Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200 (B.C.))

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST – Priority 2

DELIVERABLE NO: D6.13
Evaluation Results

Date of deliverable: 31.12.2010
Actual submission date: 14.02.2011

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **FRA**

Revision [0]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200 (B.C.))

D6.13 – EVALUATION RESULTS

FRAUNHOFER INSTITUTE FOR DIGITAL MEDIA TECHNOLOGY,
PROJECT GROUP HEARING, SPEECH AND AUDIO TECHNOLOGY
(FRA)

Abstract:

This deliverable presents the results of the evaluation of detector models from different project partners against an audio-visual data evaluation set. The evaluation set consists of recordings taken from the dirac databases assembled at three recording locations (Oldenburg, Zurich and Portland).

DIRAC
Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than things you do expect) Plautus (ca 200 (B.C.))

Table of Content

I Evaluation Results

1 Introduction

The research and development cycle within the DIRAC project, starting with the development of ideas, the description and fostering of the DIRAC paradigm of incongruency and leading to the development of detectors and models as well as the aggregation of several databases will be closed with the evaluation of these models. The databases collected by the project partners have been prepared to serve as a basis for this evaluation by annotating its content.
This deliverables describes the DIRAC detectors and models under evaluation, explains the evaluation procedure and presents the evaluation results.

2 DIRAC detectors and models to be evaluated

For the evaluation process, both audio and visual detectors provided by different project partners are evaluated against the ground truth annotation of the different databases of the DIRAC project. The following subsections provide an overview over each detector.

2.1 Acoustic Object Detection – Speech/Non-Speech Discrimination

This detector aims to classify parts of the one-channel audio signal of a recorded scene as either speech or non-speech a pre-trained model. The detector produces a binary label output every 500milliseconds indicating whether speech was detected or not.
The model used for this detection is based on amplitude modulation features coupled with a support vector machine classifier back-end. Features used for classification are modulation components of the signal extracted by computation of the amplitude modulation spectrogram. By construction, these features are largely invariant to spectral changes in the signal, thereby allowing for a separation of the modulation information from purely spectral information, which in turn is crucial when discriminating modulated sounds such as speech from stationary backgrounds. The SVM back-end allows a very robust classification since it offers good generalization performance. Further information about this detector can be found in [2] and [3].

2.2 Acoustic Localization detector

The acoustic localization detector analyzes a 2-channel audio signal of a recorded scene. It aims to give directional information of every acoustic object it detects within every time frame of 80 milliseconds. The output for every time frame is a vector of yes/no information for all 61 non-overlapping segments between 0 and 180 degrees of arrival with respect to the stereo microphone basis. The detector is capable of detecting multiple acoustic objects within one time frame, but not capable to classify any localized acoustic object.

The uses a correlation-based feature front-end and a discriminative classification back-end to classify the location-dependent presence or absence of acoustic sources in a given time frame. The features are computed on the basis of the generalized cross correlation (GCC) function between two audio input signals. The GCC is an extension of the cross power spectral density function, which is given by the Fourier transform of the cross correlation. Support Vector Machines are employed to classify the presence or absence of a source at each angle. This approach enables the simultaneous localization of more than one sound source in each time-frame. More details can be found in [3].

2.3 Tracker Tree

The tracker tree processes video information on a frame by frame basis by utilizing multiple specialized models organized in a hierarchical way (see Fig. 1). Within this deliverable, the models of Level 3, namely “Sitting”, “Walking” and “Picking up” are evaluated individually models and as part of DIRAC incongruency models.

Each model operates on a frame by frame basis and gives confidence measures for the action it was designed for, i.e.“sitting” for a sitting person, “walking” for a walking or standing person, and “picking” for a person picking up something. More information about the individual model algorithms have been presented e.g. in [11-13].

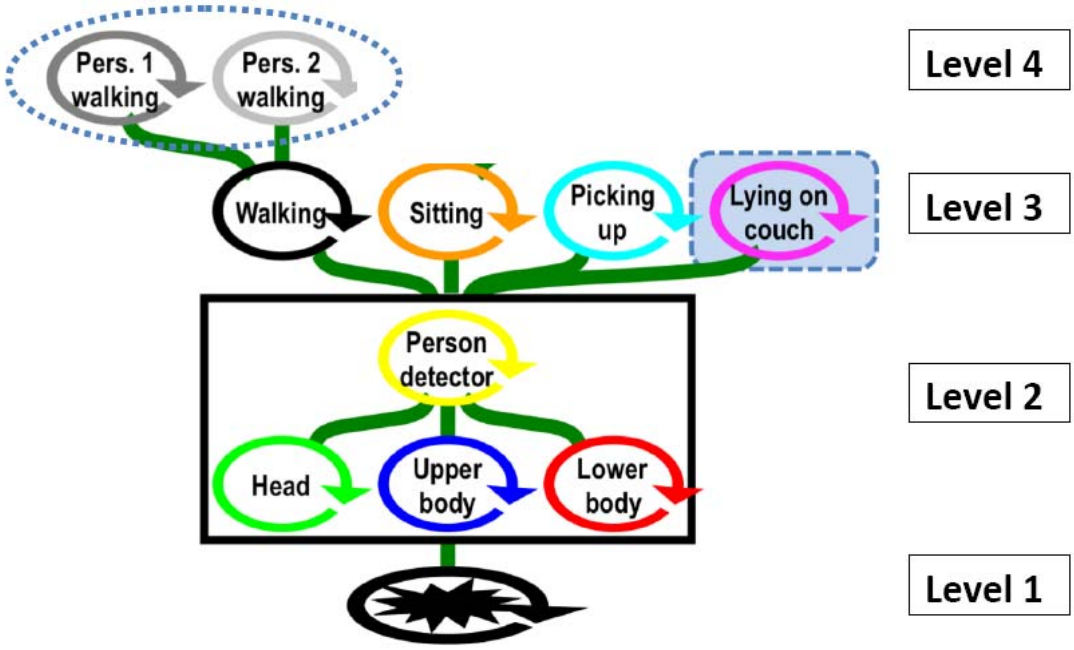

Figure 1: Visualization of the tracker tree and its dependencies over multiple levels.

2.4 Conversation Detector

The multi-modal Conversation Detector uses the DIRAC principle of incongruency detection to discriminate between normal conversation and unusual conversational behaviour, e.g. a person talking to himself. The Detector operates on output data from three detectors presented in this document: the speech/non-speech classificator mentioned (subsection 2.1), the audio localizer, the acoustic localization detector (subsection 2.2) and the tracker tree (subsection 2.3) person detector. The output signals are combined into a DIRAC incongruence model instantiating a part-whole relationship: three models on the general level (PT: person tracker, AL: audio localizer, SC: speech classificator), are combined to one conjoint model; one model on the specific level uses the fused data input of each model on the general level (see Fig. 2). The model on the specific level (Cf) utilizes a linear support vector machine and operates on the same (albeit fused) input data as the models on the general level. An incongruency is detected when the conjoint model accepts the input as conversation, whereas the specific model does not.

