Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Simple Auditory and Visual Features for Human-Robot Dialog Scene Analysis

Rujiao Yan, Tobias Rodemann, Britta Wrede

2012

Preprint:

This is an accepted article published in Int. Conf. on Intelligent Robot and Systems (IROS). The final authenticated version is available online at: https://doi.org/[DOI not available]

Simple Auditory and Visual Features for Human-Robot Dialog Scene Analysis

Rujiao Yan[1,2], Tobias Rodemann[2] and Britta Wrede[1]

***Abstract*— This paper presents a system that uses various simple auditory and visual features to achieve human-robot dialog scene analysis. Our scene analysis system is able to learn how many speakers are in the scenario, where the speakers are and who is currently speaking. Speakers are unknown in advance. A visual short-term-memory (STM) helps to memorize persons, even if they disappear from the camera's field of view for a while due to movements of persons or the robot head. In comparison to our previous work, we apply more visual features such as height, color and texture features of different upper body parts, to improve the scene representation performance. We show that our system is able to assign words to corresponding speakers. A speaker is recognized again when he leaves and enters the scene, or changes his position even with a newly appearing person.**

I. INTRODUCTION

In most real-world situations, a robot is interacting with multiple people. In this case, understanding of the dialogs is essential. However, dialog scene analysis is missing in most existing systems of human-robot interaction. In such systems, only one speaker can talk with the robot or each speaker wears an attached microphone or a headset. Our target is therefore a system of Computational AudioVisual Scene Analysis (CAVSA) that makes dialogs between humans and robots more natural and flexible. CAVSA aims at learning the number and position of interacting partners, as well as who is currently speaking. It can be used as a complementary analysis step for applications like speech recognition.

Audiovisual integration is an important issue for CAVSA. A lot of work has been done related to audiovisual integration in multi-person environments for different applications. The first group focuses on temporal coincidence between speech and visual activity (lip motion or body motion) [1]–[5]. These methods need only one camera and one microphone, and they do not make a prior measurement of auditory and visual position. Their main application is speaker indexing in multimedia data for speaker diarization (annotation of "who spoke when") or as preprocessing for multimodal person identification. Most of them can only work in constrained environments to avoid artifacts, e.g. non-talking persons are not allowed to move significantly. Therefore, these methods usually require careful tuning of quantities such as camera image resolution and temporal resolution. The second group integrates auditory and visual signals based on their position information [6]–[9]. The methods are usually proposed for multi-person tracking. They use binaural microphones or microphone arrays to localize sound, and estimate 3D visual positions using stereo cameras or even more cameras in a smart-room environment. Since most robots have a limited field of view, a speaker may leave a scenario or stand outside the field of view for a while and reappear again. In these cases, these methods can not memorize the speaker and would consider him as a new person. Bebewitz et al. [10] proposed an approach using a probabilistic belief about people in the surroundings of the robot. While the robot head is moving, a person may leave the field of view briefly, but his face is still stored in the belief and the robot can find him again later. However, since the observation assignment to stored faces is purely based on their positions, the method fails if a person greatly alters his position outside the field of view.

In our system, the scene is represented with many simple auditory and visual features using the concept of proto-objects [11], [12]. Proto-objects can combine an arbitrary number of features in a compressed form. Audio and visual proto-objects for the same speaker are then grouped together in their respective short-term-memories (STM). One advantage of using STM is to memorize persons who disappear for a while. Finally, audio and visual proto-objects are matched based on their position information. Since speakers are unknown in advance, we can not train auditory and visual features beforehand to recognize voice and face. Furthermore, high-level features for voice and face recognition exhibit an inherent computational complexity which makes real-time processing difficult. Therefore, we combine many low or mid level auditory and visual features to achieve a high performance of audiovisual integration. Additionally, while current approaches in Computational Auditory Scene Analysis (CASA) mostly deal with parallel sources using microphone arrays (see e.g. [13]), we focus on purely sequential sounds. Our system could be used after sound source separation in case of concurrently active sources.

Our previous work in [14] introduced an original version of CAVSA which was tested in online adaptation of audio-motor maps. It can correctly determine the number and position of speakers as well as the current speaker in typical human-robot dialog scenarios. However, it works under the same constrains as [10], because only spatial coincidence was used to group proto-objects in STM. In this work, we employ therefore more visual features like height, color and texture of different upper body parts to improve the capability of CAVSA. The presence of persons is detected by searching

*This work was supported by the Honda Research Institute

1 Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, 33594 Bielefeld, Germany {ryan,bwrede}@cor-lab.uni-bielefeld.de

2 Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073 Offenbach, Germany tobias.rodemann@honda-ri.de

for their faces. Based on face position and size, areas of hair, collar and clothes are estimated.

