A Modular Audio-Visual Scene Analysis and Attention System for Humanoid Robots

B. Kühn, B. Schauerte, R. Stiefelhagen
Institute for Anthropomatics,
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
e-mail: {kuehn,schauerte,rainer.stiefelhagen}@kit.edu

K. Kroschel
Fraunhofer Institute of Optronics, System
Technologies and Image Exploitation (IOSB)
Karlsruhe, Germany
e-mail: kristian.kroschel@iosb-extern.fraunhofer.de

***Abstract*—We present an audio-visual scene analysis system, which is implemented and evaluated on the ARMAR-III robot head. The modular design allows a fast integration of new algorithms and adaptation on new hardware. Further benefits are automatic module dependency checks and determination of the execution order. The integrated world model manages and serves the acquired data for all modules in a consistent way. The system has a state of the art performance in localization, tracking and classification of persons as well as exploration of whole scenes and unknown items. We use multimodal proto-objects to model and analyze salient stimuli in the environment of the robot to realize the robots' attention.**

Keywords: scene analysis system, hierarchical entity-based exploration, audio-visual saliency-based attention, world model, humanoid robot head.

I. Introduction

For humanoid robots, the multimodal perception of the environment is an essential and challenging task. Beside the perception of everyday objects and persons, the cognition of salient stimuli, i.e. acoustic events and visual attractable objects, is important for the robot attention. New and previously unknown objects have to be detected and further analyzed, to be able to recognize them again. An information storage for detected objects (and persons) and a-priori information has to be realized.

To solve the described tasks, we developed a modular audio-visual scene analysis and attention system. Acoustic and visual sensors provide the input data for all further processing steps: We combine different algorithms for person detection and identification (face detection and identification; acoustic speaker localization and identification). The object localization and classification is done using specific object characteristics (e.g., color, contour and texture). The overt attention is realized using saliency-based proto-objects. To this end, visually salient objects and acoustically salient events are detected and represented as proto-object hypotheses. Subsequently, a spatio-temporal clustering fuses the hypotheses to proto-objects. Proto-objects with a high saliency attract the attention of the robot and trigger a further analysis. A hierarchical knowledge-driven analysis in combination with an entity-centered world model provides a consequent information refinement and keeps, in combination with a particle filter-based tracking and an aging algorithm, the system up-to-date. The modular structure, in combination with automatic module dependency and execution order estimation, makes it possible to simply extent the system with new approaches and/or exchange existing algorithms with better ones.

II. Related work

The acquisition and fusion of information for scene analysis has been addressed for various application areas with different sensor setups throughout the years (e. g. [1, 2, 3, 4, 5]; cf. [6]). In this paper, we give a detailed description of our system in continuation to [2] and in combination with our previous work about attention [7]. To this end, similar to human behavior (cf. [8]), we also consider information that is currently not available in the field of view of the robot, but available in the short-term memory of our world model. The analysis of scene elements (persons and items) and salient stimuli (e.g. acoustic events, visually prominent regions) are done by using a hierarchical, knowledge-driven analysis strategy (cf. [2]; [9]), which combines a bottom-up and top-down strategy (cf. [6] and [4], respectively). It is used in combination with an entity-based world model that follows the approach described in [10]. An integrated tracking of world model entities makes it possible to detect changes as well as to distinguish novel entities from already attended entities, which both can influence the entity analysis sequence as well as the duration and is inspired by the human behavior described in [8].

III. System

In this section, we provide a detailed description of the main system components: the entity-based world model and the system modules (III.A). Afterwards, we explain the hierarchical knowledge-driven analysis process of detected objects, persons or salient stimuli (III.B). Finally, we describe the algorithms used in the system for multimodal detection, classification and tracking of objects and persons as well as saliency estimation and give a brief overview of the employed head motion controls (III.C).

