Using Robust Audio and Video Processing Technologies to Alleviate the Elderly Cognitive Decline

Vasileios Mylonakis
Athens Information Technology
0.8km Markopoulou Ave.
Peania, 19002, Greece
vmil@ait.edu.gr

John Soldatos
Athens Information Technology
0.8km Markopoulou Ave.
Peania, 19002, Greece
jsol@ait.edu.gr

Aristodemos Pnevmatikakis
Athens Information Technology
0.8km Markopoulou Ave.
Peania, 19002, Greece
apne@ait.edu.gr

Lazaros Polymenakos
Athens Information Technology
0.8km Markopoulou Ave.
Peania, 19002, Greece
lcp@ait.edu.gr

Alex Sorin
IBM Haifa Research Lab
Haifa University Campus
Mount Carmel, Haifa 31905, Israel
sorin@il.ibm.com

Hagai Aronowitz
IBM Haifa Research Lab
Haifa University Campus
Mount Carmel, Haifa 31905, Israel
hagaia@il.ibm.com

ABSTRACT

We are recently witnessing a growing interest for pervasive context-aware products and services for elderly users. This is largely due to falling fertility and rising longevity phenomena, as well as due to the proliferation of the aging population all over the world. In this paper we present a number of leading edge audio and video processing technologies, which can be exploited to build robust ambient assisted living applications for elderly groups. In particular, we discuss application requirements aiming at alleviating the cognitive decline of elderly users and present audio and video processing components that can essentially fulfill these requirements. We emphasize on technologies such as automatic speech recognition, speaker identification, face detection, person tracking, face identification, and demonstrate how mature versions of these technologies can be appropriately customized to give a significant boost to AAL applications for senior citizens. The challenges, solutions and ideas within this paper are part of the EU project HERMES, which aims at providing an integrated approach to cognitive care, based on assistive technology that reduces age-related decline of cognitive capabilities.

Keywords

J.3 Life and Medical Sciences, I.4.8 Scene Analysis

1. INTRODUCTION

The vision of pervasive computing is to transform physical spaces into computationally active and intelligent environments, which provide non-obtrusive human-centric services regardless of time and end-users location. A main characteristic of these services is that they are context-aware in the sense that they can automatically sense and perceive the status of their surrounding environment, and accordingly exploit this status in shaping their application logic [18]. Numerous instances of context-aware services have been developed in research initiatives, but also in the scope of deployments in smart homes, smart conference rooms and systems for ambient assisted living (AAL) [19].

In the area of AAL we are recently witnessing a growing interest for pervasive context-aware products and services which target elderly users. This is largely due to falling fertility and rising longevity phenomena, as well as due to the proliferation of the ageing population all over the world. AAL solutions for the elderly target a variety of assistive functionalities such as social integration and decentralized communication support (e.g., supporting interaction with friends and relatives), as well as e-health and e-care (e.g., facilitating caretakers and minimizing the need for hospitalization). Most pervasive systems and services for the elderly employ usually one of the following approaches to ubiquitous computing and context-aware systems [16]:

- Tag based systems, which read tags (e.g., Radio Frequency Identification, Active Badges) to track objects, humans and infer context.
- Wearable computing, which is based on sensors that are attached to humans and employ custom I/O mechanisms to derive and disseminate context.
- Smart spaces, which are ordinary physical environments equipped with pervasive sensors and devices that perceive and react to people in a natural and non-intrusive manner.

Smart spaces provide the less obtrusive approach to implementing human centric services for aged users. The later are not likely to be familiar with ICT technologies and devices, which makes the natural interactivity provided by smart

PETRA 2008, July 15-19,2008, Athens, Greece

spaces preferable. Nevertheless, application development in smart spaces is still a complex task, since it involves a wide range of highly distributed and heterogeneous hardware and software elements. An integral element of smart space applications are perceptual components, which provide information about the identity, location, activities and sometimes the goals of human actors through person trackers, person identification components and other situation identification elements. In addition to these development challenges, smart space applications for elderly users must extend their outreach outside the domestic environment given that user activities are not confined to the home environment. On the contrary, a variety of elderly user activities (e.g., shopping, doctor visits) takes place in outdoor environments. Hence, application development is made more difficult since developers have to deal with mobility, CPU-constrained devices, as well as their interaction with in-home systems.