Our system uses only a humanoid robot head with a pair of cameras and a pair of microphones. The head is mounted on a pan-tilt unit.

II. Auditory and Visual Features

In our system, the scene is represented with many simple auditory and visual features using the concept of proto-objects. Proto-objects are a psychophysical concept and are considered here as a compressed form of various features [15]. Proto-objects can be tracked, pointed or referred to without identification and enable a flexible interface to behavior-control in robotics. For more information about audio and visual proto-objects see [11], [12].

A. Audio Proto-Object

We employ a Gammatone Filterbank (GFB) with 100 frequency channels that span the range of 100-11000 Hz for auditory preprocessing [12]. After that, the envelope signal for every frequency channel is extracted and signal energy is calculated as the sum over envelope signals in all channels. Based on energy, audio streams are segmented and an audio segment begins when the signal energy exceeds a threshold and ends when the energy falls below this threshold. Then we derive start time, length and the mean energy over all samples of the segment, and save them in an audio proto-object. Next, a filtering of audio proto-objects based on segment length and energy is performed, since background sounds are rather short and have less energy than speech directed at the robot. Filtering of background sounds is explained in [12]. In addition, an audio proto-object contains also a position evidence vector $p_a(\theta)$ for each azimuth angle θ (-90, -80, ..., 0, ..., 80, 90). The peak position in the position evidence vector is the azimuth angle of the audio proto-object. To acquire the position evidence vector, we compute the population-coded binaural cues ITD (Interaural Time Difference) and IID (Interaural Intensity Difference), and compare them with audio-motor maps that describe the relationship between binaural cues and sound position. For more details about the sound localization system see [12] and about representation of sensory information by population coding see [16], [17].

B. Visual Proto-Object

In the camera field of view, the presence of persons is detected by searching for their faces. We assume that a person talking to the robot looks at its face most of the time. A frontal face detection algorithm developed by Viola and Jones [18] is applied for its high speed and high detection rates to extract visual proto-objects in multi-person scenarios. For each of these proto-objects, the center and size of each face in the camera image is computed.

1) Position evidence vector in spherical coordinates: To keep the same metric as audio position, visual camera positions need to be converted to position evidence vectors in spherical coordinates. For simplicity we use only one camera and keep the distance between robot head and speakers about $1m$. Using the distance information and saccade maps (see [19]), we can convert camera position into 3D world coordinates and spherical coordinates (azimuth and elevation). Instead of this fixed setting, a lot of algorithms for depth estimation using stereo cameras could be applied to obtain 3D position. The estimated azimuth angle θ_v is then converted to a position evidence vector $p_v(\theta)$ for each angle θ (-90, -85, ..., 0, ..., 85, 90). As the visual position has higher precision than the auditory one, visual angles are more densely sampled. $p_v(\theta)$ is defined by a Gaussian function:

$$p_v(\theta) = \exp\left(\frac{-(\theta-\theta_v)^2}{2\cdot\sigma_\theta^2}\right), \qquad (1)$$

where σ_θ is measured empirically to tolerate position measurement error of face detector and conversion error from camera coordinates to world coordinates. Furthermore, since position is one among many features to group visual proto-objects in STM, slight head movements should also be tolerated. In consideration of all above aspects, we set $\sigma_\theta = 5$ in our experiment.

Similarly, elevation angles are also converted to position evidence vectors for each elevation angle θ' (-30, -25, ..., 0, ..., 25, 30) and applied as height feature. The deviation $\sigma_{\theta'}$ is set to 3 under consideration that the vertical head movements are normally slighter than horizontal.

2) Face-based detection of different upper body parts: People normally do not change clothes or hairstyle during a dialog scenario. Hence we use color and texture features of different upper body parts (hair, collar and clothes). The border region between the face and hair region is more useful for differentiating between people than the bulk texture and color of a person's hair on its own [20]. Hence the selected hair area should contain most of the border region and less background. Similarly, a collar region is extracted as the border between neck and clothes which has much clothes style information. For instance, a shirt, coat and blouse can be distinguished only by collar.

These upper body parts are extracted according to the position and size of the detected faces. The width and height of each region of interest (ROI) are proportional to the one of the corresponding face ROI, and the distance to face ROI is

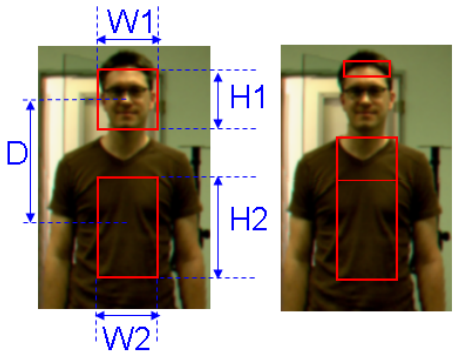

Fig. 1. Example: *left*: parameters for clothes ROI estimation; *right*: hair, collar and clothes ROI