A. System Architecture

1) World Model: Our system architecture has a modular structure, which enables a flexible, exchangeable, and extendable organization of the data flow and processing. In order to organize the data, we use an entity-based world model that manages all necessary information in a consistent structure (see Fig. 1). For a clear description, we first introduce a consistent nomenclature: The system distinguishes between the real world and a representation of it in a world model. Persons and items (e.g. books, mixer) in the real world are summa-

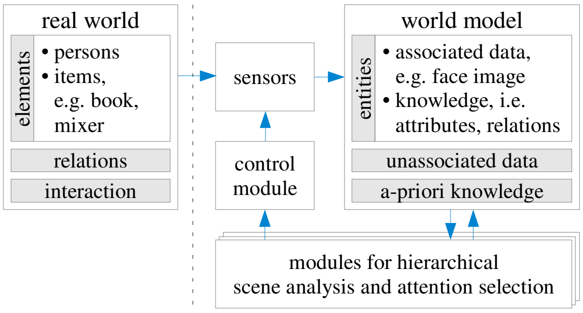
real world
elements
• persons
• items, e.g. book, mixer
relations
interaction
sensors
control module
world model
entities
• associated data, e.g. face image
• knowledge, i.e. attributes, relations
unassociated data
a-priori knowledge
modules for hierarchical scene analysis and attention selection

Fig. 1 Overview of the main system components, the naming and the relation between the real world and world model. The most modules are responsible for analyzing entities and keep the focus of attention.

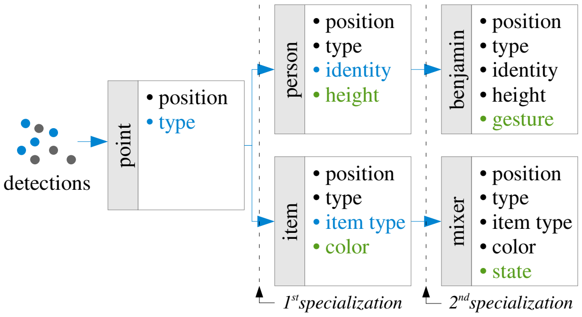

Fig. 2 Example shows the principle of the hierarchical analysis approach, where the blue attributes trigger a refinement in the hierarchy and the green attributes supply additional information about an entity.

rized as elements, whereas their representations in the world model are named objects or entities. The term "objects" is only used as a synonym for items, if the context is totally clear. A salient stimulus (i.e. acoustically salient event or visually attractable region) is represented by a proto-object and is a candidate for an entity, that can be further analyzed (see Sec. III.B).

The world model is split into several functional parts: The most important part of the world model is the set of available entities. They are representations of real world elements and consist of attributes/relations, summarized as "knowledge", and associated data (see Fig. 1). Both are generated at run-time by the system modules (see Sec. III.A.2). The characteristics of each entity are described by its attributes. All attributes have a name, value and confidence. The latter is a reliability measure for the attribute value. Relations are defined in form of geometric relations (e.g., isOn) or class relationships (e.g., isPerson). Furthermore, each entity has associated data, i.e. data that is assigned only to a specific entity and necessary to generate entity knowledge. Associated data has one producer module and multiple consumer modules. Additionally to the set of entities, the world model also includes so called unassociated data, which cannot be assigned to one specific entity. Examples for this aspect are sensor raw data or low-level data required by multiple entities. The last part of our world model is formed by the a-priori knowledge, e.g. geometric information about the environment or models to classify elements.

We decide to represent salient stimuli in our world model as proto-objects, similar to our already existing entities and use them as candidates for our entities, i.e. if a proto-object is selected through the Focus-of-Attention (FoA) selection process (see [7]), it will be converted into an entity and analyzed accordingly (see Sec. III.B). The bases for proto-objects are acoustic and visual proto-object hypotheses, which are fused using a spatio-temporal clustering (see Sec. III.C.1). The resulting multimodal proto-objects have at least two attributes: spatial location and saliency measure. The latter is important for the FoA selection process. The attention towards visually salient regions and acoustically salient events is important for a reactive robot.

An important functionality of our world model is attribute aging, which means if information about an attribute is not confirmed within a certain period of time, the confidence of the attribute decreases. If the confidence exceeds a deletion threshold and cannot be reconfirmed before, the attribute is deleted. This can also happen for e.g. the identity of a person. If we can roughly track a person, but cannot reconfirm the identity, the confidence of the identity decreases until it is deleted and the entity represents only a person without an identity.

