} | BG, X_T)$. Actually, it is $p(x^{(t)} | BG+FG, X_T)$ since it also includes the FG objects. Let $M$ be the number of Gaussian distributions used to model the pixel value, $\mu_m$ be the mean, $\sigma_m^2$ be the variance in proper dimensions and $\hat{\pi}_m$ be the weights of $m^{th}$ Gaussian. A pixel value can be found by

$$p(x^{(t)} | BG+FG, X_T) = \sum_{m=1}^{M} \hat{\pi}_m N(x^{(t)}; \mu_m, \sigma_m^2 I)$$

where $\sum_m \hat{\pi}_m = 1$.

Each parameter is estimated via a recursive update rule. After having the Gaussians, the procedure upon a new pixel is as follows: The nearest Gaussian component is checked. If not found, a new component with parameters $\hat{\pi}_{m+1} = \alpha, \mu_{m+1} = x^{(t)}$ and $\sigma_{m+1}^2 = \sigma_0^2$ is added where $\alpha$ and $\sigma_0^2$ are properly defined variables by the user. If $M_{\text{max}}$ is reached, then the component with the smallest weight is discarded. Please see [8] for further details on selecting the number of components adaptively. Background pixels will form a Gaussian with a higher weight since they stay longer time in the scene in this model whereas the foreground objects will form new Gaussians with smaller weights and leave the scene. Thus, they fire smaller values given the background model which is used to identify them.

Another issue is the existence of shadows in the environment. The shadow regions have a small amount of chromaticity and intensity difference to the background. For a given pixel value, a specific amount of chromatic distortion and intensity distortion from the background model is allowed [7].

### B. Blob Detection

Blobs are a group of pixels used to infer higher level information about the objects such as the appearance or the performed activity. We apply morphological closing to fill the holes and connect the nearby pixels to form a unified connected component since the foreground segmentation module works on individual pixels. Under the assumption of a single user, an area threshold is used and all the components having smaller area are removed. Then, the largest connected component is selected to be the person and it is tracked. In multiple people tracking a more complex model is required. We seek for good matching of blobs based on descriptors extracted from the interest points on the blobs as discussed in the following sections.

### C. Feature Point Detection and Descriptor Extraction

Due to the requirement of giving real-time feedback in the assisted living environment, FAST feature point detection is preferred to slower alternatives such as Harris or Difference of Gaussians [26]. The keypoints are extracted only from the bounding rectangles of the found foreground parts for each frame. Using bounding rectangles instead of the actual segmentation provided better results, since foreground detection can result in smaller regions than the actual foreground. After locating the keypoints, corresponding BRIEF (Binary Robust Independent Elementary Features) descriptors are extracted from the image [27]. BRIEF keypoint descriptors are preferred to SURF or SIFT again with speed concerns. BRIEF descriptors are found using relatively simple intensity difference tests and reported to be highly discriminative even when using relatively few bits. An additional advantage is that it uses binary representation that makes the comparisons efficient by using Hamming distance instead of $L_2$ norm. BRIEF is originally designed to work on greyscale images but we extended it to work on colour images to improve the descriptor power. We followed the colour boosting transformation in the opponent colour space [28]. 32 bytes are used for each channel and concatenated to form a final feature vector involving 96 bytes.

### D. Blob Matching and Tracking

For a given frame, for every blob’s every keypoint, the closest keypoint is searched in previous $n$ frames. Matched keypoint’s blob’s label count is incremented. Each label is assigned to the blob such that total number of counts is maximized. This is known as assignment problem, and solved with Hungarian algorithm. If any blob is left unlabeled, a new label is created and assigned to that blob.

### III. AUDIO-BASED MONITORING

A few condenser microphones Oktava MK-012 and one multichannel audio board M-Audio ProFire 2626 are used for digital audio signal processing. In total, four microphones are used; their positions in the room are presented in Figure 2. Each microphone has a cardioid capsule and captures sound signals in a sector of 120 degrees with approx equal amplification. The developed ASR system is multilingual one and it is able to recognize and interpret speech commands both in English and in Russian. The recognition lexicon of the system contains 5 English and 5 Russian words, plus 12 acoustic events for different types of activities. General architecture of the automatic recognition system is shown in Figure 1.