Fig. 2. Probands for visual feature testing

proportional to the face height. Fig. 1 *left* shows an example for determining clothes ROI. Three ratio parameters $r_W = \frac{W_2}{W_1}$, $r_H = \frac{H_2}{H_1}$ and $r_D = \frac{D}{H_1}$ are applied, where W_1, H_1, W_2, H_2 and D are the width and height of the face ROI, width and height of the clothes ROI, the distance between the face and clothes ROI, respectively. Using the criterion that most of the desired content and less background appears in the ROI, these three parameters are chosen by training on several images of ten subjects (see Fig. 2). r_W, r_H and r_D for each ROI are shown in TABLE I and Fig.1 *right* illustrates an example of hair, collar and clothes ROI which are estimated using these parameters. Similar methods of face-based costume detection can be found in [21], [22].

	Hair	Collar	Clothes
r_W	0.8	1	1
r_H	0.3	0.8	1.67
r_D	0.5	1	2.1

TABLE I

Parameters for estimation of hair, collar and clothes ROI

3) Histogram feature vector of color and texture: Histograms are among the most common features used in various tasks such as pattern recognition, tracking and segmentation. We calculate color histograms in RGB color space as well as Local Binary Pattern (LBP) histograms for above mentioned upper body parts. To simplify the description, we call these features Hair RGB, Hair LBP, Collar RGB, Collar LBP, Clothes RGB and Clothes LBP respectively in the following text.

For the color histograms, each color channel is discretized into six bins and there are a total of 216 (6^3) bins. The 3D histograms are described in form of a 1D vector with 216 elements. Histograms are normalized by the pixel number.

LBP is used to represent local structure for its simplicity and robustness to luminance variation. It is defined as an ordered set of binary comparisons of pixel intensities between a pixel and its eight surrounding pixels [23]. We use a subset called uniform patterns [23] which has 58 elements and represents most texture information such as corner, line and edge. Hence, the number of histogram bins is set to 58.

Then we test the performance of these six features for differentiating between ten subjects (see Fig. 2). We recorded 30 frames for each subject, 10 for training of the template and 20 as samples for testing. For each feature, every sample is compared with templates by computing their similarity s. Different metrics can be utilized to calculate vector similarity, for a survey see [24]. We use normalized inner product (NIP):

$$s = \frac{x \cdot y}{||x||\,||y||}, \tag{2}$$

where x and y stand for a feature vector in a sample and in a template, respectively. We compute the Error Rate (ER) for each feature:

$$\text{ER} = \frac{\text{total number of falsely assigned samples}}{\text{total number of samples}}.$$

The measured ERs for all histogram features and the height feature (position evidence vector of the face elevation angle) are shown in TABLE II. We can see that the best three

Feature	Collar RGB	Collar LBP	Hair RGB	Hair LBP
ER	35.11%	19.10%	37.69%	17.59%
Feature	Clothes RGB	Clothes LBP	Height	
ER	9.05%	24.62%	43.22%	

TABLE II

Error Rate (ER) for each feature in the test

features are Clothes RGB, Hair LBP and Collar LBP. Additionally, we use also the height feature for its low error correlation with Clothes RGB (falsely assigned samples with the height feature and with Clothes RGB do almost not overlap), although it has the worst performance on its own in the test. Then we save Clothes RGB, Hair LBP, Collar LBP and the height feature in a visual proto-object. The measurement of proto-object similarity S between a sample and a template is involving the fusion of feature similarity s_i ($i \in [1,F]$), where F is the number of features in a proto-object. The fusion operator could be for example the maximum, product or mean. We compute S as mean value of s_i:

$$S = \frac{1}{F} \cdot \sum_{i=1}^{F} s_i. \tag{3}$$

ER for the combined features is only 0.5%. To estimate the best similarity threshold of combined features, we calculate the TPR (true positive rate), FPR (false positive rate) and analyze the ROC (receiver operating characteristic). Fig. 3 illustrates ROC-curves for Clothes RGB, Hair LBP, Collar LBP, height and their combination. As can be seen, the curve of the combined features is closer to the perfect classification (TPR= 1 and FPR= 0). There are many criteria to compare ROC points. Here, we select the point which has the minimum distance to the perfect classification. The corresponding threshold is $\Theta_s = 0.767$. With the same criterion, we find also a pair of TPR and FPR for each individual feature as shown in TABLE III. We can see that the combination of features improves the recognition performance.

Feature	Clothes RGB	Collar LBP	Hair LBP	Height	combined
TPR	0.884	0.879	0.874	0.864	0.954
FPR	0.189	0.149	0.121	0.15	0.067

TABLE III

TPR and FPR with the minimum distance to the perfect classification on the ROC-curve for each individual feature and the combined features.