2) Modules: Another important aspect of our system is its modularity. The modules are essential for the functionality of the scene analysis approach (see Sec. III.B). The system modules are organized in five task specific groups, i.e. detection/localization, classification, tracking, information fusion, and general purpose. The interactions and dependencies between modules in such a flexible system are hard to manage without suitable automatic mechanisms that solve the problem. Every module has two types of dependencies: Knowledge dependencies specify the required information that entities need to provide in order to be processable by the module (e.g., specific entity classes, geometric relations), whereas data dependencies determine the data flow between the modules and are not necessarily associated to entities. Data dependencies then indicate an entity independent data flow, most importantly raw or pre-processed sensor data. To solve the dependencies, a directed acyclic graph is calculated for each dependency type and combined to form a global dependency graph. If the resulting graph is acyclic, a well-defined execution order of the modules exists and is calculated by applying topological sort [11]. The algorithm also determines parallel module execution paths, which are used for automatic module-level parallelization in order to optimize the system performance. An example for knowledge and data dependencies including the resulting execution order can be found in Sec. IV.C.

B. Hierarchical, Knowledge-Driven Analysis

The analysis of each entity is hierarchically organized according to a coarse-to-fine strategy. During the analysis, the level-of-detail of an entity increases (proportionally to the number of analyzed entity attributes), whilst the degree-of-abstraction decreases, i.e. the next specialization step in the hierarchy is reached (see Fig. 2). To this end, the system dynamically acquires and fuses the necessary information about the entity attributes and stores them in the world model. Most importantly, this hierarchical mechanism is efficient as well as flexibly extendable, because specialized classification (and fusion) modules can be easily added – with respect to their knowledge dependencies – to the hierarchy. Each module is then responsible for the target attribute analysis and entity specialization at the corresponding node in the hierarchy.

The knowledge-driven analysis is an efficient way to decide which module needs to be executed at a specific point in time. More precisely, from the already identified class relationships of an entity, a list of available attributes is generated. Each of these attributes can be created from one or more classification modules and the system automatically activates the corresponding classification and information fusion modules. If the attribute confidence exceeds a pre-defined classification threshold, the analysis of the next step in the hierarchy is triggered. Consequently, new modules are activated, new attributes are generated, and the level in the hierarchy increases again. This is repeated until there are no further specialization steps currently available in the hierarchy or the analysis is stopped by another high-level process.

Fig. 2 shows an example for person and item processing. The first step starts always with the detection and the creation of a basic entity ("point") with a position attribute. Now all modules that can operate on this basic entity are activated. The estimation of a new attribute ("type") is the next step. Subsequently, the confidence of the type raises and if it is high enough the next step in the hierarchical analysis is reached (in this example person or item). Depending on the type, new modules are activated for this entity, a further specialization can be done and new attributes are generated. The number of steps in the hierarchy and the available attributes depend on the available modules and can be extended in future applications.

Additionally to the previously described analysis approach, a top-down mechanism is introduced, which is responsible to keep the balance between the exploration of the whole scene and the analysis of a specific entity. This is necessary, because the complete analysis of an entity can require a considerable amount of processing resources and, most importantly, time, which is especially critical in dynamic scenes. Thus, the time available to analyze each entity can be dynamically limited depending on run-time constraints, the level of awareness, the type of each entity (e.g., person, item), and the number of available proto-objects. If the number of entities and/or proto-objects is high, a subset can be marked – supported by the integrated tracking – for a further analysis at a later time. Furthermore, descriptions of unknown items can be acquired and added to the world model, e.g., via text input or multimodal interaction (cf. [12]).

C. Algorithms

We use a combination of various algorithms in our system for item localization and classification as well as person detection and identification. The applied approaches can be easily exchanged or new algorithms can be added to increase the performance or to add new functionalities. Generally, we are using a stereo-camera setup to be able to estimate a corresponding 3-D position for a given image position. Additionally, we use a microphone array in order to perform a 3-D localization of acoustic events (speech or object sounds).