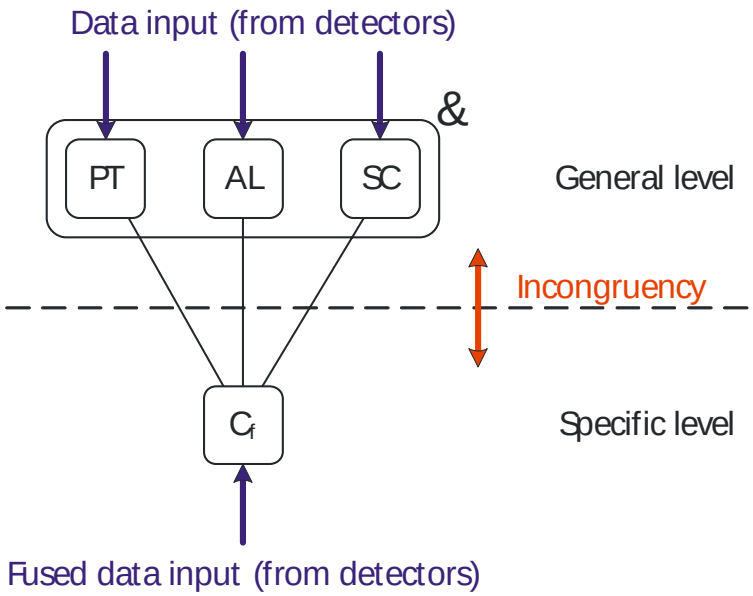

Figure 2: DIRAC incongruence model of the Conversation Detector.

All models generate a binary decision as output on a frame by frame basis:

- the Person Tracker (PT) signals "1" if any person is visible, "0" otherwise.
- the Audio source Localizer (AL) signals "1" if any sound event is detected, "0" otherwise.
- the Speech Classifier (SC) signals "1" if the sound event is classified as speech, "0" otherwise.
- the Conjoint Model (&): uses the output of PT, AL and SC as input. The output of the Conjoint Model is "1" if all three inputs are 1 (logical AND function), "0" otherwise.
- The specific Model (Cf, trained SVM) signals "1" for conversation, "0" otherwise.

A more detailed description of the Conversation Detector is given in [16].

2.5 Combined tracker tree and transfer learning

The partners IDIAP and ETHZ have combined the tracker tree algorithm from ETHZ for incongruent actions detection and the transfer learning algorithm from IDIAP for learning of the new detected action. In this combination, the ETHZ tracker tree detects an incongruent action, and asks for human annotation of few frames (from 1

to maximum 10) of it. These annotated frames are sent to the IDIAP transfer learning algorithm, which learns the new action from the few annotated samples, exploiting the prior knowledge of the system. Note that the original tracker tree algorithm would need an average of 200 annotated samples for learning such action. Once the new class has been learned, the IDIAP method acts as an algorithmic annotators, and labels data sequences sent from the ETHZ tracker tree where incongruent actions are detected. Once the number of annotation is of at least 200 frames, the data are sent to the ETHZ tracker tree, that can build the new action representation and integrate it in the tree. The position where the action is added in the tree depends on where in the hierarchy the incongruency has been detected. A thorough experimental evaluation has been reported in D5.13, and is not part of this deliverable.

3 Evaluation Databases

Starting point for the selection of relevant data to form the evaluation database has been the collection of all recordings done by the different project partners at locations in Oldenburg, Germany (FRA, OL), in Zurich (FRA, ETHZ) and in Portland (OHSU). From this overall collection of recordings, all data has been reduced showing technical drawbacks like noise that can be traced back to systematic errors in the recording sessions or to improper recording setups. In particular, 60 Hz humming noise was sometimes introduced by broken recording equipment, wind- noises and background noise levels which completely masked (even for a human listener) the intended foreground objects prevented a meaningful acoustic analysis of the scenes. Bad light conditions at the recording set or a heavy colour unbalance have been reasons for exclusion from the final evaluation set, too. More details about these issues can be found in [10]. A list of all recordings used for each detector can be found in Appendix A and B.

All recordings from the evaluation database have been pre-processed with the DIRAC pre-processing pipeline and processed by the detector models. For the evaluation ground truth, each relevant scene has been annotated by human inspection.

One important fact should be mentioned explicitly: no recordings from the evaluation dataset have been used to train the detector models. The partners only used their own recordings which have not been part of any of the databases collected within the DIRAC project. Thus, the selected evaluation dataset are completely new to the detector models.

4 Evaluation Procedure

4.1 Acoustic Object Detection, Speech-Non-Speech Discrimination

An overview of the scene-based evaluation procedure for speech-non-speech discrimination can be found in Figure 3. For a given audiovisual scene, the ground truth annotation and the acoustic object detector output is imported from the DIRAC database.

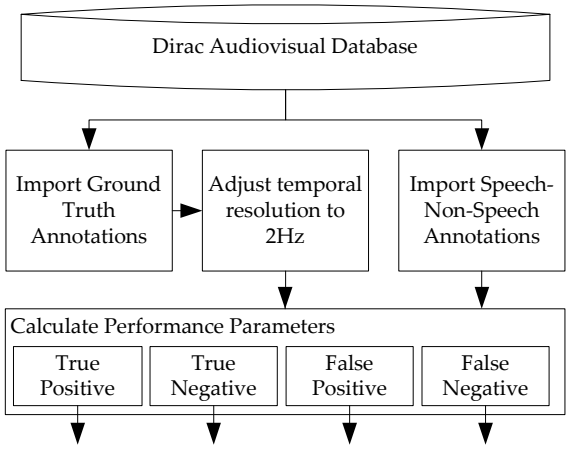

Figure 3: Evaluation Procedure for Speech-Non-Speech Discrimination

As described in [14], the ground truth annotation for speech-non-speech discrimination is represented in the HTK-Format, i.e. in a sampling frequency independent representation of the beginning and the end of a speech segment within an audio file. The original sample-index of each speech containing sample can be regenerated by using Eq. 1.

$$S = \frac{S_{HTK}}{10^7} \cdot F_S \tag{1}$$

Since the ground truth annotation for speech-non-speech discrimination was generated using the original audio data, its temporal resolution needs to be adjusted to match that of the acoustic object detector before the calculation procedure. The acoustic object detection (the speech-non-speech discrimination) generates a label every 500ms, which corresponds to a temporal resolution of 2Hz.

For the evaluation, the output of the acoustic object detection is compared with the ground truth annotation. The results of each comparison adds to the content of the confusion matrix for that detector following this rule:

- Add to TruePositive if detector output matches with positive annotation (speech detected and annotated)
- Add to TrueNegative if detector output matches with negative (no speech detected an no speech annotated)
- Add to FalsePositive if detector detects speech, but annotation says no speech
- Add to FalseNegative if detector does not detect speech, but annotation says speech

The rules are illustrated in Figure 4.