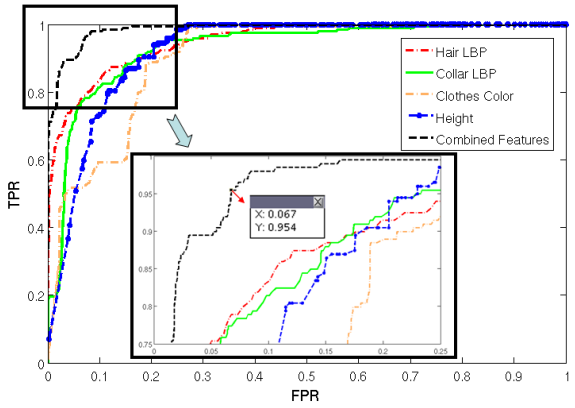

Fig. 3. ROC-curves for Clothes RGB, Hair LBP, Collar LBP, height and their combination.

III. Dialog Scene Analysis

In this section we introduce Computational AudioVisual Scene Analysis (CAVSA) for dialog scenarios. As shown in Fig. 4, we apply a visual STM together with the current audio proto-object as the inputs for audiovisual integration. We do not use an auditory STM in this work for two reasons. First, we have not yet tested auditory features to recognize people's voice. Second, grouping audio proto-objects using only position information in STM fails when people alter their positions greatly while not talking. Visual STM and audiovisual integration are described below.

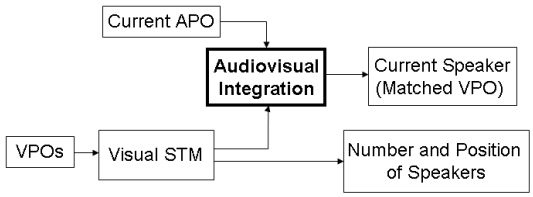

Fig. 4. System architecture of CAVSA for dialog scenarios. APO: audio proto-object, VPO: visual proto-object.

A. Visual short-term memory

We use a Short-Term Memory (STM) to group visual proto-objects for the same speaker together, and memorize proto-objects, even if their corresponding persons are out of sight for a while due to person's or robot's movement or failure of face detection. When a new proto-object appears, it is integrated with the related proto-object in STM or stored as a new entry. Proto-objects, which are not updated for more than a certain period T, are removed from the STM. The larger the parameter T, the longer a proto-object is stored in STM after the corresponding person disappears. If T is too small, STM can not remember a person who leaves for a while. Conversely, if T is too large, STM may contain too many proto-objects. We set $T = 100s$ in our experiment.

Let $X_m = (x_1, x_2, ...x_F)$ $(m \in [1, M])$ denote the m-th incoming proto-object and $Y_n = (y_1, y_2, ...y_F)$ $(n \in [1, N])$ denote the n-th proto-object in STM , where $x_i \in X_m$ (resp. $y_i \in Y_n$) stands for the i-th $(i \in [1, F])$ feature in X_m (resp. Y_n), F is the number of features in a proto-object, M the number of current incoming proto-objects and N the number of proto-objects in STM. Firstly, we compute similarity s_i between feature vector x_i and y_i based on equation 2. Next, similarity S_{mn} between proto-object X_m and Y_n is measured by:

$$S_{mn} = \frac{\sum_{i=1}^{F} s_i \cdot w_i}{\sum_{i=1}^{F} w_i}, \tag{4}$$

where w_i is the weight of the i-th feature and is defined as an activity function:

$$w_i = e^{-\frac{t_n}{T_i}}. \tag{5}$$

t_n represents the time in seconds since proto-object Y_n has been updated the last time. If Y_n is currently updated, $t_n = 0$ and thus $w_i = 1$. T_i is used to decay the activity. The longer the time a proto-object has not been seen, the more unreliable its features are. Since the position feature has a stronger variation than other features such as color and texture when people move, we set a smaller time constant $T_i = 5s$ for the position feature. For other features, T_i is set equal to $100s$. All similarities S_{mn} $(m \in [1, M],\ n \in [1, N])$ are combined into a matrix S' with M rows and N columns. Then an incoming proto-object should be linked to the proto-object with the maximum similarity, if the similarity is larger than a threshold Θ_s.

However, when more than one person appears in the camera image and more than one proto-object is stored in STM, two or more incoming proto-objects may tend to select the same proto-object in STM. To avoid the selection confusion, we use a simple greedy mapping method. First, the pair of proto-objects with the largest similarity is selected. If the similarity is larger than Θ_s, the two proto-objects are merged. Then their similarity related row and column are eliminated from similarity matrix S'. In the next step, the next most similar pair of proto-objects is selected. The process is repeated until all incoming proto-objects are assigned or the largest similarity is smaller than Θ_s. In the latter case, the remaining incoming proto-objects, which are not yet assigned, are stored in STM as new entries. A similar usage of such a greedy algorithm can be found in [2].