1) Saliency: In order to estimate the visual saliency of a scene we use quaternion discrete cosine transform (QDCT, see [13]) image signatures of a 4-channel image consisting of intensity, blue-yellow and red-green color opponents and motion (estimated as difference of two successive frames). Afterwards, we use the isophote curvature [14] to estimate regions with potentially high saliency. In contrast to that, the acoustic saliency is estimated using acoustic surprise (see [7]), which bases on Bayesian surprise and uses the detection of relevant changes in spectrogram of an audio signal. Subsequently, a localization of the sound source is performed using SRP-PHAT (see Sec. III.C.2). Both modalities create proto-object hypotheses as 3-D representations. Afterwards, a spatio-temporal fusion is used to estimate proto-objects as candidates in the Focus-of-Attention selection process (see [7]).

2) Acoustic localization/classification: In order to estimate the position of a sound source, we are using the time difference of arrival between each possible microphone pairs of a microphone array. The combination of all pairs leads to a map of weighted possible source positions. The steered response power with phase transform (SRP-PHAT) is the algorithm behind this approach. We implemented an adapted SRP-PHAT version with an additional PHAT parameter β (SRP-PHAT-β, see [15]), which leads in average to a higher localization accuracy, but has a higher computational effort. The classification of speaker and object sounds are based on mel-frequency cepstral coefficients (MFCC) extracted from the audio signal of each microphone channel. Subsequently, the classification is performed using Gaussian Mixture models (see [16]) for objects. Persons are identified using adapted universal background models (UBM; see [16]). The general UBM represents an universal model for speakers. The model for a new person is generated by adapting the UBM with person specific MFCC features.

3) Visual person identification/item classification: In order to detect a person, we use a MCT-based face detector (see [17]) instead of the widely used Viola-Jones approach, because the first approach has a lower false alarm rate and the implementation is faster. The identification is done with a DCT-based approach comparable to [18], which is suitable for a defined closed data set scenario. For item detection and classification we use a combination of different algorithms, e.g. color histograms, textures and location-based approaches. But it can easily be extended with new or more powerful algorithms that may have higher computing requirements.

4) Tracking: Another important part is the tracking of entities which depends on the entity type. For items, a very effective and simple median tracker is used. According to our experience, this tracker type is sufficient in most encountered situations in the lab. For persons, an interacting Markov Chain Monte Carlo Particle Filter (IMPF) is used [19], which is more suitable for quick and unpredictable person movements. The number of particles depends on runtime and accuracy constrains. More particles normally generate better results, but have higher computational costs.

5) Head Control: The gaze direction of the robot head is essential for the exploration of an unexplored environment. It is characterized by a two-dimensional vector of pointing angles, which can be expressed as a function of the head kinematics. Using the associated differential kinematics, a gaze control concept is implemented, which allows the visual tracking of moving targets with unknown and arbitrary trajectory, using all available degrees-of-freedom (DoF) of a robot head, in order to obtain a natural human-like behavior. The realized control concept is presented in detail in [20] and is composed by a feedback control action proportional to the pointing error and a feedforward control based on the predicted motion of the target.

IV. Experimental Evaluation

A. Setup

We tested our system on two different hardware platforms (see Fig. 3). First, we have an experimental platform with two stereo cameras, a microphone array and a pan-tilt unit (PTU). Second, we evaluated the system using the ARMAR-III humanoid robot head, which has a comparable sensor setup. We are using two stereo cameras in order to perceive the scene with different levels-of-detail. One stereo camera is used for near-field and the other for far-field image acquisition. A referencing between the stereo cameras is able to find a corresponding element, detected in one camera, in the other camera. Detailed information for each platform is summarized in Table I.

B. Procedure

The complexity of the system and the number of components is high, so that a detailed evaluation of every aspect of the presented system would go beyond the scope of this paper. Instead, we want to show selected functionalities of the system in a comprehensive description. We have evaluated our system with various combinations of items and persons in a scene and achieved good results. In this contribution, the multimodal perception of a person is the example for explaining all subsequent functionalities. In order to demonstrate the effective system approach, we finally show typical identification steps for a person with real data and the corresponding results for tracking.