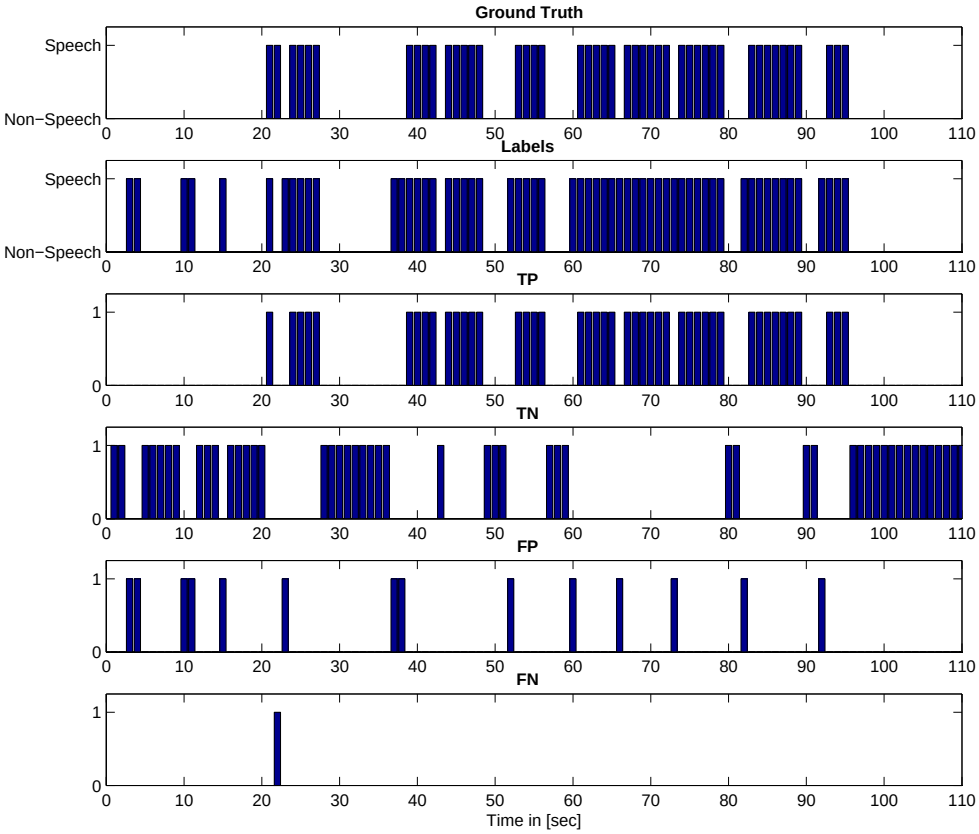

Figure 4: Illustration of the calculation of the performance parameters True Positive, True Negative, False Positive and False Negative

4.2 Acoustic Object Localization – Direction of Arrival

An overview of the scene-based evaluation procedure for acoustic object localization is depicted in found in Figure 5. For a given audiovisual scene out of the DIRAC database, the ground truth annotations and the acoustic object localization output is used. As ground truth annotation, the annotation for a persons head from the video modality, i.e. labels that have been originally generated for the evaluation of the tracker tree is used. The location of a person is given as pixel coordinates based on an un-projected, spherically distorted video frame, which makes a conversion necessary.

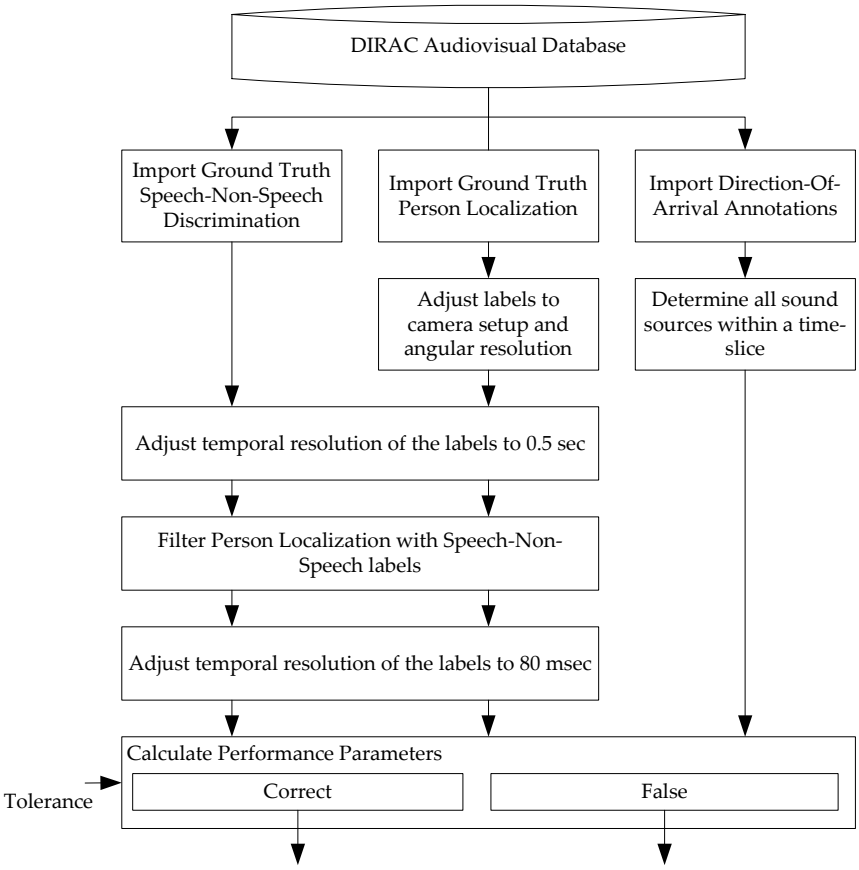

Figure 5: Evaluation Approach for Acoustic Source Localization

In order to use these labels for the evaluation of acoustic object localization, the coordinates first need to be projected in a way that the spherical projection of the image introduced by the cameras matches the angular resolution of the acoustic object localization. Within the DIRAC project, several tools and functions for MATLAB were developed for exactly this purpose.

For evaluation, only a speaking person is of interest, as a visible sound source is needed to use the video annotation as ground truth. So, the head position annotation data is combined with the ground truth speech annotation. The temporal resolution of both ground truth annotations and the filtered data is then adapted to match the audio localization output. As a result, segments that contain a speaking person only are obtained.

The output of the acoustic object localization consists of three-dimensional data, i.e. a matrix that shows us the location of a sound source (speech) over time. Each sound source in a time-slice is represented by a "1" at its angular position in the matrix. This now is the basis for the evaluation of the acoustic object localization.