Assuming $X_{m'} = (x_1, x_2, ...x_F)$ and $Y_{n'} = (y_1, y_2, ...y_F)$ are the selected pair in time step t and their similarity $S_{m'n'}$ is larger than Θ_s, feature vector y_i in $Y_{n'}$ is updated by

$$y_i(t) = y_i(t-1) \cdot w_i + x_i(t) \cdot (1 - w_i), \tag{6}$$

where w_i is defined by equation 5. The duration of one time step is $0.1s$, since the frame rate of our visual system is 10 frames per second.

B. Audiovisual association

We link the current audio proto-object A to a visual proto-object V_j ($j \in [1,N]$) from the STM based on the similarity of their position evidence vectors p_a and p_{v_j}, to find the current speaker. First, auditory position evidence vector p_a is linearly interpolated to have the same length as the visual position vector p_{v_j}. The interpolated auditory vector is denoted as p'_a. Then the probability that A and V_j are caused by the same person is computed as the similarity of p_{v_j} and p'_a, as follows:

$$P_{common}(A,V_j) = \frac{p'_a \cdot p_{v_j}}{||p'_a||\,||p_{v_j}||}, \tag{7}$$

We also consider the situation where the current sound is related to a person who has not yet been seen. For convenience, we denote the proto-object of the unseen person as V_{N+1}. The probability that the current audio proto-object A is associated with V_{N+1} is described by:

$$P_{common}(A,V_{N+1}) = \frac{\sum_\theta U(\theta)\cdot p'_a(\theta)}{\sum_\theta p'_a(\theta)}, \tag{8}$$

where $U(\theta)$ is a view memory with elements between 0 and 1 for each azimuth angle θ.

$$U(\theta) = 1 - e^{-\frac{t_\theta}{T_u}}, \tag{9}$$

where t_θ indicates the time in seconds since position θ has been viewed the last time. The initial value of t_θ is set to ∞, so that $U(\theta) = 1$ if position θ has never been attended. The larger the value $U(\theta)$ is, the longer the time since position θ has not been attended. $U(\theta) = 0$ means that position θ is currently in the field of view. T_u is a time constant to decay the activity and it is set to $100s$ to match the parameter T.

Assume that visual proto-object V_{jMax} has the maximum probability $P_{common}(A,V_j)$ ($j \in [1,N]$). If $P_{common}(A,V_{N+1})$ is larger than $P_{common}(A,V_{jMax})$, the matched visual proto-object is considered to have not yet been seen. Otherwise, the audio proto-object is associated with visual proto-object V_{jMax}. More information about audiovisual integration can be found in our previous work [25].

IV. Results

To evaluate the performance of our audiovisual scene analysis, we carried out experiments in a small room. Three dialog scenarios were designed as shown in Fig. 5 to test if our system is able to correctly determine the number and position of speakers, as well as the current speaker. The distance between robot and speakers is about $1m$. Person 1, 2 and 3 are represented with red, blue and black icons. Red, blue and black dashed lines stand for moving trajectories of person 1, 2 and 3, respectively.

For each scenario we compared our system using position and introduced visual features (Hair LBP, Collar LBP, Clothes RGB and height) with our previous work in [14] which used only position to group visual proto-objects in STM. To simplify the description let us denote the previous system using only position as "*CAVSA-Pos*" and the current version using multiple visual features as "*CAVSA-Multi*".

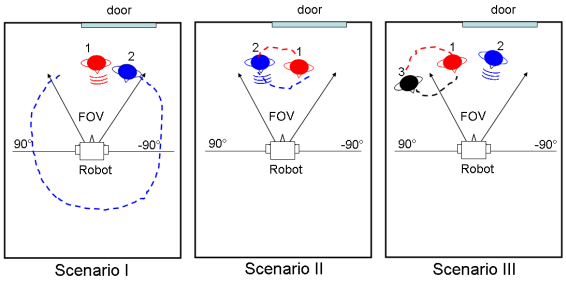

Fig. 5. Sketch of three scenarios. **Scenario I**: two participants talk to the robot alternatingly at the beginning. Then speaker 2 leaves the field of view and passes behind the robot and returns; **Scenario II**: two speakers change their positions; **Scenario III**: person 1 and person 2 talk to the robot alternatingly, while person 3 is outside the field of view and remains silent. Then person 1 and person 3 change their positions.