In this paper we present a number of ambient assisted living applications for elderly groups, which are built based on leading edge audio and video processing technologies. In particular, we discuss application requirements aiming at alleviating the cognitive decline of elderly users and present audio and video processing components that can essentially fulfil these requirements. The presented audio and video processing applications can be used to derive context and accordingly build context-aware human-centric services for aged users. Hence, we present technologies such as automatic speech recognition, speaker identification, face detection, person tracking, face identification, and demonstrate how mature versions of these technologies can be appropriately customized to give a significant boost to AAL applications for senior citizens. Note that customization is very important given that the involvement of elderly users poses unique requirements for audio-visual components. As a prominent example conventional automatic speech recognition systems can not reach high-performance levels when applied on aged users. In addition to audio-video processing technologies, we also discuss some basic structuring principles that can be employed to combine the underlying audio and visual processing applications into meaningful applications and services. From a service viewpoint the presented applications are clustered into three functional categories, namely memory aid services, ambient calendar services, as well as properly designed cognitive training games. Towards implementing these functionalities the respective services must function in both indoor and outdoor environments. As a result, we discuss the instantiation of the presented technologies and structuring principles in both indoor and outdoor environments.

The challenges, solutions and ideas presented in this paper are part of the European Commission co-funded project HERMES [2], which aims at providing an integrated approach to cognitive care, based on assistive technology that reduces age-related decline of cognitive capabilities, while also supporting the aged users when necessary. In line with already discussed requirements HERMES will support the elderly both in their home environment, as well as in outdoor environments (i.e. through mobile devices). The rest of this paper is structured as follows: Section 2 presents our target users and environment in order to acquaint the readership with the operational environment of the presented technologies and services. Section 3 presents relevant audio processing technologies, emphasizing on research challenges stemming from the need to deal with elderly users' speech. Likewise, section 4 discusses video processing technologies and their use within elderly applications in a way that renders the pervasive environment as non obtrusive as possible. Section 5 presents architectures and structuring principles for putting together audio and video processing technologies towards added value human centric services. Finally section 6 summarizes the paper and draws basic conclusions.

2. TARGET USERS AND APPLICATIONS

Ageing is associated with a wide range of problems for seniors. In this paper we emphasize on applications targeting cognitive changes that start usually at the age of thirty and are quite normal in older people. The impact of the cognitive decline is usually manifested in terms of slower reaction, slower reasoning and thinking capability, as well as in aspects relating to working memory such as the capacity for maintaining some information and mentally operating with it at the same time.

To alleviate these cognitive problems we envisage assistive technology that focuses on facilitation of episodic memory, cognitive training through games, support for prospective memory, as well as conversation support and interactive reminiscence. In particular, our scenarios span three complementary axes:

- Memory aid services enabling the aged users to conveniently query about past events and access useful responses or memory cues facilitating their recollection.
- Ambient calendar services facilitating users in keeping a context-aware diary in an automated manner.
- Cognitive training services, based on appropriately designed games that help senior citizens to improve their cognitive capabilities.

Memory aid services can be built through collecting and indexing contextual information, which aged users could conveniently recall in future moments. Context-awareness and collection can be realized based on a variety of sensing and context acquisition technologies in both indoor and outdoor environments. These sensing and acquisition technologies, along with time-stamping provide the so-called W5 context (Who?, Where?, What?, Why?, When?).

In the scope of indoor environments, audio and visual technologies for smart spaces (comprising visual and acoustic sensors) come into foreground. These include technologies for detecting and recognizing people (e.g., face recognition, faceID, speakerID), technologies tracking end-users location (e.g., visual person tracking), as well as technologies recognizing speech (i.e. automatic speech recognition and transcription). Based on these technologies memory aid services are capable of keeping track of people entering and leaving specific rooms within the smart space, as well as identifying speakers and "what is said" within the house. The advantage of these technologies is that they can be non intrusive. In the outdoor environment, aged users have to use a mobile device (e.g., mobile phone or Personal Digital Assistant (PDA)) in order to capture context. ASR technology can be employed to capture and stored outdoor conversations. Furthermore, standard sensing technologies like Global Positioning Service (GPS) can be used to capture locations. Also, mobile devices can be used to capture context like photos, images. Overall, the above technologies, which are discussed in later sections enable users to capture and retrieve past information (including memory aiding cues) about people, visitors, places, conversations. While additional audio-visual processing technologies (e.g., scene analysis, activity recognition, object recognition) could enhance the context acquisition and search functionality, these add-on technologies are out of the scope of this paper.

Having these technologies at hand, aged users can support a variety of memory aid, calendar applications, as well as cognitive training games. Memory aid applications can be based on the recording and post-processing of audio-visual information. Characteristic examples of such audio-visual information are the audio conversations of the old person with friends, relatives and doctors, as well as images of these people. Such kind of information can be tagged based on W5 context-acquisition components including GPS and time-stamping technologies. Accordingly, the post-processing of this information can allow storage of context cues that can be later provided to the elderly user in the scope of a memory aid service. For example the old person can query the memory aid service regarding past conversations, summaries of past events, the identity of people, as well as the context-based reminding of important moments. Furthermore, it can access information based on specific keywords, names, emotions and/or other contextual cues that are explicitly provided to the system by end-user. In addition to memory aid services, elderly users can take advantage of ambient calendar services that help the user enter key moments within an automated calendar. Calendar entries can be inserted to the systems in a conventional (i.e. manual) fashion, but also in a context-aware (i.e. automatic) manner. Also, a hybrid semi-automated mode, where some information is filled manually and other in a context-aware fashion is possible.

