Robust Multimodal Audio-Visual Processing for Advanced Context Awareness in Smart Spaces

Aristodemos Pnevmatikakis, John Soldatos, Fotios Talantzis
and Lazaros Polymenakos
Athens Information Technology
19,5 km Markopoulou Peania, Ave.
{apne, jsol, fota, lcp}@ait.edu.gr
http://www.ait.edu.gr

Abstract. Identifying people and tracking their locations is a key prerequisite to achieving context-awareness in smart spaces. Moreover, in realistic context-aware applications, these tasks have to be carried out in a non-obtrusive fashion. In this paper we present a set of robust person identification and tracking algorithms, based on audio and visual processing. A main characteristic of these algorithms is that they operate on far-field and un-constraint audio-visual streams, which ensures that they are non-intrusive. We also illustrate that the combination of their outputs can lead to composite multimodal tracking components, which are suitable for supporting a broad range of context-aware services. In combining audio-visual processing results, we exploit a context-modeling approach based on a graph of situations. Accordingly, we discuss the implementation of realistic prototype applications that make use of the full range of audio, visual and multimodal algorithms.

1 Introduction

The emerging Ubiquitous Computing paradigm aims at exploiting casually accessible sensors, devices and networks to transparently provide computing services, regardless of time and location of the user [1]. A core characteristic of ubiquitous computing environments is context sensitivity, which refers to the ability of ubiquitous devices/systems to react to their environment and adapt their behavior accordingly [2]. To this end, ubiquitous computing services are essentially context-aware, since they acquire and process information about their surrounding environment. This information is derived implicitly, without requiring end-users to provide explicit input. There are several approaches to derive implicit information. As a prominent example there are tag-based approaches, where tags are read to track

Please use the following format when citing this chapter:
Pnevmatikakis, Aristodemos, Soldatos, John, Talantzis, Fotios, Polymenakos Lazaros, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 290–301

objects and infer context [3]. Another approach is the wearable computing paradigm, where sensors and customized input-output operations are used to instrument humans [4]. Context-awareness can be also realized in smart spaces [5], which leverage sensors and effectors to achieve natural interaction between the humans and the environment.

In this paper we emphasize context-awareness in smart spaces. Context derivation in smart spaces relies on sophisticated audio-visual processing algorithms, which leverage audio and video sensors to derive context. For example, audio processing may be used to derive the location of speakers (i.e. acoustic localization). Similarly, processing of video streams can be used to track people location (i.e. visual person tracking), to detect faces (i.e. face detection), as well as recognize people (i.e. face recognition). A key challenge for these audio-visual processing algorithms is that they should be as non-intrusive as possible, which is in-line with the unobtrusive human-centric nature of context-aware applications in smart spaces. The audio and visual processing techniques introduced in this paper are aligned to this requirement, as they operate on far-field and un-constraint audio-visual streams. Hence the operation of the proposed algorithms does not require that human actors are under specific lighting conditions or have a particular orientation.

Another challenge is to combine context cues from both audio and video signal processing towards identifying more complex contextual states. Non-trivial context-aware applications need to identify composite situations based on combinations of elementary context cues about people identity and location, as well as context from other sources (e.g., sensors, other perceptual components). Along with the audiovisual processing algorithms, we also illustrate a context modeling approach for fusing context from these algorithms. This modeling approach is based on a graph of situations defining the situation of interest and the allowed transitions between them. While this approach is quite static, it can enable a wide range of context-aware applications, for example, relating to support for meeting and conferences, as well as security and surveillance. As prominent examples, we illustrate the implementation of context-aware actuating services, group activities recognition and memory aids in the scope of lectures, meeting and conferences. The rest of the paper is structured as follows: section 2 presents algorithms for deriving context based on far-field visual processing, while section 3 elaborates on audio processing algorithms for speaker localization and tracking. Section 4 illustrates our context modeling approach based on the network of situations techniques. Section 5 discusses some real-life prototype applications that leverage both the presented algorithms and the introduced situation modeling technique. Finally section 6 concludes the paper.