Threshold of similarity S_{mn} (in equation 4) was set to $\Theta_s = 0.78$ in "*CAVSA-Pos*" to tolerate slight movements of speakers such as head shaking. With $\Theta_s = 0.78$, about 8 cm face position difference in 0.04 s (between two frames) is tolerated, the corresponding movement speed is ca. 2m/s. For "*CAVSA-Multi*", we already get a threshold for combined visual features, without consideration of the position feature, as $\Theta_s = 0.767$ by analyzing the ROC in section II-B.3. Since above mentioned two thresholds are comparative, we approximately set $\Theta_s = 0.767$ for the similarity threshold of all combined features. The three scenarios and the corresponding results are described below.

A. Scenario I

At the beginning of the first scenario, person 1 stood at ca. $0°$ of azimuth angle and person 2 at $-25°$. They talked to the robot alternatingly. At time step 205, person 2 got out of the field of view, passed behind the robot. During this time only person 1 spoke. At time step 356, person 2 reappeared at ca. $25°$ and they talked to the robot alternatingly again. Fig. 6 (1) and (2) show the result of scene representation over time using *CAVSA-Pos* and *CAVSA-Multi*, respectively. Person identity is visualized using different colors. Red, blue and black are used for person 1, person 2 and person 3, respectively. A solid line means that the related person is currently in the field of view, while dashed lines indicate disappeared persons. Fig. 7 (for scenario II) and Fig. 8 (for scenario III) visualize results in the same manner. As can be seen in Fig. 6 (1) and (2), when person 2 left the field of view, both *CAVSA-Pos* and *CAVSA-Multi* stored the corresponding visual proto-objects in the STM. However, *CAVSA-Pos* considered person 2 as a new person ("person 3"), when he disappeared briefly and returned from the different azimuth angle, because similarity between the new and saved position was too small. In comparison, *CAVSA-Multi* was able to correctly represent the visual scene, since it integrated also visual features. Fig. 6 (3) visualizes the result of audiovisual integration based on position information. *CAVSA-Multi* was used here for grouping visual proto-objects in STM. "Audio" stands for sound position, while "Current Speaker" represents the visual position of the detected current speaker and "GT-

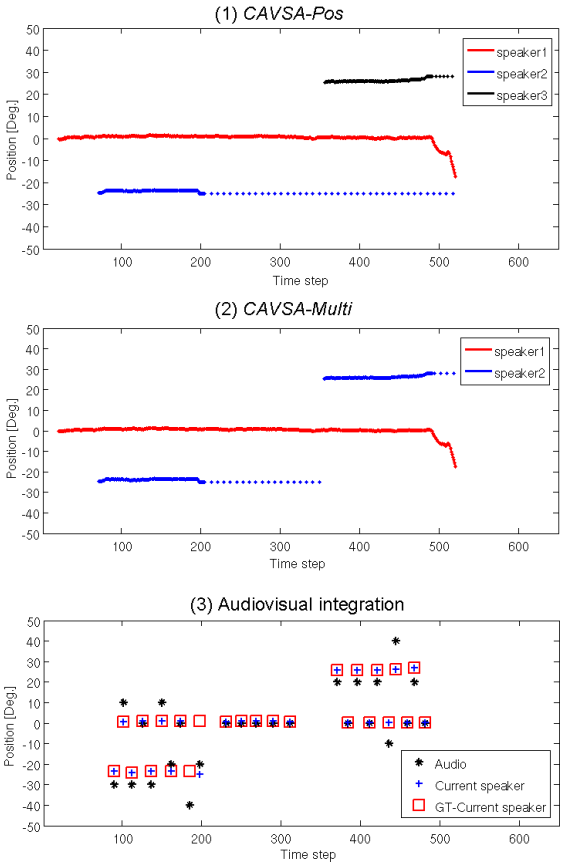

Fig. 6. Scenario I: (1) visual scene representation with *CAVSA-Pos*; (2) visual scene representation with *CAVSA-Multi*; (3) result of audiovisual integration.