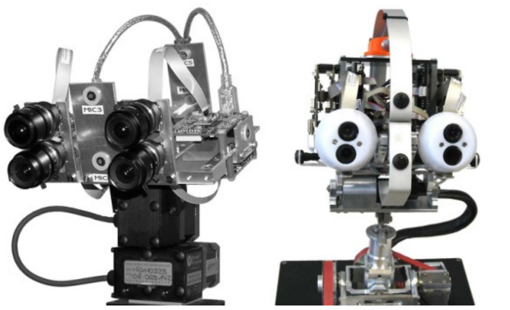

Fig. 3 Experimental hardware platforms used for the evaluation of the system. A stereo camera setup with microphones and pan-tilt unit (left) and the ARMAR-III humanoid robot head (right).

TABLE I
HARDWARE AND SOFTWARE PARAMETERS OF BOTH PLATFORMS

	PTU sensor setup	ARMAR-III head
stereo cameras	2	2
- focal lengths	3.5 mm / 6 mm	4 mm / 12 mm
- resolution	640×480 pixels	640×480 pixels
- frame rate	30 fps	30 fps
microphone array		
- microphones	6 omi-directional	6 omi-directional
- sampling rate	48 kHz	48 kHz
degree-of-freedom	2	7

C. Results and Discussion

1) Attention: The ability to react on salient stimuli is an important part of your system. However, we described and evaluated this behavior already in our previous work (see [7]). Summarizing it can be said, that we are able to detect visual salient regions using spectral whitening of images and detect acoustically salient events using acoustic surprise. We showed that a subsequent alignment of the sensor towards salient stimuli helps to improve the perception quality. In the last part of the evaluation, we used a complex scene as an example for detection, tracking and prioritization of salient stimuli.

2) Dependencies and Execution order: In this section, we want to show how the knowledge and data dependencies can be used to estimate the module dependencies and the execution order including parallel execution paths. An overview of the whole process shows Fig. 4. The first step is the acquisition of the sensor data, which is in our example the acquisition of a scene image from the camera and an audio frame from the microphone array. The raw sensor data is stored in the entity-independent unassociated short-term memory of the world model. The next steps are the detection of faces in

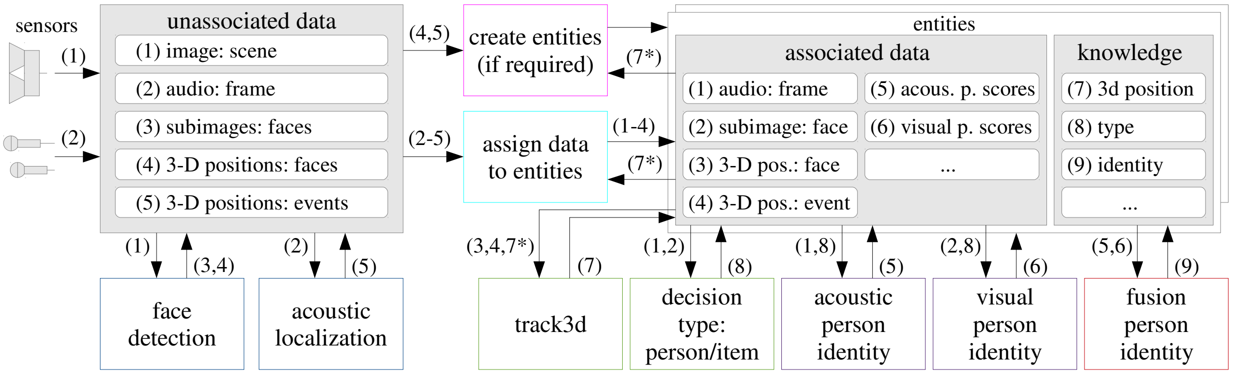

Fig. 4 An example for person identification that shows the relationships between the modules, the data flow and the data as well as knowledge dependencies (*=value of previous step). This example shows the functional principle of the system and represents a part of the whole system.