In order to train the speech recognizer a speech corpus has been recorded in office conditions with an acceptable level of background noise. Totally, we have recorded over 1.3 hours of audio data of one subject.

![Fig. 1. Architecture of speech recognition system.](image-url)
uses the Viterbi-based token passing algorithm [30].

IV. EXPERIMENTS AND RESULTS

A. Database

In order to validate the methods, we have installed two omni-directional cameras and four microphones in a room involving two chairs and two tables of use, and a sink. The design of the assisted living environment is given in Figure Error! No bookmark name given.2.

One camera is mounted at the ceiling and the other one is mounted on the side wall. The frame resolution is $640 \times 480$ pixels and the frame rate is around 8 fps for both cameras. The cameras offer higher resolution but with a lower fps which was not sufficient for tracking. Input frame samples taken from the cameras can be seen in figure 3. The microphones are located on a grid so as to take the input clearly from the person with the nearest microphone.

Fig. 2. Design of the assisted room used for database collection.

Fig. 3. Sample inputs of side camera (left) and up camera (right).

We designed two scenarios to form a prototype to basic actions performed at a smart home. First scenario involves single person and finishes with an emergency state. The latter one is a more challenging setup which involves three subjects where the subjects occlude each other at several frames. Scenario 1 involves audio whereas scenario 2 does not. Both scenarios are stated below.

**Scenario 1 (audio/video):**

1. Enter room from door (open & close)
2. Walk to table 1
3. Pick up glass of water from table 1
4. Walk to chair 1
5. Sit on chair 1
6. Drink water
7. Cough after drinking
8. Stand up
9. Walk to table 1
10. Release glass
11. Walk to sink
12. Wash hands
13. Exit room (open & close)
14. Enter room (open & close)
15. Walk to chair 2
16. Sit on chair 2
17. Phone rings on table 2
18. Say “Answer phone”
19. Talk on the phone
20. Say “Hello”
21. Say “I’m fine”
22. Say “Don’t worry”
23. Say “Bye”
24. Stand up
25. Walk to table 1
26. Pick up metallic cup from table 1
27. Free walk
28. Fall the cup on the floor and leave it there
29. Free walk
30. Fall
31. Cry for “help”

**Scenario 2 (video):**

1. P1 (person 1) enters the room
2. P1 walks to chair 1
3. P1 sits on chair 1
4. P2 enters the room
5. P2 walks to table 1 passing from the right-side of P1
6. P1 stands up
7. P1 walks to table 1 (on the left-side of P2)
8. P1 and P2 shake hands
9. P3 enters the room
10. P1 walks to middle as P3 walks to middle
11. P1 and P3 meet and shake hands
12. P2 leaves the room by passing from the right-side of the room
13. P1 and P3 walks to table 2
14. P1 and P3 stays in front of table 2
15. P1 walks to the door and exits
16. P3 walks to the door and exits

1) Video database

We gathered samples belonging to Scenario 1 from 6 different subjects. They were free to perform the fall as they are used to do. In case of Scenario 2, we collected a total of 18 samples from 5 different subjects by interchanging the roles of P1, P2, and P3 in the scenario. All videos are recorded in
MJPEG, which is the original format provided by the cameras.

2) Audio database

A training audio database has been recorded in the office room in normal acoustic conditions, where a factor of entering new people into the room during the recording session was avoided that allows removing major external noises. It was decided to apply a speaker-dependent recognition system taking into account system’s aim and usability issues.

Figure 4 presents a tree classification of user’s speech commands and acoustic events, which were collected for the database.

Table I

<table>
<thead>
<tr>
<th>Speech command</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer phone</td>
<td>aph</td>
</tr>
<tr>
<td>Help</td>
<td>hlp</td>
</tr>
<tr>
<td>No</td>
<td>N</td>
</tr>
<tr>
<td>Problem</td>
<td>pro</td>
</tr>
<tr>
<td>Yes</td>
<td>y</td>
</tr>
<tr>
<td>Da (Да)</td>
<td>d</td>
</tr>
<tr>
<td>Net (Нет)</td>
<td>nt</td>
</tr>
<tr>
<td>Otvetit (Ответить)</td>
<td>ot</td>
</tr>
<tr>
<td>Pomogite (Помогите)</td>
<td>pmg</td>
</tr>
<tr>
<td>Problema (Проблема)</td>
<td>bm</td>
</tr>
</tbody>
</table>