For the evaluation, both the generated ground truth data and the detector output is compared and the result is added to the confusion matrix. A sound source (in this case speech) is counted as correctly tracked if the angular position of the ground truth annotation equals the one of the detected sound sources. The sound source is also considered as correctly tracked if its position is within a defined range of tolerance. The tolerance parameter is introduced to compensate for the errors introduced by the projection and to take care of the fact that a person can move freely within a room while speaking instead of constantly standing in a frontal position to the camera and microphone array. Thus, the positions of the mouth does not necessarily correspond to the coordinates of the centre of the head of a person. The tolerance parameter is currently set to 25 pixels, which corresponds to the mean width of a head in the audio-visual scenes.

4.3 Tracker Tree

An overview of the scene-based evaluation procedure for the tracker tree can be found in Figure 6.

In general, the evaluation procedure can be distinguished into the evaluation of actions (such as walking, standing, picking something up, sitting, etc.) and positions (such as the position of body parts, persons and moving objects and their presence as well) as it is described in [14].

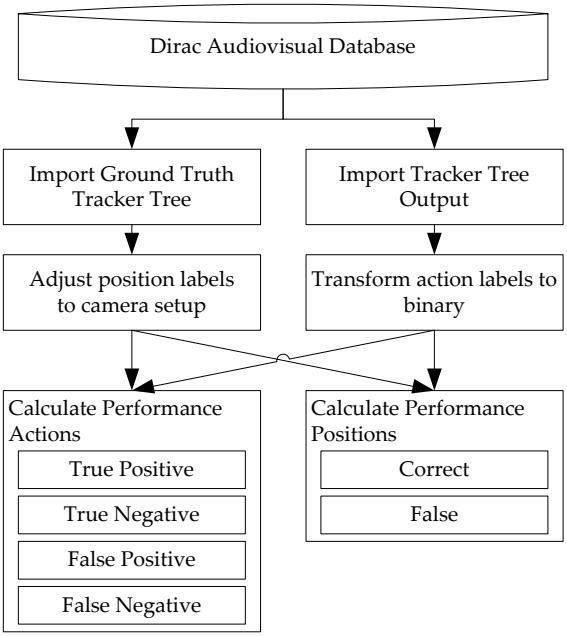

Figure 6: Evaluation Procedure for the Tracker Tree

For a given audiovisual scene out of the DIRAC database, the ground truth annotations and the tracker tree output is used. Both datasets cover both the actions and the positions. Similar to the evaluation procedure of the acoustic object localization described in the previous section, the ground truth annotation for the positions needs to be projected accorded to utilized camera setup. For this purpose, tools and MATLAB functions developed in the DIRAC project are used.

At the same time, the continuous confidence output of the tracker tree for actions needs to be transformed into a binary representation. Otherwise, an evaluation according to [4] would not be possible. Therefore, a threshold parameter T is introduced. Despite that this parameter should be chosen for each audiovisual recording individually to take environmental conditions and a variety of potential audiovisual impairments into account, this would prevent a comparison of the detector performance on a global level. The parameter T is set to -30dB to obtain a binary representation of the detector output following Equation 2.

$$AT[n] = \begin{cases} 1 & AT[n] > T \\ 0 & otherwise \end{cases} \tag{2}$$

Now, the performance reference is the same for all data sets.

Once all this data is available, the actual evaluation procedure is started, which is basically the same as for the speech-non-speech discrimination and the acoustic object localization as described in the two previous sections.

4.4 Conversation Detector

Prior to evaluation, the output signals of the audio localization detector, the speech /non-speech classificator and the video-based person locator are pre-processed. In a first step, time resolution for all input data streams is interpolated. In a second step, the video localization coordinates are mapped to the angle resolution of the audio localization coordinates using a piecewise linear mapping function derived from calibration data recorded in WP6. In a third step, the position of each person and each sound localization is mapped into an interval of bins with fixed width.

After the pre-processing steps, the localization information of both modalities (video and audio) is given as a vector of 257 entries each on a time basis of 12 frames per second. The entries are “1” for object localized and “0” otherwise. The speech/non-speech information is given as a binary information for each frame. The fused data vector has dimension 515 (2x257+1) with a time basis of 12 frames per second. The fused data, i.e. the input data of the specific model, is labelled automatically. Additionally, all frames were labelled w.r.t. the expected incongruency, the monologue of a single person. This second labelling formed the ground truth for the evaluation of the DIRAC model producing the incongruency between the general and specific model.

5 Results and Discussion

In the following subsections, separate results for every evaluated model and for each of the three recording locations is given. The positions of the True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN) in table representing the confusion matrix is as follows:

TP	FP
FN	TN

5.1 Acoustic Object Detection – Speech-Non-Speech Discrimination

For the speech/non-speech discrimination, recordings from locations Oldenburg, Portland and Zurich have been evaluated. From the Oldenburg data set, 74,4% of 1928 frames have been detected correctly.

Oldenburg data set:

1295	383
109	141

From the Portland data set, 76,2% of 15312 frames have been detected correctly.

Portland data set:

7478	2559
1073	4202

From the Zurich data set, 80,8% of 3240 frames have been detected correctly.

Zurich data set:

1327	591
28	1294

For the joined data set, 76,8% of all 20480 frames have been detected correctly.

Joined data sets:

10100	3533
1210	5637

5.2 Acoustic Localization

For the audio localization, recordings from Zurich have been evaluated. Due to the special evaluation scheme, only hit and miss rates are available. For 87,7% of the 3756 frames, a localized position matches the annotated position of a speaking person.

Zurich data set:

hits	misses
3297	459

5.3 Tracker Tree

Detector "Picking up"

For the tracker tree detector "Picking up", recordings from locations Oldenburg and Zurich have been evaluated. From the Oldenburg data set, 77,5% of 1152 frames have been detected correctly.

Oldenburg data set:

41	93
166	852

From the Zurich data set, 87,5% of 4596 frames have been detected correctly.

Zurich data set:

157	367
227	3845

For the joined data set, 84,2% of all 4596 frames have been detected correctly.

Joined data sets:

198	460
393	4697

Detector "Sitting"

For the tracker tree detector "Sitting", recordings from locations Oldenburg, Portland and Zurich have been evaluated. From the Oldenburg data set, 76,3% of 3868 frames have been detected correctly.

Oldenburg data set:

372	244
672	2580

From the Portland data set, 76,4% of 69884 frames have been detected correctly.

Portland data set:

208	7738
8733	53205

From the Zurich data set, 80,4% of 16021frames have been detected correctly.

Zurich data set:

3213	1840
1224	9744

For the joined data set, 77,2% of all 89773 frames have been detected correctly.

Joined data sets:

3793	9822
10629	65529

Detector "Walking"

For the tracker tree detector "Walking", recordings from locations Oldenburg, Portland and Zurich have been evaluated. From the Oldenburg data set, 74,6% of 16785 frames have been detected correctly.

Oldenburg data set:

6355	1251
3007	6172

From the Portland data set, 75,8% of 91343 frames have been detected correctly.