The main difference between memory aids and ambient calendar applications is that memory aids focus on reactive functionalities, while calendar applications are also proactive (i.e. context-aware). Reactive reminding examples include aids in order to recall indoor and outdoor conversations, conversations derived from telephone, as well as conversation derived from social contact. For instance, reminding conversations between the old person and the doctor can facilitate the episodic memory of the old person. Calendar reminder examples include reminders about shopping lists during shopping session (i.e. context-aware reminders), reminders associated with points of interest (based on GPS based location awareness), as well as recollection of past pictures and photos.

In addition to memory aids and ambient calendar services, the presented audio and visual processing algorithms can be used to implement cognitive training games. The later games can be used/play by users in order to reduce their cognitive decline. Audio based games can involve short sentences that are displayed on a screen for short time intervals and then disappear. Accordingly the game asks the end-user to recall/repeat the short sentence text, in which case the voice of the user is recorded. Based on ASR technology the system can provide an automated indication of the user's success or failure in faithfully repeating the text. In such a game the collected voice data can be used to improve the acoustic model of the end user, without any intervention or even knowledge of the end-user.

Another audio based cognitive training game could be based on playing back short excerpts extracted from the doctor's audio recording, as the later are stored in the Ambient Assisted Living system. Similar to the previous game the old user could be asked to type what the doctor said. The game can assess the answer and provide to the user feedback for self evaluation. Note that the data collected from this exercise can be used for an improvement of doctor's acoustic model.

Cognitive training games can also be constructed based on the visual technologies, which are described in this document. For example, users may be prompted to match faces with names, with the system auditing the user's match. This game could be based on a static set of know faces. However, face identification technology can be employed to make this matching system dynamic: new faces-images can be added to the system based on everyday monitoring and associated visual processing activities. In this case the registration of new faces into the system must be made only when there is high-confidence regarding the on the similarity of the face with the set of existing faces (that are used for training). This tactic along with the robust face recognition algorithm (described in Section 4) can guarantee a high success rate, without imposing essential constrains on the set of images.

The following sections describe audio and video processing technologies which can be used to implement the already discussed memory aids, ambient calendar and cognitive training functionalities.

3. AUDIO PROCESSING

3.1 Automatic speech recognition & transcription

Figure 1 depicts the basic structure of an Automatic Speech Recognition System, which represents audio signal as a sequence of acoustic feature vectors (approx. 70-100 vectors per second) and accordingly extracts instantaneous features based on the perceptual mechanisms of the human auditory system. Furthermore, addition of dynamic features (time derivatives or other types of the instantaneous features blending) is used for capturing the temporal behavior of speech. Key to improving an ASR system's performance is the adaptation of the system to speaker voice and recording conditions. To this end, certain transformations of the acoustic features derived from the target data and of the acoustic model are made. The adaptation process can be either supervised (i.e. based on audio and exact transcripts) or unsupervised (i.e. on the fly, without any transcript as input).

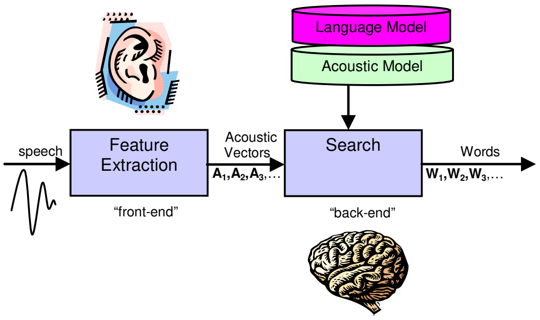

Figure 1: Basic structure of an ASR system.

ASR performance is calculated and evaluated on the basis of Word Error Rate (WER), which signifies the percentage of words that were either substituted or deleted or inserted in comparison to the actual speech. Towards supporting scenarios requiring recording and data mining conversations, such as those described in Section 2 above, an WER than 35% is required. WER in leading ASR systems can be less than 10% when speakers talk in close-talking microphones. As a prominent example the system developed by IBM in the scope of the EU TC-STAR project featured a 8.6% WER [14], when it was evaluated based on speeches delivered at European Parliament Plenary Sessions (EPPS). WER rates deteriorate however when evaluating systems that operate on the far-filed (i.e. based on distant microphones). For instance the systems developed in the scope of the CHIL (Computers in Human Interaction Loop) project [1], involved transcription of meetings based on distant microphones, and rendered a WER close to 45%. In the same meeting environment close talking microphones gave a WER below 35% which could be significantly lower in the absence of the cross-talking effect.