Table II

<table>
<thead>
<tr>
<th>Acoustic event</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applause</td>
<td>ap</td>
</tr>
<tr>
<td>Chair moving</td>
<td>cm</td>
</tr>
<tr>
<td>Cough</td>
<td>co</td>
</tr>
<tr>
<td>Moan</td>
<td>cr</td>
</tr>
<tr>
<td>Steps</td>
<td>st</td>
</tr>
<tr>
<td>Throat cleaning</td>
<td>th</td>
</tr>
<tr>
<td>Fall</td>
<td>fa</td>
</tr>
<tr>
<td>Key jingle</td>
<td>kj</td>
</tr>
<tr>
<td>Paperwork</td>
<td>pw</td>
</tr>
<tr>
<td>Phone ring</td>
<td>pr</td>
</tr>
</tbody>
</table>

Fig. 4. Classifications of acoustic events and speech commands.

The test audio data has been collected during experiments with a test scenario described in 3.3.5.

3) Scenario 1 experiments and results

For implementation and evaluation Open Computer Vision Library is used [31, 32]. The evaluation metrics used for person tracking is Multiple Object Tracking Precision and Accuracy, MOTP and MOTA, as described in [33]. MOTP (overlap) metric can be interpreted as the average ratio of intersection and the union of the true object and hypothesized boundaries.

At development of the training audio database, five check points were selected in the room space. Four of them located on the floor under each microphone and the fifth point was in the centre of the room. At each check point, one expert pronounced a speech command or simulated an acoustic event. All speech commands and acoustic events have been simulated and recorded 100 and 200 times, correspondingly. Tables I and 2 describe speech commands and acoustic events with their textual labels.
For all 10 videos in Scenario 1, MOTP (overlap) and MOTA were calculated as 0.78 and 64.85% respectively where the miss, false alarm and mismatch rates are 2.92%, 32.11% and 0.10%.

4) Scenario 2 experiments and results

For all 36 videos in Scenario 2, MOTP (overlap) and MOTA were calculated as 0.73 and 72.31% respectively where the miss, false alarm and mismatch rates are 23.47%, 3.74% and 0.47%.

Scenario 1 false positive rate is very high due to persistent unsuccessful foreground detection that is present in two of the videos. Because there are only 10 videos, the problematic two videos dominate the results. Since Scenario 2 contains multiple people occluding each other, miss ratio is dramatically high compared to Scenario 1. When the performance of the tracker is investigated in detail, it was observed that miss ratio is significantly higher for the videos acquired from top camera in both scenarios. This can be explained by the extended duration of the occlusion that is present in videos of top camera.

Fig. 5. Scenario 1 results
ultimodal assistive environment was in the centre of the room. A sub-
ject performed the following sequence of actions: (1) come to a
check point in the room; (2) give a speech command five
times with one second pause between utterances; (3) move to
the next point.

At the second experiment, 2811 audio wave files have been
recorded, which include 1226 files with acoustic events and
1585 files with speech commands, correspondingly. Ta-
bles 3 and 4 present the obtained results on recognition accu-
Racy in the form of confusion matrices (first column shows
commands and events issued by a user with numbers of sam-
ple files).

5) Experiments on audio modality recognition

Evaluation of the speech and audio modality recognition
system was made in multimodal assistive environment. Several
test scenarios simulating real situations were prepared for the
experiments.

For the first scenario five check points were selected in the
room. Such points were located under each of four micro-
phones and the last one was in the centre of the room. A sub-
ject performed the following sequence of actions: (1) come to
a check point in the room; (2) give a speech command five
times with one second pause between utterances; (3) move to
the next point.

At the second experiment, 2811 audio wave files have been
recorded, which include 1226 files with acoustic events and
1585 files with speech commands, correspondingly. Ta-
bles 3 and 4 present the obtained results on recognition accu-
Racy in the form of confusion matrices (first column shows
commands and events issued by a user with numbers of sam-
ple files).

These results show that most of speech commands were
recognized with a good accuracy over 90%, but there were
some mistakes as well. For example, the system recognizes the
English command “Problem” as the Russian word “Проблема”,
because there are many equal phonemes in both words, but these commands have the same meanings.