Portland data set:

16860	84
21981	52418

From the Zurich data set, 71,9% of 26185 frames have been detected correctly.

Zurich data set:

6961	594
6742	11888

For the joined data set, 74,9% of all 1343313 frames have been detected correctly.

Joined data sets:

30176	1929
31730	70478

5.4 Conversation Detector

For the Conversation detector, recordings from location Zurich have been evaluated. The Evaluation has been twofold: the trained specific model has been evaluated as such, and the DIRAC model using the general and specific model has been evaluated.

For the specific model, 97,7% of 5484 frames have been detected correctly.

Specific model, Zurich:

985	0
124	4375

The DIRAC model, generated correct incongruencies for 98,4% of all 5484 frames.

DIRAC model, Zurich:

1515	85
0	3884

5.5 Summary of evaluation results

The different detectors evaluated for both audio and video modality show a good performance over the evaluation data set from all three recording locations. The results per location never fall below 71% for a detector per location set, and there are only subtle differences in detector performance between recordings of different locations. Over all sets and all detectors, a remarkable average of 75,3% of all frames have been correctly detected.

6 Conclusion

Within this deliverable, detector models developed by different project partners have been evaluated against the joint DIRAC audio-visual data bases. The evaluation data base has been processed with the detectors, and ground truth has been annotated for evaluation. The detectors under evaluation performed well on recordings from all three recording locations on both the audio and video input data.

7 References

[1] HTK Speech recognition toolkit. Available online, <http://htk.eng.cam.ac.uk/>, 26th August 2010.

[2] J. Bach, B. Kollmeier, and J. Anemüller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010, pp. 41-44.

[3] J. Bach, H. Kayser, and J. Anemüller, "Audio Classification and Localization for Incongruent Event Detection", in Proc. ECML PKDD 2010 Workshop, Detection and Identification of Rare Audio-Visual Cues (DIRAC), Barcelona, Spain, September 2010

[4] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.11 "Testing and evaluation plan", 29. June 2010

[5] Hengel, P.W.J. van, Andringa, T.C.: Verbal aggression detection in complex social environments. In Proceedings of AVSS 2007, (2007)

[6] Hengel, P.W.J. van, and Anemüller, J.: Audio Event Detection for In-Home-Care. In NAG/DAGA International Conference on Acoustics, Rotterdam, 2326 March 2009, (2009)

[7] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.1 "Application Scenarios"

[8] DIRAC Project Wiki Page: <https://dirac.uni-oldenburg.de/DIRAC>

[9] Official DIRAC Project Web page: <http://www.diracproject.org/>

[10] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.12 "Catalogue of basic scenes containing incongruent events"

[11] F. Nater, H. Grabner, T. Jaeggli, and L. van Gool, "Tracker trees for unusual event detection," in Proc. ICCV 2009 Workshop on Visual Surveillance, 2009.

[12] P. Felzenszwalb, D. McAllester, D. Ramaman. A Discriminatively Trained, Multiscale, Deformable Part Model. Proceedings of the IEEE CVPR 2008.

[13] G. Bradski Computer Vision Face Tracking For Use in a Perceptual User Interface Intel Technology Journal, Q2, 1998

[14] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.14, II Annex: "Generation of Ground Truth Annotations"

[15] Hollosi, Danilo; Wabnik, Stefan; Gerlach, Stephan; Kortlang, Steffen: Catalog

of basic scenes for rare/ incongruent event detection, DIRAC Workshop at the European conference on machine learning and principles and practice of knowledge discovery in databases (ECMLPKDD2010), Barcelona, Spanien, September 2010

[16] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.13, II Annex: "The Conversation Detector"

Appendix

A Audio based methods and detectors

1.1.1 Speech-Non-Speech Detection

Audio-visual Recordings	True Positive	True Negative	False Postive	False Negative
HdH office (Oldenburg, 01 Apr 2010)				
1_HDH_office_bird_a	80	12	32	6
1_HDH_office_bird_b	66	12	34	12
1_HDH_office_bird_c	97	5	18	8
1_HDH_office_bird_d	90	9	28	15
1_HDH_office_birdnoise_a	88	6	39	21
1_HDH_office_tel_a	70	26	27	3
1_HDH_office_tel_b	73	5	41	9
1_HDH_office_tel_c	89	8	30	3
1_HDH_office_tel_d	100	12	50	6
1_HDH_office_tel_f	85	14	40	5
1_HDH_office_telnoise_a	104	16	32	14
03b_07Apr2009_KAS_Walking				
1_KAS_Walking_2	183	7	9	7
1_KAS_Walking_3	170	9	3	0
Sum	1295	141	383	109
22_17Mar2010_Zurich_living_lab				
scene_01_woman_telephone_take_c	71	69	22	0
scene_01_woman_telephone_take_d	72	64	27	1
scene_01_woman_telephone_light_take_a	78	45	22	1
scene_01_woman_telephone_light_take_b	81	71	17	1
scene_02_knocking_light_take_c	12	39	17	0
scene_02_knocking_take_a	14	37	12	1
scene_02_knocking_take_b	6	32	14	0

scene_03_enter_room_fab_dan_light_take_a	37	32	27	0
scene_03_enter_room_fab_dan_light_take_b	46	48	18	0
scene_03_enter_room_fab_dan_take_c	42	38	28	0
scene_03_enter_room_fab_dan_take_d	37	43	32	0
scene_03_enter_room_fab_dan_take_e	43	37	34	0
scene_04_dan_radio_active_take_a	62	34	7	1
scene_04_dan_radio_active_take_b	71	40	4	1
scene_04_dan_radio_active_light_take_c	82	53	5	0
scene_04_dan_radio_active_light_take_d	87	45	4	2
scene_05_dan_radio_remote_take_d	40	45	1	0
scene_05_dan_radio_remote_light_take_a	48	41	9	4
scene_05_dan_radio_remote_light_take_b	40	40	4	4
scene_07_fab_standup_talkshimself_light_take_d	45	50	14	1
scene_07_fab_standup_talkshimself_light_take_e	44	48	15	1
scene_07_fab_standup_talkshimself_take_b	48	54	14	0
scene_07_fab_standup_talkshimself_take_c	48	52	17	1
scene_16_fab_oov_couch_take_e	31	33	12	4
scene_16_fab_oov_couch_take_f	29	28	18	3
scene_16_fab_oov_couch_light_take_c	28	29	23	2
scene_20_woman_hits_limping_speech_light_take_f	10	24	32	0
scene_20_woman_hits_limping_speech_take_d	10	31	23	0
scene_20_woman_hits_limping_speech_take_e	13	23	32	0
scene_20_fab_hits_limping_speech_light_take_a	18	20	32	0
scene_20_fab_hits_limping_speech_light_take_b	18	26	30	0
scene_20_fab_hits_limping_speech_take_c	16	23	25	0
Sum	1327	1294	591	28
21_06Mar2010_OHSU_walk_walk_with_oov_walk				
scenario9_take3_20100303	226	109	21	73
scenario9_take4_20100303	246	136	37	37
20_05Mar2010_OHSU_walk_lay_on_couch_walk				
scenario8_take4_20100303	57	177	47	14
scenario8_take3_20100303	40	190	72	10