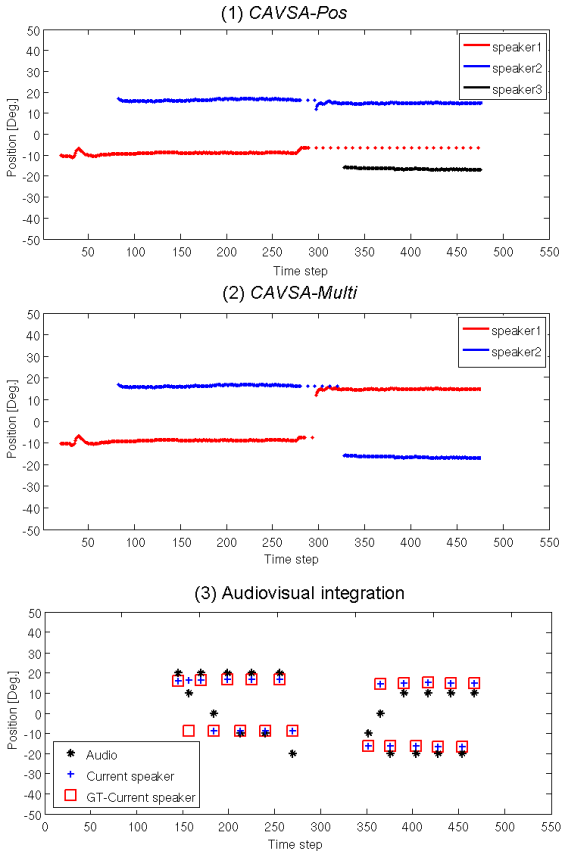

Fig. 7. Scenario II: (1) visual scene representation with *CAVSA-Pos*; (2) visual scene representation with *CAVSA-Multi*; (3) result of audiovisual integration.

Current Speaker" the visual position of the true current speaker. Ground truth data was edited by hand. We can see that audiovisual integration worked most of the time. But the current speaker was considered to have not been seen at time step 185 and was falsely detected at time step 198. Both errors were due to inaccurate sound localization.

B. Scenario II

In the second scenario, person 1 stood initially at ca. $-10°$ of azimuth angle and person 2 at $20°$. They talked to the robot alternatingly. At time step 280, speaker 1 moved from $-10°$ to $20°$ and speaker 2 from $20°$ to $-20°$. Then they continued to talk in the same manner. Fig. 7 shows the result of scene representation using *CAVSA-Pos* and *CAVSA-Multi*, as well as the result of audiovisual integration. *CAVSA-Pos* understood that speaker 2 did not move, speaker 1 exited and a new speaker ("person 3") came into the scenario, while *CAVSA-Multi* correctly represented the scene over time. Fig. 7 (3) shows that only one error of audiovisual integration occurred at time step 157 also due to an erroneous sound localization.

C. Scenario III

In the third scenario, person 1 stood initially at ca. $17°$ of azimuth angle, person 2 at $-15°$ and person 3 was out of the field of view. Person 1 and person 2 talked to the robot alternatingly. At time step 173, person 1 exited at $30°$ (the border of the field of view). Person 3 appeared at $30°$ at time step 187 and moved to $15°$ at time step 195. Then person 2 and person 3 talked to the robot alternatingly. As can be seen from Fig. 8 (1), *CAVSA-Pos* detected speaker 1 as a new person "person 3" at time step 171 because of his fast movement. Then person 1 exited from the border of the field of view and speaker 3 appeared from the same side, so that they were considered as the same person ("person 3"). At time step 190, person 3 moved to the old position of person 1, and was detected as person 1, while "person 3" was detected to be out of the field of view. As shown in Fig. 8 (2), *CAVSA-Multi* represented the dynamic scene correctly. Fig. 8 (3) illustrates that all the audiovisual integrations were successful.

V. Summary and Outlook

We have introduced a system that uses various simple auditory and visual features to achieve human-robot dialog scene analysis in multi-person environments. We have shown that our system is able to correctly determine the number and position of speakers as well as who is currently speaking. In comparison to previous work, our current system, integrating more simple visual features, is able to recognize persons again in more complicated scenarios where speakers leave

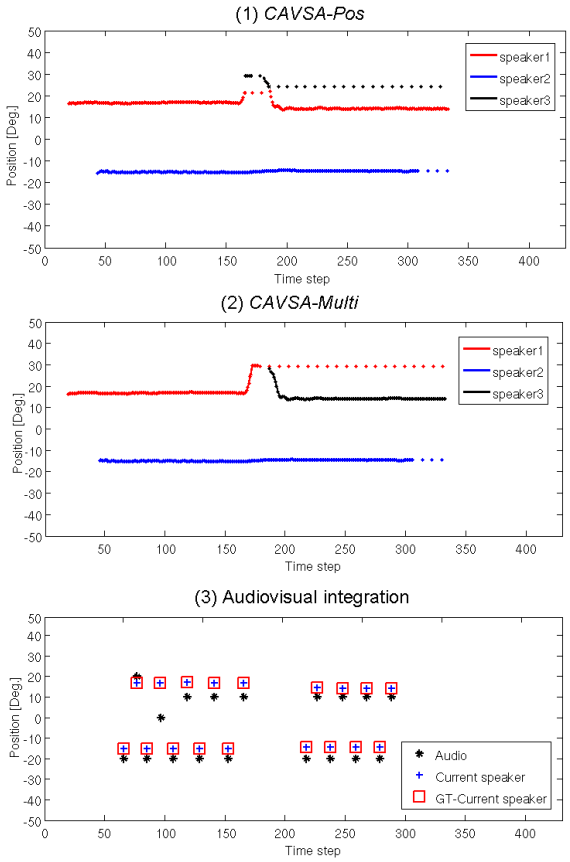

Fig. 8. Scenario III: (1) visual scene representation with *CAVSA-Pos*; (2) visual scene representation with *CAVSA-Multi*; (3) result of audiovisual integration.

and enter the scene, or change their positions even with a newly appearing person.