the image and the detection of speech and other sounds in the audio frame using the appropriate modules (blue). Subsequently, we estimate the individual 3-D positions for a later tracking. All generated data is also stored in the unassociated memory. The third step is to decide whether the data belongs to new entities (and create them; magenta) or to already existing ones. In both cases, we assign the suitable data to the corresponding entities (cyan). All following modules use only associated data as well as knowledge and they are executed for each entity once. The next step is to decide whether the entity is a person or an item (green). This is the first specialization step (see Fig. 2) and can be easily done for persons in the case of enough face detections and/or acoustic person classification, with a high confidence. In parallel, the position of each entity is tracked (green) over time in order to follow the movements and be able to assign the unassociated data in the next processing loop. In order to decide which person is visible in the camera or is speaking at the moment, two further modules for classification are executed (purple). They generate normalized classification scores for all available person models. The scores are used in the fusion process to estimate to persons' identity (red). This is the second specialization step (see Fig. 2).

When we consider the dependencies shown in Fig. 4, we can see the different types: All arrows from and to the unassociated data block are the generated and required unassociated data dependencies of the modules. The same applies for the associated data and knowledge dependencies of the entities. The numbers in brackets refer to the fields in the blocks and a star indicates data from the previous analysis loop. With this information, we can create a global dependency matrix and use topological sort (see [11]) to estimate the execution order shown in Fig. 5. Each line represents one time step and parallel executable modules are shown side by side. All entity dependent modules can be executed for each entity in parallel and within these also parallel, if the modules are independent (e.g., acoustic/visual person identity).

3) Entity Analysis and Tracking: In order to demonstrate the previously described functionalities, we provide the following results of the real system. As above, the task is the detection, identification and tracking of a person in a scene as an example for the whole approach.

First, we want to show the detection and identification functionality. Therefore, we plotted in Fig. 6 the class attributes over time of the detected person (Benjamin). As one can see, the entity creation is done within a second and the confidence rises quickly. After two seconds, the entity type (person) can be estimated and the identification starts. In the next seconds, the system tries to identify the person, which is at the beginning sometimes wrong, because of only a few or bad detections, but the confidence of this statement is very low. Three seconds later, the correct person can be identified. When the person starts moving, the confidence of the identity may decrease by a small amount, because of the noisier sensor data.

Second, the confidences for the most important person attributes over time are shown in Fig. 7. Generally, the confidence of all attributes is high and the increase at the beginning is sharp. The identity is plotted (purple) to show the pure confidence and the relation to other attributes. After five seconds the confidence exceeds the specified threshold. The estimated height of the person (cyan) starts also quickly and is nearly constant until the person is going for a short period into a kneeling position. The next attributes represent acoustic (red) and visual (green) presence of detections. If the person is not perceivable, the confidence starts to decrease during the attribute aging. The last curve (blue) represents the robust confidence of the person position.

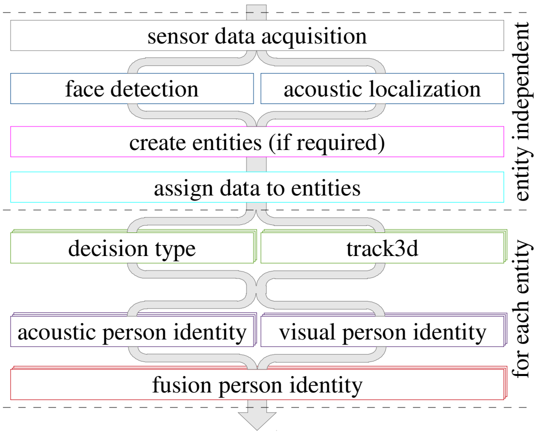

Fig. 5 Module execution order and parallel execution paths are estimated using topological sort of data and knowledge dependencies. Only-entity-dependent modules can be additionally parallelized, by executing the paths for each entity at the same time.

Finally, we evaluated the tracking of the person. In Fig. 8, the trajectory of the person (red) for each dimension can be found, including the separate visual and acoustic detections (blue and green). Because of the higher accuracy of the visual detections and the higher variance of the acoustic localization, the particle filter has a sensor specific weighting of the detections. As it can be seen easily, a tracking of the person during the whole sequence is achieved. In Fig. 9 the same complex trajectory is shown from a different point of view.