19_05Mar2010_OHSU_walk_pick_something_up_walk				
scenario7_take3_20100303	54	170	16	10
scenario7_take4_20100303	55	145	18	23
Watching television (Portland, 12 Jul 2010)				
take1	46	165	144	2
take2	98	177	119	9
take3	89	219	92	12
Falling picking up object from floor (Portland, 01 Jul 2010)				
take1	74	76	25	17
take2	78	67	25	11
OHSU_inprogress3				
s39t10	30	20	45	0
s39t3	10	136	39	1
s39t5	33	22	33	5
s39t6	35	22	44	1
s39t7	38	13	46	1
s39t8	54	21	30	3
s39t9	36	50	38	1
s40t10	22	79	49	12
s40t5	2	59	52	3
s40t7	3	64	52	1
s40t9	22	62	82	11
s41t1	101	16	28	25
s41t10	63	38	27	7
s41t2	94	16	11	17
s41t3	93	29	14	19
s41t4	78	38	19	9
s41t5	52	44	27	16
s41t6	56	36	31	13
s41t7	60	31	29	12
s41t8	58	31	31	10
s41t9	64	22	36	12
s42t1	41	49	45	26

s42t10	98	35	20	9
s42t3	60	30	37	17
s42t5	76	54	17	17
s42t6	77	59	35	9
s42t7	95	57	23	7
s42t8	79	53	30	3
s42t9	100	11	13	1
s43t1	314	20	23	32
s43t10	366	14	1	52
s43t2	252	20	23	9
s43t3	227	9	10	25
s43t5	123	27	17	8
s43t6	243	35	3	32
s43t7	246	43	16	34
s43t8	227	18	9	26
s45t1	18	43	58	8
s45t10	41	50	34	7
s45t2	8	47	66	1
s45t3	21	61	52	2
s45t4	22	61	62	3
s45t5	18	49	46	4
s45t6	23	44	42	3
s45t7	29	35	34	13
s45t8	38	28	31	8
s45t9	18	51	43	1
s46t1	182	21	29	12
s46t10	128	43	11	13
s46t2	232	28	18	33
s46t3	412	3	7	46
s46t4	460	11	7	45
s46t5	223	6	4	32
s46t6	353	10	1	66
s46t7	92	28	25	5

s46t8	138	25	11	12
s46t9	129	29	10	13
s48t1	15	107	35	1
s48t10	27	49	34	13
s48t3	17	62	29	5
s48t4	24	52	35	4
s48t5	27	60	35	4
s48t7	27	68	35	2
s48t8	29	53	26	5
s48t9	36	64	38	8
Sum	7478	4202	2559	1073

1.1.2 Acoustic Object Localization

22_17Mar2010_Zurich_living_lab		
	Correct	False
scene_01_woman_telephone_light_take_a	435	39
scene_01_woman_telephone_take_c	394	32
scene_01_woman_telephone_take_d	367	71
scene_02_knocking_light_take_c	39	33
scene_02_knocking_take_a	42	48
scene_02_knocking_take_b	6	30
scene_07_fab_standup_talkshimself_light_take_d	261	15
scene_07_fab_standup_talkshimself_light_take_e	225	45
scene_07_fab_standup_talkshimself_take_b	280	8
scene_07_fab_standup_talkshimself_take_c	291	3
scene_16_fab_oov_couch_light_take_c	165	15
scene_16_fab_oov_couch_take_e	194	16
scene_16_fab_oov_couch_take_f	176	16
scene_20_fab_hits_limping_speech_light_take_a	90	18
scene_20_fab_hits_limping_speech_light_take_b	87	21
scene_20_fab_hits_limping_speech_take_c	84	12

scene_20_woman_hits_limping_speech_light_take_f	49		11	
scene_20_woman_hits_limping_speech_take_d	45		15	
scene_20_woman_hits_limping_speech_take_e	67		11	
	3297		459	

B Video based methods and detectors

1.1.3 Tracker Tree - Head

29_27Jul2010_KAS_Elementary_Configuration				
	True Positive	True Negative	False Postive	False Negative
scene_01_DH_take_a	56	0	14	6
scene_01_DH_take_b	93	19	5	3
scene_02_DH_take_a	98	8	8	10
scene_02_DH_take_b	96	7	9	9
scene_03_DH_take_a	130	21	16	13
scene_03_DH_take_b	135	9	11	15
scene_04_DH_take_a	149	5	24	10
scene_04_DH_take_b	137	3	12	10
scene_05_DH_take_a	107	1	1	9
scene_05_DH_take_b	131	1	8	10
scene_06_DH_take_a	149	6	8	7
scene_06_DH_take_b	149	6	10	10
scene_07_DH_take_a	149	4	12	10
scene_07_DH_take_b	141	0	10	14
scene_08_DH_take_a	135	2	6	12
scene_08_DH_take_b	141	2	9	6
scene_09_DH_take_a	129	2	3	21
scene_09_DH_take_b	124	9	0	17
scene_10_DH_take_a	128	4	8	16
scene_10_DH_take_b	128	3	6	19

scene_11_DH_take_a	143	9	7	4
scene_11_DH_take_b	146	7	8	4
scene_12_DH_take_a	158	5	8	4
scene_12_DH_take_b	158	5	8	4
scene_13_DH_take_a	114	2	9	15
scene_13_DH_take_b	53	0	5	72
scene_14_DH_take_a	116	11	6	7
scene_14_DH_take_b	112	6	7	11
scene_15_DH_take_a	72	40	13	15
scene_15_DH_take_b	66	38	12	24
scene_16_DH_take_a	86	36	11	13
scene_16_DH_take_b	69	42	6	18
scene_17_DH_take_a	182	8	14	16
scene_17_DH_take_b	170	5	15	10
scene_18_DH_take_a	181	4	10	5
scene_18_DH_take_b	172	19	2	17
scene_19_DH_take_a	140	3	5	17
scene_19_DH_take_b	137	3	5	15
scene_20_DH_take_a	137	6	10	7
scene_20_DH_take_b	140	7	6	10
scene_21_DH_take_a	147	3	18	12
scene_21_DH_take_b	137	9	17	7
scene_22_DH_take_a	132	5	11	12
scene_22_DH_take_b	141	4	14	6
scene_23_DH_take_a	99	6	5	30
scene_23_DH_take_b	96	4	4	30
scene_24_DH_take_a	101	3	8	28
scene_24_DH_take_b	107	4	9	28
scene_25_DH_take_a	120	2	0	28
scene_25_DH_take_b	102	12	0	30
scene_26_DH_take_a	110	7	11	22
scene_26_DH_take_b	113	7	1	25
scene_27_SG_take_a	60	3	3	24