Presently, the current speaker is selected correctly most of the time. However, audiovisual integration sometimes fails if sound location is falsely estimated. Therefore, we plan to use more auditory features in future work to recognize voices. We expect that once an audio and a visual proto-object are integrated, the integration can be maintained for a certain time even if sound position is sometimes inaccurate. Moreover, we chose visual features based on our training database. However, the performance of individual features for differentiating between people is scenario dependent. For example, if people have very different hair colors, Hair RGB could be the best feature. Thus a future issue is to fuse features adaptive to scenarios.

ACKNOWLEDGMENT

Sincere thanks to all the colleagues at Honda Research Institute Europe, who participated in the experiments.

REFERENCES

[1] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," *Advances in Neural Information Processing Systems*, vol. 12, pp. 813–819, 2000.

[2] H. Hung and G. Friedl, "Towards audio-visual on-line diarization of participants in group meetings," *In Proc.of European Conference on Computer Vision (ECCV)*, 2008.

[3] A. Noulas and B. J. A. Krose, "On-line multi-modal speaker diarization," *In Proc. Int. Conf. on Multimodal interfaces (ICMI)*, 2007.

[4] A. L. Casanovas, G. Monaci, and P. Vandergheynst, "Blind audiovisual source separation using sparse representations," *IEEE Transactions on Multimedia*, pp. 358 – 371, 2010.

[5] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 6, pp. 406–413, 2004.

[6] H. Kim, K. Komatani, T. Ogata, and G. Okuno, "Auditory and visual integration based localization and tracking of humans in daily-life environments," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2007.

[7] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for humanoids," *Proc. of 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

[8] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud, "Audio-visual clustering for multiple speaker localization," *5th International Workshop on Machine Learning for Multimodal Interaction*, 2008.

[9] K. Bernardin and R. Stiefelhagen, "Audio-visual multi-person tracking and identification for smart environments," *Int.Proc. of the 15th Int. Conf. on Multimedia*, 2007.

[10] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke, "Enabling a humanoid robot to interact with multiple persons," *In Proc. Int. Conf. on Dextrous Autonomous Robots and Humanoids (DARH)*, 2005.

[11] B. Bolder, M. Dunn, M. Gienger, H. Janssen, H. Sugiura, and C. Goerick, "Visually guided whole body interaction," *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2007.

[12] T. Rodemann, F. Joublin, and C. Goerick, "Audio proto objects for improved sound localization," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2009.

[13] H. Liu and M. Shen, "Continuous sound source localization based on microphone array for mobile robots," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, 2010.

[14] R. Yan, T. Rodemann, and B. Wrede, "Computational audiovisual scene analysis for dialog scenarios," *IROS 2011 Workshop on Cognitive Neuroscience Robotics, http://www.honda-ri.de/intern/Publications/PUBA-22*, 2011.

[15] R. A. Rensink, "Seeing, sensing and scrutinizing," *Vision Research*, vol. 40, pp. 1469–1487, 2000.

[16] B. B. Averbeck, P. E. Latham, and A. Pouget, "Neural correlations, population coding and computation," *Nature Neuroscience*, vol. 7, pp. 358–366, 2006.

[17] M. Jazayeri and J. A. Movshon, "Optimal representation of sensory information by neural populations," *Nature Neuroscience*, vol. 9, pp. 690 – 696, 2006.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[19] T. Rodemann, F. Joublin, and C. Goerick, "Continuous and robust saccade adaptation in a real-world environment," *KI-Künstliche Intelligenz*, 2006.

[20] P. Corcoran, G. Costache, and R. Mulryan, "Automatic indexing of consumer image collections using person recognition techniques," *IEEE Int. Conf. on Consumer Electronics (ICCE)*, 2005.

[21] G. Jaffre, P. Joly, I. Universit, and P. Sabatier, "Costume: A new feature for automatic video content indexing," *In Proc. Int. Conf. on Adaptivity, Personalization and Fusion of Heteogeneous Information (RIAO)*, 2004.

[22] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer, "Audiovisual person tracking with a mobile robot," *In Proc. Int. Conf. on Intelligent Autonomous Systems*, 2004.

[23] S. Marcel, Y. Rodriguez, and G. H. Marcel, "On the recent use of local binary patterns for face authentication," *International Journal on Image and Video Processing Special Issue on Facial Image Processing*, 2007.

[24] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, 2007.

[25] R. Yan, T. Rodemann, and B. Wrede, "Learning of audiovisual integration," *In Proc. of IEEE Int. Conf. on Development and Learning (ICDL)*, 2011.