Note that when it comes to supporting transcription scenarios involving elderly users, ASR systems become much more challenging. Researchers report 80%-150% relative increase in WER on elderly speech. Retraining on elderly data helps but WER remains high. In the scope of the MALACH (Multilingual Access to Large Spoken Archives) project [3], 50,000 interviews with Holocaust survivors spanning 32 languages and 100,000 hours of speech were transcribed. The ASR system had to deal with the challenges of variable level of background noise (in the home environment), as well as with the fact that speakers featured elderly, emotional and accented English speech. Baseline ASR systems that never have seen MALACH data performed at the 60% WER level. Only extensive training on hundreds of hours produced by hundreds of speakers allowed to achieve 25% WER.

The above experience reveals the following challenges with respect to transcription and mining of elderly speech:

- Far-field ASR for elderly in the scope of in-home scenarios is particularly challenging and difficult to render acceptable WER.

- Systems based on recording with close talking microphones can have acceptable performance. In such a case senior citizens can use wireless microphone in the home environment. Given this setting, the elderly users must become accustomed to speak out the things that the system has to remind in the scope of the ambient calendar and/or memory aid applications.

- The use of mobile devices for conversation recordings creates a setting similar to far field recordings. Hence, a high quality microphone, capturing uncompressed audio (16 kHz at least) is essential.

- Existing ASR baseline systems must be retrained for elderly speakers and based on the recoding conditions of the target environments.

- Adaptation of the Acoustic Model (AM) to specific speaker voices could make the system more robust. To this end, the target system could be built as a personalized system through: (a) Enrolment of the elderly users, (b) Automatic unsupervised adaptation of the involved speaker AMs behind the scenes and (c) Automatic supervised AMs adaptation based on data acquired during the operation of some applications (e.g., data from cognitive training games).

We believe that in order to achieve high accuracy in the transcription of speech produced by an elderly speaker in different environments two ASR acoustic models per speaker have to be trained: one for the in-door environment with wall mounted microphones, and another one for out-door environments using the microphones of the terminal device. Accordingly the ASR language model can be tailored for the lexicon pertinent to target application scenarios.

3.2 Speaker identification

The speaker identification problem relies on a set of known target speakers, along with training data for each target speaker. The set of speakers for the problem at hand can be either closed (in which case the speaker is always one of the target speakers) or open (in which case any unknown speaker must be rejected). In the case where a single speaker is talking in the recording, speaker identification identifies who is speaking. A related problem is the speaker tracking one, which tracks the talking person in cases where several speakers are talking in the recording.

The vast majority of research in speaker identification is performed on telephone speech. Indicative performance metrics in terms of EER (equal error rate) are: 4% to 9% on clean telephony speech. While such a performance is acceptable for many applications, the problem becomes much more challenging when it comes to run the identification algorithms in noisy conditions and based on distant microphones. Channel mismatch, environmental noise, as well as reverberation cause the recognition accuracy to degrade significantly. Furthermore, there are no adequate publicly available corpora for system development. In an AAL environment with elderly users these challenges have to be addressed. A promising direction would be to adapt the target speaker models repeatedly ”behind the scenes” using audio data accumulated by the system and speaker identity information deduced from a ”who's speaking” cognitive game.

3.3 Emotion detection

Speech is a major channel for communicating human emotions [11]. Hence, speech is important source for detection of speaker's emotional state. The problem of speech-based emotion detection lies in partitioning the speech signal to homogenous segments, each conveying distinct emotion from a predefined set of emotional categories, such as anger, fear, frustration, sadness, surprise and joy. Features indicative of emotional state that are used in emotion detection algorithms include pitch (tone) and energy (loudness) contour statistics, spectral shape descriptors, speech rate and certain keywords. Support Vector Machine (SVM), Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) are commonly used for modelling and classification. Note that some very useful features (e.g. average pitch) are noticeably speaker-dependent. A main problem associated with speech emotion detection research is the lack of real life data for experimentation and training. Hence, most of the research work is conducted using actors play data.

Despite the research challenges, emotion detection can be used to implement added-value functionalities for the scenarios discussed above. In particular, it can serve as an additional dimension for information retrieval in the scope of memory aids (e.g., answering the question: 'What it was that excited me yesterday so much?').

The robust emotion detection challenge can be handled in the context of AAL system by development of speaker dependent emotion recognition, given that the majority of the speakers are known in terms of their inherent characteristics (pitch, spectral tilt, speech rate) in neutral state. As soon as these characteristics are measured and modeled, any deviation from the neutral state can be detected.