scene_27_SG_take_b	72	5	1	18
scene_28_SG_take_a	87	12	12	9
scene_28_SG_take_b	64	13	15	8
scene_29_SG_take_a	120	1	5	24
scene_29_SG_take_b	138	0	0	22
scene_30_SG_take_a	108	2	6	24
scene_30_SG_take_b	121	2	1	26
scene_31_SG_take_a	76	0	15	29
scene_31_SG_take_b	96	0	4	25
scene_32_SG_take_a	71	0	12	31
scene_32_SG_take_b	91	0	23	26
scene_33_SK_take_a	162	0	11	133
scene_33_SK_take_b	156	2	10	132
scene_34_SK_take_a	148	0	14	98
scene_34_SK_take_b	149	0	10	91
scene_35_SK_take_a	189	2	11	138
scene_35_SK_take_b	140	0	9	153
scene_36_SK_take_a	207	0	10	93
scene_36_SK_take_b	132	0	15	153
scene_37_SK_take_a	158	0	11	151
scene_37_SK_take_b	145	0	7	148
scene_38_SK_take_b	158	0	22	130
scene_39_DH_take_a	219	6	11	24
scene_39_DH_take_b	238	9	15	23
scene_40_DH_take_a	256	9	5	20
scene_40_DH_take_b	262	12	2	29
scene_41_DH_take_b	293	1	6	25
scene_42_DH_take_a	279	4	6	21
scene_43_SG_take_a	195	29	24	155
scene_46_SG_take_a	206	0	11	181
scene_46_SG_take_b	223	0	14	173
scene_47_SG_take_a	261	53	78	72
scene_49_SK_take_a	243	0	0	175

Sum	11985	599	844	3357
22_17Mar2010_Zurich_living_lab				
scene_01_woman_telephone_take_c	888	0	0	32
scene_01_woman_telephone_take_d	921	0	0	9
scene_03_enter_room_fab_dan_light_take_b	632	0	0	12
scene_03_enter_room_fab_dan_take_c	609	0	0	21
scene_03_enter_room_fab_dan_take_e	649	0	0	11
scene_07_fab_standup_talkshimself_light_take_d	604	0	1	15
scene_07_fab_standup_talkshimself_light_take_e	598	0	0	25
scene_07_fab_standup_talkshimself_take_b	635	0	0	5
scene_07_fab_standup_talkshimself_take_c	666	0	0	4
scene_02_knocking_light_take_c	348	39	6	27
scene_02_knocking_take_a	352	26	4	14
scene_02_knocking_take_b	203	12	96	13
scene_03_enter_room_fab_dan_light_take_a	421	28	1	138
scene_03_enter_room_fab_dan_take_d	636	26	3	19
scene_04_dan_radio_active_light_take_c	765	60	6	21
scene_04_dan_radio_active_light_take_d	767	58	2	13
scene_04_dan_radio_active_take_a	527	48	3	58
scene_04_dan_radio_active_take_b	490	34	140	44
scene_05_dan_radio_remote_light_take_a	552	51	9	12
scene_05_dan_radio_remote_light_take_b	460	58	2	20
scene_05_dan_radio_remote_take_d	437	48	6	37
scene_10_picksup_dan_light_take_c	126	40	5	21
scene_10_picksup_dan_light_take_d	223	32	4	41
scene_10_picksup_dan_take_a	120	37	5	18
scene_10_picksup_dan_take_b	118	41	4	29
scene_10_picksup_dan_take_e	217	49	2	44
scene_11_fab_picking_up_light_take_a	182	36	15	19
scene_11_fab_picking_up_light_take_b	158	28	11	19
scene_11_fab_picking_up_take_c	170	37	8	25
scene_11_fab_picking_up_take_d	181	44	4	35

scene_13_fab_tie_laces_light_take_a	120	40	8	180
scene_13_fab_tie_laces_light_take_b	228	35	4	165
scene_13_fab_tie_laces_take_c	235	52	8	197
scene_13_fab_tie_laces_take_d	198	55	5	210
scene_14_ste_stumbling_light_take_c	156	47	7	6
scene_14_ste_stumbling_light_take_d	170	47	10	13
scene_14_ste_stumbling_take_a	163	42	9	14
scene_14_ste_stumbling_take_b	72	49	2	81
scene_15_ste_falling_complete_light_take_a	193	58	11	50
scene_15_ste_falling_complete_light_take_b	123	52	2	147
scene_15_ste_falling_complete_take_c	140	57	3	148
scene_16_fab_oov_couch_light_take_c	171	18	5	310
scene_16_fab_oov_couch_take_e	379	25	7	81
scene_16_fab_oov_couch_take_f	399	27	8	46
scene_17_ste_falling_down_light_take_c	63	100	29	72
scene_17_ste_falling_down_light_take_d	75	127	8	126
scene_17_ste_falling_down_take_a	67	87	12	62
scene_17_ste_falling_down_take_b	67	144	6	119
scene_18_fab_picking_up_falling_down_limping_light_take_c	243	72	48	93
scene_18_fab_picking_up_falling_down_limping_take_a	277	74	16	41
scene_18_fab_picking_up_falling_down_limping_take_b	242	102	9	67
scene_20_fab_hits_limping_speech_light_take_a	332	59	4	37
scene_20_fab_hits_limping_speech_light_take_b	337	63	9	47
scene_20_fab_hits_limping_speech_take_c	296	62	13	25
scene_20_woman_hits_limping_speech_light_take_f	307	40	14	47
scene_20_woman_hits_limping_speech_take_d	232	68	7	89
scene_20_woman_hits_limping_speech_take_e	267	64	5	84
scene_21_fab_walk_behind_couch_light_take_c	216	37	11	0
scene_21_fab_walk_behind_couch_light_take_d	183	32	7	30
Sum	19606	2567	624	3388

OHSU_inprogress_3				
s39t10/	892	350	21	178
s39t3	479	323	164	1841
s39t5	750	331	302	41
s39t6	1163	307	12	77
s39t7	1064	307	17	106
s39t8	1149	326	31	131
s39t9	1513	279	36	67
s40t2	1611	118	15	30
s40t3	591	184	0	1200
s40t5	0	271	0	1490
s40t6	292	195	12	1108
s40t8	643	227	25	1147
s40t9	1458	415	13	792
s41t1	1866	432	8	274
s41t10	1712	230	4	108
s41t3	1706	462	1	184
s41t4	1500	245	3	430
s41t5	1169	212	0	721
s41t6	1747	233	2	83
s41t7	1727	209	1	73
s41t8	1706	182	1	84
s41t9	1698	239	6	92
s42t1	1798	536	7	102
s42t10	2134	181	37	106
s42t3	1671	429	10	79
s42t5	2141	246	33	59
s42t6	2263	327	20	107
s42t7	2224	376	22	126
s42t8	2152	285	38	18
s42t9	1548	281	24	52
s43t1	5184	472	6	196
s43t2	3786	478	2	324