2 Visual Processing

The block diagram of the visual system is shown in Fig. 1. It comprises a detector that operates on video streams and a recognizer that provides the identity of the faces detected over a time interval. The detector begins with a tracker that segments the bodies from the video streams. This is based on Stauffer's [6] adaptive background

estimation. Compared to a static background, the adaptive approach fades into the background lighting variations and furniture movements. This is coupled with a shadow detector [7] to get rid of shadows that deform the extracted body shapes and cause false alarms of target collisions. In order not to fade the bodies into the background when the people remain stationery, the gated approach of [8] is used to treat the found bodies as targets and the region around a target in a frame as a gate, i.e. as the region where it is expected to find the target in the next frame. Should the background adaptation fade an immobile body by diminishing the foreground pixels inside the gate, the target is not lost; the estimated body position and the gate remain the same as in the previous frame.

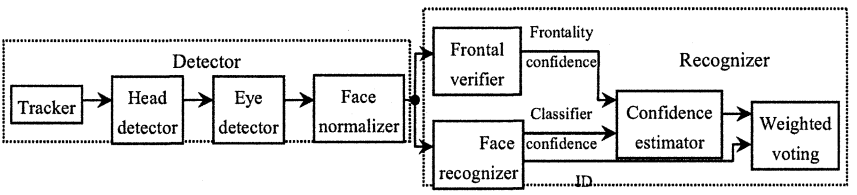

Fig. 1. Block diagram of the complete face detection and recognition system

Target collisions are handled both while they occur, and after they finish. The former is important for smart room applications, where people can be collaborating for some time and have their targets overlapping. At collisions, the involved targets are merged into a super-gate, in which the known number of bodies is sought. Handling during collision is the attempt to keep the targets approximately separated using *k*-means. When a collision finishes, it is important not to swap the tracks, since face recognition is based on each track. This is done by matching the 2D histogram of normalized red and green color components of each of the post-collision targets to those of the pre-collision targets.

Heads are detected by processing the outline of the body. The derivative of the width of the body as a function of its height at shoulder level increases significantly, indicating the beginning of the head. This approach fails at profile views and at bent-over bodies; in those cases just a fixed percentage of the body height over width is passed as head, usually resulting to overestimated head regions. This degrades both the speed and the accuracy of the eye detector that follows. The latter is not very troublesome, as profile faces are useless for recognition.

The eye detector operates inside the head region. The approach followed is based on vector quantization of the colors and face geometry constraints. Since thick, raised hair with highlights deteriorates detection performance, the face region inside the head is estimated by processing the edge image of the head region (Fig. 2.a). The vertical and horizontal sums of pixels in the edge image drop rapidly outside the face region and the face is finally confined within sums that are above 85% of the sum means across every direction. The colors of the pixels of the face are vector-quantized to 6 colors, to identify distinct regions inside the face (Fig. 2.b). Properties of the regions are used to identify eye candidates. The properties related to color are

brightness and resemblance to human skin. The latter is evaluated using the human skin color histogram of [9]. Both these quantities should be low, since, at the targeted resolution, eyes usually are a blur. The properties related to shape are the extent and the roundness of the regions. Finally, the properties related to face geometry are the center locations of the regions, which are mildly constrained inside the head region. These constraints can lead to misses at profile views, but this is not troublesome, since profiles are not useful for recognition. The eye positions are estimated by searching for the darkest spots near the candidate region centers. Based on them, the faces are normalized to standard size. Eye detection errors of more than 10% of the eye distance usually result to misclassifications. Only smaller errors are regarded a hit. The hit rate as a function of eye distance increases abruptly in the 16 to 18 pixels range. Hit rates below 30% for eye distances close to 10 pixels are not unexpected; one pixel of error is not unusual for human annotators! As expected, the RMS eye detection error relative to the eye distance drops as the face resolution increases.

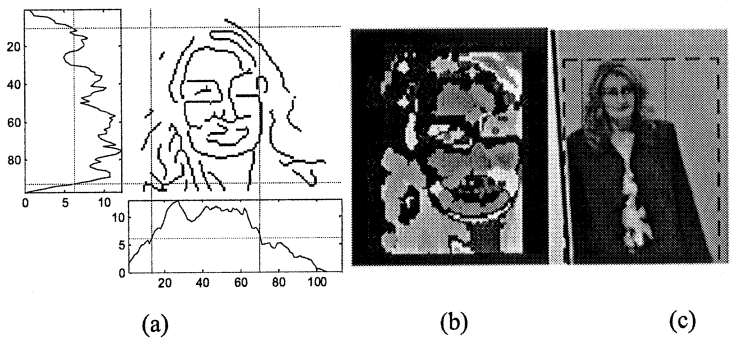

Fig. 2. Eye detector. (a) Estimation of the face inside the head region. (b) Regions after vector quantization. (c) Body, head and eyes