V. Conclusion And Future Work

We presented a system for audio-visual scene analysis, which enables a saliency-driven exploration. The concept of a world model centered approach, combined with an intelligent module-based processing, is the basis for the exploration and analysis process. The automatic integration of new algorithms with the help of the proposed data and knowledge dependencies enables a flexible and extendable analysis approach. An automatic estimation of the execution order is a great benefit for a flexible extension of the system. As future work, we plan to add further aspects of attention and integrate an inquiring behavior.

Acknowledgment

This work has been supported by the German Science Foundation (DFG) within the Collaborative Research Centre 588 "Humanoid Robots – Learning and Cooperating Multimodal Robots".

REFERENCES

[1] B. Schauerte, J. Richarz, T. Plötz, C. Thurau and G. A. Fink, "Multi-Modal and Multi-Camera Attention in Smart Environments," in *Proc. Int. Conf. on Multimodal Interfaces*, 2009.

[2] B. Kühn, A. Belkin, A. Swerdlow, T. Machmer, J. Beyerer and K. Kroschel, "Knowledge-driven opto-acoustic scene analysis based on an object-oriented world modelling approach for humanoid robots," in *Proc. Joint Conf. of 41st Int. Symp. Robotics and 6th German Conf. Robotics*, 2010.

[3] I. Essa, "Ubiquitous sensing for smart and aware environments," *IEEE Personal Communications,* vol. 7, pp. 47-49, 2000.

[4] J. Holsopple and S. Yang, "Designing a data fusion system using a top-down approach," in *Proc. Int. Conf. on Military Com.*, 2009.

[5] H. Wersing, S. Kirstein, M. Götting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J. Steil, H. Ritter and E. Körner, "Online Learning of Objects in a Biologically Motivated Visual Architecture," *Int. Journal of Neural Systems,* pp. 219-230, 2007.

[6] D. Hall and J. Linas, Handbook of Multisensor Data Fusion: Theory and Practice, CRC Press, 2008.

[7] B. Schauerte, B. Kühn, K. Kroschel and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *Proc. Int. Conf. on Intelligent Robots and Systems*, 2011.

[8] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends in Cog. Science,* pp. 498-504, 2003.

[9] T. Machmer, A. Swerdlow, B. Kühn and K. Kroschel, "Hierarchical, knowledge-oriented opto-acoustic scene analysis for humanoid robots," in *Proc. Int. Conf. on Robotics and Automation*, 2010.

[10] M. Baum, I. Gheta, A. Belkin, J. Beyerer and U. D. Hanebeck, "Data Association in a World Model for Autonomous Systems," in *Proc. Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems*, 2010.

[11] A. B. Kahn, "Topological sorting of large networks," *Commun. ACM,* vol. 5, no. 11, pp. 558-562, November 1962.

[12] B. Schauerte and G. A. Fink, "Focusing Computational Visual Attention in Multi-Modal Human-Robot Interaction," in *Proc. Int. Conf. on Multimodal Interfaces*, 2010.

[13] B. Schauerte and R. Stiefelhagen, "Predicting Human Gaze using Quaternion DCT Image Signature Saliency and Face Detection," in *Proc. of IEEE Workshop on the Applications of Computer Vision (WACV)*, 2012.

[14] J. Lichtenauer, E. Hendriks and M. Reinders, "Isophote Properties as Features for Object Detection," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, 2005.

[15] K. D. Donohue, J. Hannemann and H. G. Dietz, "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments," *Signal Process.,* vol. 87, pp. 1677-1691, 2007.

[16] A. Swerdlow, T. Machmer, B. Kühn und K. Kroschel, „Robust sound source identification for a humanoid robot," September 2008.

[17] B. Froba and A. Ernst, "Face detection with the modified census transform," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 2004.

[18] H. K. Ekenel and R. Stiefelhagen, "Local appearance-based face recognition using discrete cosine transform," in *FRGC 2.0 Database, Face Recognition Grand Challenge Workshop (FRGC)*, 2006.