4. VISUAL PROCESSING

4.1 Visual 3D tracking

Our system developed visual tracking in three dimensions using multiple cameras is detailed in this section. Two approaches towards 3D tracking can be followed [10]. According to the data-driven approach, 2D trackers operate independently on the camera views; then the tracks of the same target are collected into a 3D one [9]. According to the model-based approach, a 3D model is maintained by rendering it onto the camera views, searching for supporting evidence per view, and based on that, updating the 3D model [7].

If the model-based approach is followed, then rendering can be implemented in a way that it mimics the real image formation process, including effects like perspective distortion and scaling, lens distortion, etc. In the context of multi-body tracking this is particularly advantageous, since occlusions can be handled at the rendering level. This way, the update is done by looking for supporting evidence only in the image parts where the different models are visible, thus occlusions are handled in a systematic manner. The disadvantage is that you have to initialize and occasionally update the models, which in some situations may be tricky. For that reason, we follow the data-driven approach, where target initialization is done per camera view based on the human body evidence collected there.

A body tracker [13] provides the bodies of humans present in the scene. These are used for search regions for the face tracking system [9]. The latter comprises a synergy of detectors and trackers. Three face detectors [22] for frontal and left/right profile faces provide candidate face regions in the body areas. The face candidates are validated using the probability scores from a Gaussian Mixture Model. The surviving candidates are checked for possible merging, as both the profile detectors and the frontal one can detect different portions of the same face if the view is half-profile. The resulting face candidates are associated with faces existing in the previous frame and also with tracks that currently have no supporting evidence and are pending to either get an association, or be eliminated. Any faces of the previous frame that do not get associated with candidate faces at the current frame have a tracker initiated to attempt to track similarly colored regions in the current frame. CAM-Shift [6], Kalman [8] and Particle Filter [4] trackers have been utilized. The CAM-shift tracker can only work with color cues and its performance is limited under lighting changes. Kalman filters need linear state dynamics and measurements. The Particle Filter tracker is not limited in its dynamics and gives the best results. Finally, all active face tracks are checked for duplicates, i.e. high spatial similarity.

We follow the approach of [12] for our color-based Particle Filter tracker. The chosen object model makes only a weak assumption for the state evolution. Motion smoothness is guaranteed by a Gaussian random walk model, while lock recovery after erratic motion or occlusion is aided by uniform component. The measurement model uses color cues. A reference color likelihood ratio is built around the face detection that initializes the target. An object histogram is first built using the pixels inside the detection rectangle, with more confidence placed on the central than the peripheral pixels. Then a background histogram is built using the pixels around the detection rectangle. The ratio of the two forms the reference color likelihood, which carries the color information that discriminates the tracked face from the background. The color histogram of the image portion represented by the state is also calculated by weighting all pixels equally and without any background histogram. The Bhattacharyya distance of the two histograms is used in the exponential distribution assumed by the measurement model. The proposal distribution has the Gaussian random walk component of the object model, plus the contribution of the measurement model in the form of a sum of Gaussian densities. To do so, we search in a grid around all the particles for the locations that have a Bhattacharyya distance from the reference color likelihood smaller than a threshold. These locations of good colour match are used to bias the proposal distribution towards them. The reference color likelihood is updated with memory when a validated face detection is associated with the track.

Multiple targets [21] can be addressed either independently using multiple trackers, or jointly using approaches like Multiple Hypotheses Tracking, Joint Probabilistic Data Association, or Approximate Bayesian [10]. The joint approaches are much more computationally intensive; to be as close as possible to real-time operation we initialize an independent tracker upon first detection of a new target. Target interaction is resolved by our two-level approach: While some of the targets tracked by the body tracker can be confused due to interaction, this confusion is resolved at the face tracking level.

For 3D tracking, the views of the face of the same person from the different cameras are associated following a 3D space to 2D image planes approach. The space is spanned by a 3D grid. Each point of the grid is projected onto the different image planes. Faces whose centers are close to the projected points are associated to the particular 3D point. 3D points that have more than one face associated to them are used to form possible associations of views of the face of the same person from the different cameras. The same face in a camera view cannot be a member of different valid associations. This renders some of the associations mutually exclusive. After eliminating duplicate associations, the remaining ones are grouped into possible sets of mutually exclusive associations and sorted according to a weight that depends on the distance of each association from the face center and on the number of other associations that contradict it. All the mutually exclusive sets of possible associations are validated using a Kalman filter in the 3D space. For each new frame, all possible solutions are compared to the state established on the previous frame, penalizing solutions which fail to detect previously existing targets, or in which there are detections of new targets in the scene. While this strategy reduces the misses and false positives, it does not prevent new targets from appearing, as in the case of new people entering the room, all solution pairs will include that new target and thus will be equally penalized.