The recognizer is based on the PCA+LDA combination [10]. Even though there are methods less sensitive to eye detection [11], these are much slower and hence not suitable for real-time operation. The system is trained to recognize 16 people. This is a reasonable number for smart room applications. Although intensity preprocessing helps under illumination changes [10], it is destructive under pose and expression changes [10], so it is not used. To account for the effect of imperfect eye detection [11], the manually annotated eye positions of the ten training faces per person are deliberately perturbed to every one of their eight neighbors, in total creating 81 training faces form every original one [8]. This produces an RMS eye perturbation of 2% of the eye distance, effectively matching the training and testing conditions.

After recognition, each detected face is associated with an identity *ID* and two confidence values *W* and *D*. *W* is the confidence the classifier has about the decision

that person k_1 is the best-matching, compared to the second-best, k_2. This can be expressed as the ratio of the distance d_2 of the testing face from the center of k_2, over its distance d_1 from the center of k_1:

$$W = d_2/d_1 \quad (1)$$

D is related to the certainty of the system that the face is frontal. This can be expressed as the Distance From Face Space (DFFS) [12] when the projection is done on 12.5% of the eigenfaces with the largest eigenvalues. W and D are combined into a single confidence value C as:

$$C = W^6 \cdot 2^{-D/3277} \quad (2)$$

As the detector is not perfect and suitable faces are not always available, the system accumulates identities and confidences over some time interval T and fuses them into a single identity using the sum rule [13]. This decision fusion scheme renders the system very robust to eye misalignments and pose changes. The system is tested two months after training, with three videos per person, using three different camera zoom settings. As a result the eye distances are between 8 and 28 pixels. Even though 57.8% of the individual recognitions failed, the system always yields the correct identity from on average the 2nd frame onwards, as the average of the sum of the confidences of the correct decisions is 72%. The proposed visual system processes 20 faces per second on an Intel Zeon at 2.8GHz, with 2Gb of RAM, SUSE 9.3 Linux and using the Intel IPP libraries.

3 Audio Processing

The processing of audio streams provides a set of functionalities that facilitate the localization, recognition and context interpretation problems. For the purposes of the present work we restrict our analysis to the Audio Source Localization (ASL) system. Collection of audio data is performed using a total of 80 microphones located in different places inside the acoustic enclosure and organized in different topologies. More analytically, there is a 64 channel linear microphone array and four smaller clusters of microphones, each containing four microphones. Each of the microphone clusters has the microphones organized in an inverted T topology.

A dominant requirement in the dynamic environments in which the microphones are employed is the localization of speakers. This is generally dealt with the estimation of the direction of arrival (DOA) of the acoustic source by means of time delay estimation (TDE) algorithms. Estimation of DOA essentially provides us with the direction from which sound is arriving from. Typically, audio data is collected in frames so that the current TDE estimate can be provided. Combination of several DOAs can then provide us with the actual source position.

The practical and, in many ways, severely restricting disadvantage of traditional methods for TDE [14] is that if the system is used in reverberant environments, the returned estimate could be a spurious delay created by the ensuing reflections. For the purposes of our system, we have proposed [15] a new mathematical framework that resolves to great amount the reverberation issues and generates robust estimations. It is thus of interest to briefly investigate the used model.

Consider two of the microphones with a distance d between them. The sound source is assumed to be in the far field of the array. For the case in which the environment is non-reverberant, the assumption of a single source leads to the following discrete-time signal being recorded at the m^{th} microphone (where m=1, 2):

$$x_m(k) = s_m(k - \tau_m) + n_m(k) \tag{3}$$

where τ_m denotes the time in samples that it takes for the source signal to reach the m^{th} microphone, and n_m is the respective additive noise (assumed to be zero mean and uncorrelated with the source signal). The overall geometry of the corresponding system can be seen in Fig. 3. Without loss of generality, this considers m_1 to be the reference microphone, i.e., τ_1=0. The delay at m_2 is then the relative delay between the two recorded signals, and thus, the relationship is reduced to $x_1(k)=x_2(k-\tau_2)$. The DOA is defined with respect to the broadside of the array as a function of any delay τ as:

$$\theta = \arcsin\left[\frac{\tau c}{f_s d}\right] \tag{4}$$

where f_s is the sampling frequency, and c is the speed of sound (typically defined as 343 m/s). Thus, DOA estimation methods rely on successful estimation of τ. However, in a real reverberant environment, each of the microphone recordings are a result of a convolution operator between the speech signal and a reverberant impulse response of significant length (depending on the reverberation level).