[19] L. Ying and P. Vadakkepat, "Interacting MCMC particle filter for tracking maneuvering target," *Digitial Signal Processing,* vol. 20, no. 2, pp. 561-574, 2010.

[20] G. Milighetti, A. De Luca and L. Vallone, "Adaptive Predictive Gaze Control of a Redundant Humanoid Robot Head," in *Proc. Int. Conf. on Intelligent Robots and Systems*, 2011.

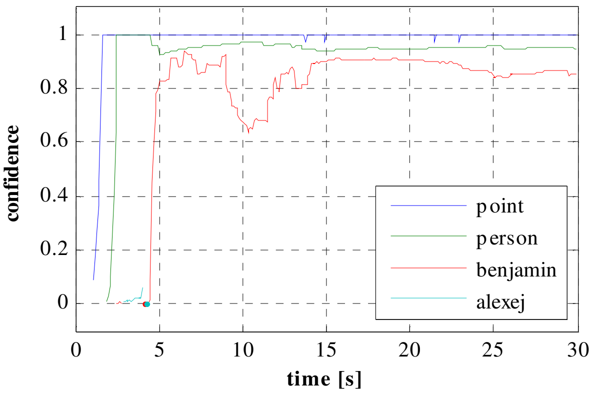

Fig. 6 Class confidences for the detected person (benjamin) within the first 30 seconds (includes false classifications with low confidence).

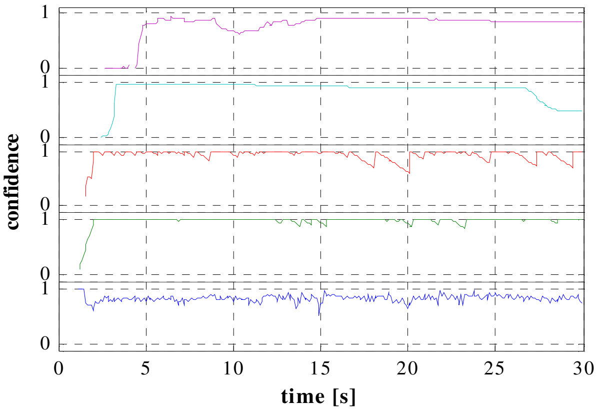

Fig. 7 Attribute confidences for identity, height, acoustic observation, visual observation and position (from the top to the bottom).

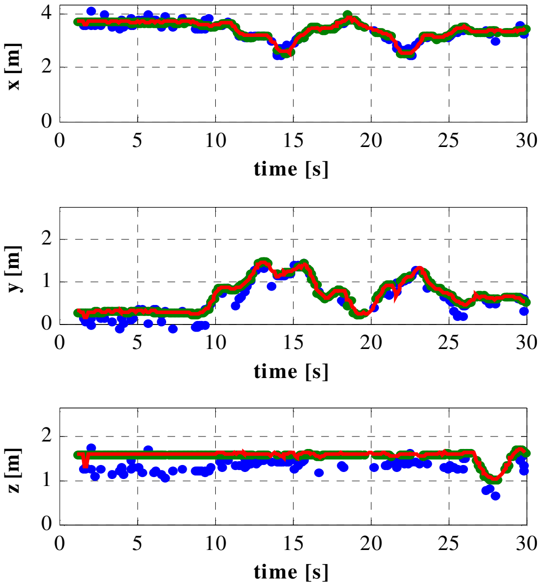

Fig. 8 The persons' trajectory for each dimension including the acoustic localizations (with a higher inaccuracy) and face detections.

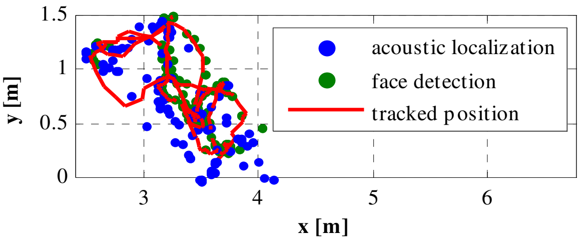

Fig. 9 The persons' trajectory including the acoustic localizations and the face detections (top view).