4.2 Face recognition

The goal of the far-field video-based face recognition system is to extract features from faces that are robust to low resolution, motion artifacts and large pose, expression and illumination changes. To this effect, a face preprocessing scheme is used, followed by feature extraction in the linear subspace obtained by sub-class LDA projection. Then, a classifier is proposed, based on the Bayesian approach of modeling intrapersonal differences. The classifier yields the postulated identity per facial image, and an associated confidence. The confidences are used to fuse the individual identities into a single one per video sequence. The following subsections detail each of the components of the system.

The faces are intensity normalized to 127 mean and 40 standard deviation. Then they are processed by an 8x8 block DCT. Only 9 of the 64 DCT coefficients per block are kept and are normalized to zero-mean, unity standard deviation, prior to being concatenated together with those form the other blocks into a vector representing the face.

Features are extracted from the face vectors in a subspace trained by subclass Linear Discriminant Analysis. The difficulty of automatic generation of the subclasses, handling faces cropped from videos, no matter their pose, expression or illumination, is addressed as in [20]. That algorithm exploits the training face manifold to build a hierarchical clustering tree that automatically clusters the training faces of a particular person in subclasses, whose number differs from one person to another. The classifier employed is the nearest neighbour with cosine distance. The identities obtained per detected and validated face are fused across time using the sum rule to obtain the identity per probe video.

4.3 Visual identity tracking: Door watcher

For visual identity tracking, the face recognition and face tracking modules are merged into a single system that monitors people entering or leaving the house. The developed door watcher utilizes two cameras on the door-frame, one facing outwards and another facing inwards. This way the faces of people both entering and leaving the house are tracked and recognized. The faces collected by the detectors from each track are the input to the face recognizer. The reason for not using the tracked image regions but only those returned by the detectors is that the tracked regions can be partial faces or just head skin patches due to occlusions. Fusion is carried out per track.

We have tested our system trained with 18 individuals and have obtained 100% recognition rate per track for the people entering or leaving without paying attention to the cameras monitoring them. The same system has been used to capture the faces used for training. Hence the training videos depict people in the same location, only they are recorded three months earlier.

5. APPLICATION ARCHITECTURES

Towards supporting the application scenarios outlined in Section 2, audio and visual technologies must be combined in a synergetic fashion in order to form an integrated Ambient Assisted Living (AAL) System. A possible architecture for structuring these components into an integrated pervasive AAL system for in-door environment is depicted in Figure 2. This architecture comprises a number of distinct groups of components, which are structured more or less in a layered fashion. Specifically, these components groups are as follows:

- Sensing Infrastructure: This group consists of the sensors that support context extraction, automatic information tagging and indexing. A non-intrusive sensing infrastructure (hidden microphones, cameras) is required for controlled in-door environments (house, office, hospital). However, additional sensors embeded on the terminal devices (cameras, microphones, GPS) are needed for sensing in out-door environments.

- Visual Processing: This group includes the range of visual processing components, which are described in Section 4. Visual 3D tracking, face recognition and door watcher components consume the underlying camera streams. These algorithms need to be configured, adapted and deployed in the operational environment of the AAL applications.

- Audio Processing: Another group of components enables information indexing and annotation based on audio processing. Automatic Speech Transcription, Speaker Identification and Voice-based Emotion Detection systems consume the audio streams. Similarly to the visual processing systems, these components are adapted to the operational environment of the AAL application.

- Low-Level Information Fusion: Low-Level Information Fusion components combine contextual cues from both audio and visual processing systems with a view to improving the recognition accuracy obtained from unimodal recognition systems. Multimodal analysis can generally lead to a better recognition accuracy rate compared to unimodal analysis.

- Context Modeling: Several of the application scenarios and use cases require situation identification beyond the contextual cues provided in by the audio and video extraction components. In identifying more sophisticated context, the architecture prescribes the implementation of context modeling scripts that define composite contextual states as combinations of underlying elementary acquisition components. For example, to identify that an elderly user is seeing his/her doctor several contextual components need to be combined. Hence, the application scenarios must be defined based on a set of non-sequential collections of contextual states and their allowable transitions e.g., following the network of situations approach [17]. Contextual states are triggered by combinations of the outputs of elementary (e.g., A/V processing) components. Upon identification of a composite contextual state (e.g., person visits his/her doctor, person fell down) service or action logic will be executed. Furthermore, at the same time instant selected information is annotated, indexed and stored.