In order to overcome the problems introduced by reverberation we make use of the concept of mutual information (MI) by tailoring it appropriately to the tracking of an acoustic source. A review of the concept can be found in the work of Bell et al. [16].

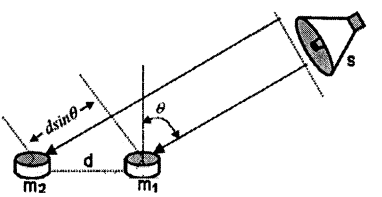

Fig. 3. Geometry of the recording system

Most of the DOA estimation techniques are required to operate in real time. We must, therefore, assume that data at each sensor m are collected over t frames $\mathbf{x}_m=[x_m(tL), x_m(tL+1),\ldots, x_m(tL+L-1)]$ of L samples. Since the analysis will be independent of the data frame, we can drop t to express frames simply as $\mathbf{x}_m$ for any t. In the context of our model, and for any set of frames, we may then write

$$\mathbf{x}_1 = \mathbf{x}_2(\tau) \tag{5}$$

where $\mathbf{x}_m(\tau)$ denotes a delayed version of $\mathbf{x}_m$ by τ samples. Thus, the problem is to estimate the correct value of and the DOA by processing two frames $\mathbf{x}_1$ and $\mathbf{x}_2(\tau)$ only.

If we were to neglect reverberation, only a single delay is present in the microphone signals. Thus, the measurement of information contained in a sample l of $\mathbf{x}_1$ is only dependent on the information contained in sample l-τ of $\mathbf{x}_2(\tau)$. In the case of the reverberant model, though, information contained in a sample l of $\mathbf{x}_1$ is also contained in neighboring samples of sample l-τ of $\mathbf{x}_2(\tau)$ due to the fact that the model is now convolutive. The same logical argument applies to the samples of $\mathbf{x}_2(\tau)$. In order to estimate the information between the microphone signals, we use the marginal MI that considers jointly N neighboring samples and can be formulated as follows [17] for the case where the recordings exhibit Gaussian behavior

$$I_N = -\frac{1}{2}\ln\frac{\det[C(\tau)]}{\det[C_{11}]\det[C_{22}]} \tag{6}$$

with the joint covariance matrix $C(\tau)$ given as

$$C(\tau) \approx \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_1(1) \\ \vdots \\ \mathbf{x}_1(N) \\ \mathbf{x}_2(\tau) \\ \mathbf{x}_2(\tau+1) \\ \vdots \\ \mathbf{x}_2(\tau+N) \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_1(1) \\ \vdots \\ \mathbf{x}_1(N) \\ \mathbf{x}_2(\tau) \\ \mathbf{x}_2(\tau+1) \\ \vdots \\ \mathbf{x}_2(\tau+N) \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} C_{11} & C_{12}(\tau) \\ C_{21}(\tau) & C_{22} \end{bmatrix} \tag{7}$$

If N is chosen to be greater than zero, the elements of $C(\tau)$ are themselves matrices. In fact, for any value of τ, the size of $C(\tau)$ is always $2(N+1)\times 2(N+1)$. For the purposes of the present letter, we call N the *order* of the tracking system. When $C(\tau)$ reaches a maximum as a function of at a specific time shift τ, then there is at this point a joint process with a maximum transport of information between $\mathbf{x}_1$ and $\mathbf{x}_2(\tau)$. According to the presented information-theoretical criterion, this is the delay that synchronizes the two recordings. In the context of DOA, this delay returns the correct angle θ, at which the signal coincides with the microphone array.

The last step in the ASL process is the combination of several DOA estimates, in order to get the actual 3D coordinates of the speaker. This is performed by calculating the crossing points between the lines defined by the estimated DOAs. In most cases these lines cross in more than one point and thus, the speaker lies within some area defined by these points. The speaker position is found by, employment of a closed-form source location estimator as found in [18]. This estimator represents a tremendous computational saving over other exhaustive search methods.