Presented results show that the lowest recognition rate
(63%) was observed for the audio event “Fall”. In one third of
such cases this event was recognized as “Steps”. Such mistake
can happen because this event was simulated by dropping a
big bag filled by some clothes.

V. CONCLUSION

The developed speech and audio event recognition system
is multilingual one and it is able to recognize commands both
in English and in Russian. At the experiments, 2811 wave files
with user’s speech commands and simulated acoustic events
have been recorded in total. Recognition rates for speech
commands and non-speech acoustic events were 96.5% and
93.8%, correspondingly.

In the experiments conducted, the multiple object tracking
precision (overlap) of the proposed video monitoring system
was 0.78 and 0.73 for single person and three people scenario
respectively. The precision is measured as 62.81% and
72.31% correspondingly.

ACKNOWLEDGMENTS

This work is supported by TÜBİTAK under project
108E161; by the Ministry of Education and Science of Russia
in the framework of the Federal Targeted Programme «R&D
in Priority Fields of S&T Complex of Russia for 2007-2013»
(contract No. 11.519.11.4025) and by the grant of the Presi-
dent of Russia (project No. MK-1880.2012.8).

REFERENCES

[1] H. Ö. Alemdar, G. R. Yavuz, M. O. Özen, Y. E. Kara, Ö. D. Incel,
L. Akarun, C. Ersoy, “Multi-modal fall detection within the WeCare
framework”. Proceedings of the 9th ACM/IEEE International Confer-
ence on Human-Robot Interaction (HRI 2014)


Lale Akarun received the BS and MS degrees in Electrical Engineering from Bogazici University, Istanbul, Turkey, in 1984 and 1986, respectively, and the PhD degree from Polytechnic University, Brooklyn, New York, in 1992. From 1993 to 1995, she was Assistant Professor of Electrical Engineering at Bogazici University, where she is now Professor of Computer Engineering. Her current research interests are in image processing, computer vision, and computer graphics.

Alexander Ronzhin received the M.Sc. diploma from St. Petersburg State University of Airspace Instrumentation (SAI) in 2010. His main research interests include processing audio-visual signals for intelligent spaces, text-to-speech systems, user localization, tracking and video surveillance in the smart room, audio-visual speech analysis and synthesis. Currently he is a PhD student of Speech and Multimodal Interfaces Laboratory of SFIPIAS. He is the (co)author of 15 papers in proceedings of international conferences.

Alexey Karpov received the M.Sc. diploma from SUAI and Ph.D. degree in computer science from SFIPIAS, in 2002 and 2007, respectively. His main research interests are in ASR, TTS, HCI multimodal user interfaces. He is the senior researcher of the Speech and Multimodal Interfaces Laboratory of SFIPIAS. He has been the (co)author of more than 100 papers in journals and proceedings of top international conferences. Dr. Karpov is the co-chairman of the organizing committee of the series of International conferences on Speech and Computer SPECOM and member of ISCA, EURASIP (Local Liaison Officer in Russia), IAPR associations. He was awarded with the medal of RAS for the best research work of young Russian scientists in 2011.

Barış Evrirm Demiröz received the BS degree in Computer Engineering from Istanbul Technical University, Istanbul, Turkey. He is currently pursuing PhD degree at the Perceptual Intelligence Laboratory (PILAB), Bogazici University, Istanbul, Turkey. His main research interests are computer vision, omnidirectional cameras, and image retrieval systems.
İsmail Arı received the BS and MS degrees in Computer Engineering from Bogazici University, Istanbul, Turkey. Currently he is a PhD candidate and teaching assistant in the Department of Computer Engineering in Bogazici University. His research interests include Bayesian machine learning, computer vision, and music processing. Arı has an MS degree from the same department.

Dr. Hulya Yalcin is an assistant professor at Faculty of Engineering in Istanbul Technical University. Prior to that, she was a postdoctoral fellow in the Robotics Institute at Carnegie Mellon University. She received her PhD at Brown University in Division of Engineering in 2004. She received her BSc and MSc degrees in Electrical & Electronics Engineering from Bogazici University in 1996 and 1999 respectively. Her current research interests include computer vision, image processing, active vehicle safety systems, human-computer interaction and ambient intelligence.