- Indexing, Annotation and Knowledge Conceptualization components: The contextual information provided by the presented A/V components serves as a basis for tagging, annotating, indexing and subsequently querying content. Since this information stems from a variety of distributed and heterogeneous components, some

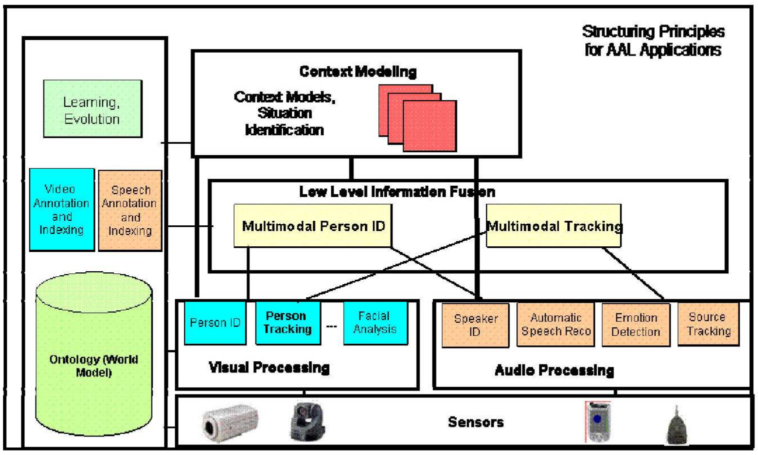

Figure 2: Structuring principles for building AAL applications (indoor and outdoor environments).

harmonization is required. This is achieved based on a common knowledge structure and related meta-data specifying how information should be formatted, stored and disseminated throughout the various applications. To this end, a knowledge base (e.g., to be implemented as a W3C ontology [5]), which comprises all objects, concepts and relationships of interest, needs to be specified. Moreover, data models and interfaces for accessing the knowledge based towards indexing and annotating information are required. Having a knowledge base at hand, additional knowledge extraction and data mining mechanisms can be implemented to allow discovery of additional situations and context of the user and its surrounding environments.

Note that the computational and storage components comprising this architecture are hosted in one or more servers of the in-door environment (i.e. home servers).

Figures 3 and 4 explore the more specific integration of audio components for real-time situation detection at home, as well as for off-line spoken data indexing and retrieval. In particular, Figure 3 depicts the operation of the audio-technologies described in Section 3 (i.e. speaker identification, ASR, emotion detection) in the scope of indoor environments (i.e. homes of elderly person). To increase efficiency, Figure 3 assumes the use of Voice Activity Detection (VAD) [15] technology prior to recording user input for transcription, identification or emotion detection.

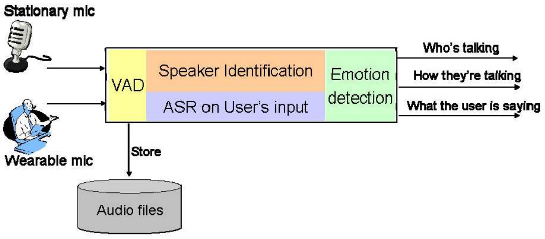

Figure 3: Nearly real time situation detection at home.

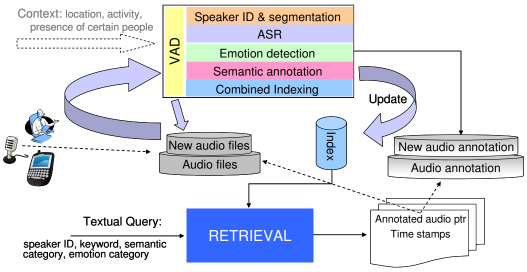

Figure 4: Spoken data indexing and retrieval for off-line scenarios.

Figure 4 illustrates the process of populating the knowledge base (depicted as ”index” in the figure) with audio-based context acquired in outdoor environments. In this case audio information (e.g., audio recordings) is collected within the mobile device of the user. However, this information is not processed on the device, since the later is CPU-constrained and cannot accommodate the computational demanding audio processing components. Therefore, the old user has to process this information off-line, using the home server that hosts the audio processing components. The whole process can be carried out in an automatic manner (e.g., as soon as the end-user recharges his/her mobile device), in order to make it as non-intrusive to the elderly user as possible.

6. CONCLUSIONS

In this paper we have presented a number of audio and video based perceptual processing components, which can be used to implement a wide range of ambient assisted living functionalities. At the same time we have also introduced some indicative application architectures which define structuring principles for integrated these perceptual technologies into added value applications for senior citizens. We have described three families of such added value applications spanning the areas of memory aids, ambient calendar and cognitive training games. These applications can contribute to minimizing the cognitive decline for elderly users.