The DOAs that feed the estimation of the 3D coordinates are provided by considering a series of microphone pairs in the enclosure. The system is able to provide coordinates in space because as discussed earlier, the inverted T arrays have microphones in different planes.

4 Context Modeling

Our context modeling approach relies on the network of situations paradigm [19,20]. According to this paradigm the contextual states of interest are structured into a graph where the nodes denote the target states *ST* and the arcs *ed* the possible

transitions between contextual states. Specifically, edge ed_{ij} denotes that it is possible to reach state ST_j from state ST_i.

Contextual states may be arbitrarily complex in terms of their defining cues. The situation model is accompanied by a truth table, which depicts the underlying combination of audio or visual processing component outputs that trigger each one of the composite contextual states. To formally specify how the situation transitions occur, assume that the smart space is supported by m components having $k_1, k_2,\ldots, k_m$ outputs respectively and let $k = \max(k_1, k_2,\ldots, k_m)$. Without any loss of generality we can represent the observed outputs of all perceptual components at a given time instant t using the matrix:

$$P_{out}(t) = \{p_{ij}(t)\}, \text{ where } 1<i<m \text{ and } 1<j<k \quad (8)$$

where $p_{ij}(t)=0$ for $j>k_i$, since there are perceptual components providing less than k outputs. As a result, $P_{out}(t)$ contains the observations of the perceptual components outputs at time instant t.

For each situation ST_l ($1<l<m$) targeted by a situation model we define a matrix $\mathbf{S}_l$ comprising the target values of the perceptual components that according to the situation modeling lead to ST_l, as follows:

$$\mathbf{S}_l = \{s_{ij}\}, 1<i<m, 1<j<k \quad (9)$$

where $s_{ij} \neq 0$ if the j-th output of the i-th perceptual component contributes in the triggering the state ST_l and $s_{ij} = 0$ otherwise. Towards associating with the non-zero s_{ij} values with the observed outputs $p_{ij}(t)$, we perform an element-wise multiplication of $P_{out}(t)$, with the following matrix:

$$\mathbf{A}_l = \{a_{ij}\}, 1<i<m, 1<j<k \quad (10)$$

where $a_{ij}=1$ if $s_{ij} \neq 0$ and $a_{ij}=0$ otherwise. The result of the element-wise multiplication is a $\mathbf{P}_l$ matrix filtering the observed outputs in a way that only values defining the state ST_l are retained:

$$\mathbf{P}_l = \{a_{ij} \cdot p_{ij}(t)\}, 1<i<m \text{ and } 1<j<k \quad (11)$$

ST_l occurs when all the elements of the matrices $\mathbf{P}_l$ and $\mathbf{S}_l$ coincide i.e.:

$$\mathbf{S}_l - \mathbf{P}_l = \{s_{ij} - a_{ij} p_{ij}\} = \mathbf{O}_{mk} \quad (12)$$

where $\mathbf{O}_{mk}$ corresponds to the matrix having all its elements equal to zero. In practice, due to perceptual component inaccuracies (i.e. measurement errors) it is rare to achieve a total agreement between target and observed values. Therefore, the triggering of the situation may be defined as the case when the elements of the two matrices almost coincide, thus allowing the observed outputs to somewhat deviate from the target values:

$$\mathbf{S}_l - \mathbf{P}_l = \{s_{ij} - a_{ij} p_{ij}\} = \mathbf{E}_l \{e_{ij}\} \quad (13)$$

where $|e_{ij}|<thr$. In order for situation ST_l to be triggered, (13) has to be fulfilled, while at the same time the situation model has to be on a state ST_p that allows transition to ST_l. Therefore:

$$ST_p \rightarrow ST_l \text{ occurs whenever } |s_{ij} - a_{ij} p_{ij}| <thr \text{ and } ed_{pl}=1 \quad (14)$$

Note that this network of situations approach is general and applicable not only to audio-visual processing outputs but also to more general class of observations, which may include any sensor signals. In practice the matrix $\mathbf{S}_l$ is likely to be very sparse, which can greatly simplify (13). It is also noteworthy that (8)-(13) assume numeric values for perceptual components. While this may sound limiting, it is in general possible to encode several other domains as numbers.

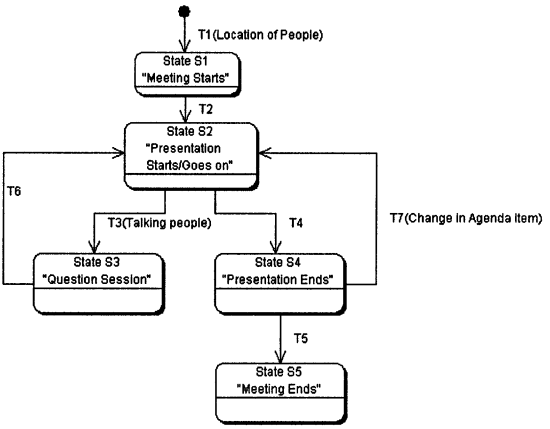

Fig. 4. Situation Model tracking different states within a meeting

Fig. 4 depicts a sample situation model for tracking activities in a meeting. The situation model consists of five states corresponding to the commencement of a meeting, the start of a presentation during the meeting, a question on the presentation, the end of a presentation and the end of a meeting. The arcs in Fig. 4 denote the possible transitions. For example a question can only occur, while a presentation is in progress, since there is no means to reach state ST_3 unless the model is in ST_2. Table 1 illustrates how particular situation states are triggered based on underlying perceptual components. NIL denotes the starting state. As an example the start of the meeting (i.e. the transition NIL→ST_1) occurs when an expected number of people are speaking very close to the table. Given a number of perceptual components (i.e. TablePeopleCount (based on Face Detection), WhiteBoardPeopleCount (based on Face Detection), Speech Activity Detection, Acoustic Localization) and their APIs, Table 1 can be mapped to equation (13) in a straightforward way. Nevertheless, the mapping is in general service specific since the elements of the matrices $\mathbf{E}_1$, $\mathbf{E}_2$,..., $\mathbf{E}_5$ (see equation (13)) are likely to be defined based on the problem at hand.

Table 1. Mapping Perceptual Component Outputs to Situation Transitions

Situation Transition	Combinations of Perceptual Components Outputs
NIL → S1	*TablePeopleCount=N (N people in table area),* *Speech Activity Detection=1*
S1 → S2	*WhiteBoardPeopleCount=1 (1 in board area),* *TablePeopleCount=N-1 (N-1 in table area),* *Acoustic Localization = (X, Y) within the board Area*
S2 → S3	*Acoustic Localization = (X, Y) within the Table Area*
S3 → S2	*Acoustic Localization = (X, Y) within the board Area*
S2 → S4	*TablePeopleCount=N,* *WhiteBoardPeopleCount=0*
S4 → S2	*WhiteBoardPeopleCount=1,* *TablePeopleCount=N-1* *Acoustic Localization = (X, Y) within the board Area*
S4 → S5	*TablePeopleCount=0 (everybody has left)*

5 Prototype Applications

A number of prototype context-aware applications have been developed based on the algorithms and the context modeling approach illustrated above. These applications are fully functional within our prototype smart space, which consists of a variety of sensors, devices and perceptual components.

The intelligent display service selects the optimal display device according the location of the target person(s) within a smart room. A smart space may have more than one means to display information. The service selects the device that is more convenient for the room participants. This is accomplished through examining display requests it in relation to the current users' context. The context of interest includes the location and orientation of participants, which are tracked based on the audio/visual processing techniques detailed above. The algorithm attempts to provide a satisfactory view for as many participants as possible.

The intelligent meeting recorder is a recording service, which can be instantiated for any of the cameras within the smart space. A realistic meeting recording service is expected to operate like an automated intelligent camera-man. In particular, an ambient recording selects the optimal camera view based on the location and orientation of the participants, as well as based on their activities and role within the group interaction. A thorough description of this service can be found in [21].

The memory jog pervasive service aims at providing non-obtrusive assistance to humans during meetings, lectures and presentations in the smart space [22]. The service identifies participants and tracks their locations within the smart room. It also keeps track of meeting progress based on a known agenda. Moreover, it records the event based on the best-camera selection mechanism. The recording is tagged with meta-data from the situation model of the service, to allow selective retrieval of the recording. The memory jog can also provide context-aware assistance through displaying relevant information from past meetings. The memory jog relies on all the audio-visual perceptual components presented above and exploits the situation model of Fig. 4 to follow higher level situations such as agenda tracking, questions tracking, meeting commencement and meeting finish.