We have also highlighted stringent requirements and constraints associated with building context-aware applications for elderly users. Specifically, speech processing applications and speech based emotion detection require customization to the peculiarities of elderly speech. Also, a number of usability issues are raised, given that elderly people are not accustomed to using devices and context-aware applications. In the scope of the paper we have attempted to make some suggestions regarding possible solutions to these problems. A more systematic and complete resolution of these important issues asks however for a thorough and consistent understanding of end-user requirements. Nevertheless, the availability of robust perceptual components and application architectures open up new opportunities in the area of the ambient assisted living applications for senior citizens. We envisage that the unique partnership between leading edge technology providers with usability and gerontology experts (realized in the scope of the HERMES project) will capitalize on these opportunities towards novel added value products.

7. ACKNOWLEDGMENTS

This work is part of the HERMES Specific Targeted Research Project (Contract No: FP7-216709), partially funded by the European Commission under the Information Society Technology (IST) program. The authors acknowledge help and contributions from all partners of the project.

8. REFERENCES

[1] Chil (computer in the human interaction loop) eu fp6 integrated project. http://chil.server.de/.

[2] Hermes eu fp7 specific targeted research project. http://www.fp7-hermes.eu.

[3] Malach (multilingual access to large spoken archives). http://malach.umiacs.umd.edu/.

[4] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, Feb. 2002.

[5] N. Baumgartner and W. Retschitzegger. A survey of upper ontologies for situation awareness. In *Knowledge Sharing and Collaborative Engineering (KSCE 2006)*, St. Thomas, US Virgin Islands, November 2006.

[6] G. Bradski. Computer vision face tracking for use in a perceptual user interface, 1998.

[7] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia. A generative approach to audio-visual person tracking. In R. Stiefelhagen and J. S. Garofolo, editors, *Multimodal Technologies for Perception of Humans, First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006, Southampton, UK, April 6-7, 2006, Revised Selected Papers*, volume 4122 of *Lecture Notes in Computer Science*, pages 55–68. Springer, 2007.

[8] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[9] G. Karame, A. Stergiou, N. Katsarakis, P. Papageorgiou, and A. Pnevmatikakis. 2d and 3d face localization for complex scenes. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'07)*, pages 371–376, London, UK, September 2007. IEEE Computer Society.

[10] O. Lanz. Approximate bayesian multibody tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1436–1449, 2006.

[11] C. M. Lee, S. S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. In *Seventh International Conference on Spoken Language Processing (INTERSPEECH 2002 - ICSLP)*, pages 873–876, Denver, CO, USA, 2002.

[12] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proceedings of IEEE*, 92(3):495–513, 2004.

[13] A. Pnevmatikakis and L. Polymenakos. Robust estimation of background for fixed cameras. In *15th International Conference on Computing (CIC '06)*, pages 37–42, Mexico City, Mexico, November 2006. IEEE Computer Society.

[14] B. Ramabhadran, O. Siohan, L. Mangu, G. Zweig, M. Westphal, H. Schulz, and A. Soneiro. The ibm 2006 speech transcription system for european parliamentary speeches. In *Ninth International Conference on Spoken Language Processing (INTERSPEECH 2006 - ICSLP)*, Pittsburgh, PA, USA, September 2006.

[15] E. Rentzeperis, C. Boukis, A. Pnevmatikakis, and L. Polymenakos. Combining finite state machines and lda for voice activity detection. In *Artificial Intelligence and Innovations 2007: from Theory to Applications*, pages 323–329, Peania, Athens, Greece, September 2007.

[16] J. Soldatos, N. Dimakis, K. Stamatis, and L. Polymenakos. A breadboard architecture for pervasive context-aware services in smart spaces: middleware components and prototype applications. *Personal and Ubiquitous Computing*, 11(3):193–212, 2007.

[17] J. Soldatos, I. Pandis, K. Stamatis, L. Polymenakos, and J. L. Crowley. Agent based middleware infrastructure for autonomous context-aware ubiquitous computing services. *Computer Communications*, 30(3):577–591, 2007.

[18] V. Stanford. Pervasive computing goes to work: Interfacing to the enterprise. *IEEE Pervasive Computing*, 01(3):6–12, 2002.

[19] V. M. Stanford. Pervasive computing: Applications - using pervasive computing to deliver elder care. *IEEE Distributed Systems Online*, 3(3), 2002.

[20] A. Stergiou, A. Pnevmatikakis, and L. Polymenakos. The ait multimodal person identification system for clear 2007. In *CLEAR'07 Evaluation Campaign and Workshop - Classification of Events, Activities and Relationships*, Baltimore, MD, USA, May 2007.

[21] L. D. Stone, C. A. Barlow, and T. L. Corwin. *Bayesian Multiple Target Tracking*. Artech House Radar Library, 1999.

[22] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, pages 511–518, Kauai, HI, USA, December 2001.