6 Conclusions

In this paper we presented a set of robust audiovisual processing systems able to support the emerging wave of ubiquitous computing services. These systems acquire information implicitly and unobtrusively; further processing provides context-awareness. Thus the audio-visual streams are processed in order to answer the questions who, where and what. With the systems presented we can robustly identify people, find their location, track their movement and activities in an in-door multi-sensor environment. Situation modeling based on the audiovisual information acquired supplies the middleware foundation for advanced ubiquitous computing services. We presented 3 such services - the intelligent display, the intelligent meeting recorder and the memory jog. Future work will investigate more dynamic

situation modeling, scalability (more participants and sensors), extendibility (more perceptual components) and the enhancement of user experience in the smart spaces.

Acknowledgements

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909.

References

1. M. Weiser, "The Computer for the 21st Century" Scientific American, vol. 265, no. 3, 1991, pp. 66–75.
2. D. Anind, D. Salber and G. Abowd, 'A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications', Human-Computer Interaction, 16,. 2001.
3. R. Want, A. Hopper, V. Falcao and J. Gibbons, 'The Active Badge location System', ACM Transactions on Information Systems 10(1), pp. 91-102, January 1992.
4. A. Smailagic, D.P. Siewiorek "Application Design for Wearable and Context-Aware Computers", IEEE Pervasive Computing Vol. 1 No. 4 Dec. 2002 pp. 20-29.
5. B. Johanson, A. Fox and T. Winograd. 'The Interactive Workspaces Project: Experiences with Ubiquitous Computing Rooms', IEEE Pervasive Computing Magazine Vol. 1 No. 2, April-June 2002.
6. C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking", CVPR, pp. 246–252, 1999
7. L-Q. Xu, J. Landabaso and M. Pardàs, "Shadow Removal with Blob-Based Morphological Reconstruction for Error Correction", ICASSP 2005
8. A. Pnevmatikakis and L. Polymenakos, "An Automatic Face Detection and Recognition System for Video Streams", *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Edinburgh, July 2005
9. M. Jones and J. Rehg. "Statistical color models with application to skin detection", *Computer Vision and Pattern Recognition*, pp. 274–280, 1999
10. E. Rentzeperis, A. Stergiou, A. Pnevmatikakis and L. Polymenakos, "Impact of Face Registration Errors on Recognition", AIAI 2006, accepted for publication
11. A. Pnevmatikakis and L. Polymenakos, "A Testing Methodology for Face Recognition Algorithms", *Lecture Notes in Computer Science*, Vol. 3869, 2005
12. M. Turk and A. Pentland, "Eigenfaces for Recognition", *J. Cognitive Neuroscience*, pp. 71-86, March 1991
13. J. Kittler, M. Hatef, R.P.W. Duin and J. Matas, "On combining classifiers", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 20, No. 3 pp. 226–239, March 1998
14. C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.
15. F. Talantzis, A. G. Constantinides, and L. Polymenakos, "Estimation of Direction of Arrival Using Information Theory," IEEE Signal Processing, vol. 12, no. 8, pp. 561-564, Aug. 2005.
16. A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," Neural Comput., vol. 7, pp. 1129–1159, 1995.
17. T. M. Cover and J. A. Thomas, Elements of Information Theory. New York: Wiley, 1991.

18.J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-35, pp. 1661–1669, Dec. 1987.
19.J. L. Crowley, 'Context Driven Observation of Human Activity', in the Proc. of the European Symposium on Ambient Intelligence, Oct. 2003.
20.J. Soldatos, I. Pandis, K. Stamatis, L. Polymenakos, J. Crowley, 'A Middleware Infrastructure for Autonomous Context-Aware Computing Services', Computer Communications Magazine, special Issue on Emerging Middleware for Next Generation Networks, to appear 2006.
21.S. Azodolmolky, N. Dimakis, V. Mylonakis, G. Souretis, J. Soldatos, A. Pnevmatikakis, L. Polymenakos, 'Middleware for In-door AmbientIntelligence: The PolyOmaton System', in the Proc. of the 2nd NGNM workshop, Networking 2005, Waterloo, Canada, May 2005.
22.J. Soldatos, L. Polymenakos, A. Pnevmatikakis, F. Talantzis, K. Stamatis and M. Carras, 'Perceptual Interfaces and Distributed Agents supporting Ubiquitous Computing Services', in the Proc. of the Eurescom Summit 2005, April 2005, pp. 43-50.