s44t10	322	137	2	1269
s44t4	1866	166	9	55
s44t6	239	140	1	1462
s44t7	287	217	0	1644
s44t8	221	91	4	1360
s44t9	173	113	3	1848
s39t10	892	350	21	178
s39t3	479	323	164	1841
s39t5	750	331	302	41
s39t6	1163	307	12	77
s39t7	1064	307	17	106
s39t8	1149	326	31	131
s39t9	1513	279	36	67
s40t2	1611	118	15	30
s40t3	591	184	0	1200
s40t5	0	271	0	1490
s40t6	292	195	12	1108
s40t8	643	227	25	1147
s40t9	1458	415	13	792
s41t1	1866	432	8	274
s41t10	1712	230	4	108
s41t3	1706	462	1	184
s41t4	1500	245	3	430
s41t5	1169	212	0	721
s41t6	1747	233	2	83
s41t7	1727	209	1	73
s41t8	1706	182	1	84
s41t9	1698	239	6	92
s42t1	1798	536	7	102
s42t10	2134	181	37	106
s42t3	1671	429	10	79
s42t5	2141	246	33	59
s42t6	2263	327	20	107

s42t7	2224	376	22	126
s42t8	2152	285	38	18
s42t9	1548	281	24	52
s43t1	5184	472	6	196
s43t2	3786	478	2	324
s44t10	322	137	2	1269
s44t4	1866	166	9	55
s44t6	239	140	1	1462
s44t7	287	217	0	1644
s44t8	221	91	4	1360
s44t9	173	113	3	1848
	54445	10552	892	19064

1.1.4 Tracker Tree – Lower Body

29_27Jul2010_KAS_Elementary_Configuration				
	True Positive	True Negative	False Postive	False Negative
scene_01_DH_take_a	60	0	14	2
scene_01_DH_take_b	92	11	13	4
scene_02_DH_take_a	103	8	8	5
scene_02_DH_take_b	90	12	4	15
scene_03_DH_take_a	132	13	24	11
scene_03_DH_take_b	143	6	14	7
scene_04_DH_take_a	154	11	18	5
scene_04_DH_take_b	146	4	11	1
scene_05_DH_take_a	102	1	1	14
scene_05_DH_take_b	125	1	8	16
scene_06_DH_take_a	135	12	2	21
scene_06_DH_take_b	144	11	5	15
scene_07_DH_take_a	141	4	12	18
scene_07_DH_take_b	133	2	8	22

scene_08_DH_take_a	126	8	0	21
scene_08_DH_take_b	130	10	1	17
scene_09_DH_take_a	136	3	2	14
scene_09_DH_take_b	125	6	3	16
scene_10_DH_take_a	125	5	7	19
scene_10_DH_take_b	138	1	8	9
scene_11_DH_take_a	141	7	9	6
scene_11_DH_take_b	141	6	9	9
scene_12_DH_take_a	154	6	7	8
scene_12_DH_take_b	153	7	6	9
scene_13_DH_take_a	103	4	13	20
scene_13_DH_take_b	89	7	7	27
scene_14_DH_take_a	96	15	11	18
scene_14_DH_take_b	98	15	7	16
scene_15_DH_take_a	74	41	12	13
scene_15_DH_take_b	74	37	13	16
scene_16_DH_take_a	87	42	5	12
scene_16_DH_take_b	74	41	7	13
scene_17_DH_take_a	181	17	5	17
scene_17_DH_take_b	154	12	8	26
scene_18_DH_take_a	176	14	0	10
scene_18_DH_take_b	169	19	2	20
scene_19_DH_take_a	138	3	5	19
scene_19_DH_take_b	132	2	6	20
scene_20_DH_take_a	130	13	3	14
scene_20_DH_take_b	141	11	2	9
scene_21_DH_take_a	145	8	13	14
scene_21_DH_take_b	136	11	15	8
scene_22_DH_take_a	132	13	3	12
scene_22_DH_take_b	137	17	1	10
scene_23_DH_take_a	101	5	6	28
scene_23_DH_take_b	99	4	4	27
scene_24_DH_take_a	101	11	0	28

scene_24_DH_take_b	112	12	1	23
scene_25_DH_take_a	124	2	0	24
scene_25_DH_take_b	111	5	7	21
scene_26_DH_take_a	110	9	9	22
scene_26_DH_take_b	118	6	2	20
scene_27_SG_take_a	50	2	4	34
scene_27_SG_take_b	65	3	3	25
scene_28_SG_take_a	64	11	1	44
scene_28_SG_take_b	50	10	0	40
scene_29_SG_take_a	105	6	0	39
scene_29_SG_take_b	123	0	0	37
scene_30_SG_take_a	91	7	1	41
scene_30_SG_take_b	116	1	2	31
scene_31_SG_take_a	66	0	15	39
scene_31_SG_take_b	91	0	4	30
scene_32_SG_take_a	58	8	4	44
scene_32_SG_take_b	92	14	9	25
scene_33_SK_take_a	144	0	11	151
scene_33_SK_take_b	135	2	10	153
scene_34_SK_take_a	142	5	9	104
scene_34_SK_take_b	140	3	7	100
scene_35_SK_take_a	158	2	11	169
scene_35_SK_take_b	118	0	9	175
scene_36_SK_take_a	194	5	5	106
scene_36_SK_take_b	131	4	11	154
scene_37_SK_take_a	144	0	11	165
scene_37_SK_take_b	123	0	7	170
scene_38_SK_take_b	144	4	18	144
scene_39_DH_take_a	231	9	8	12
scene_39_DH_take_b	248	11	13	13
scene_40_DH_take_a	268	10	1	11
scene_40_DH_take_b	274	13	1	17
scene_41_DH_take_b	165	110	8	42

scene_42_DH_take_a	158	127	3	22
scene_43_SG_take_a	181	0	8	214
scene_46_SG_take_a	114	113	0	171
scene_46_SG_take_b	130	119	0	161
scene_47_SG_take_a	223	0	8	233
scene_49_SK_take_a	124	102	0	192
	11071	1252	563	3899
22_17Mar2010_Zurich_living_lab				
scene_01_woman_telephone_take_c	794	0	0	126
scene_01_woman_telephone_take_d	852	0	0	78
scene_03_enter_room_fab_dan_light_take_b	580	0	0	64
scene_03_enter_room_fab_dan_take_c	586	0	0	44
scene_03_enter_room_fab_dan_take_e	641	0	0	19
scene_07_fab_standup_talkshimself_light_take_d	586	0	1	33
scene_07_fab_standup_talkshimself_light_take_e	575	0	0	48
scene_07_fab_standup_talkshimself_take_b	631	0	0	9
scene_07_fab_standup_talkshimself_take_c	656	0	0	14
scene_02_knocking_light_take_c	131	43	2	244
scene_02_knocking_take_a	188	30	0	178
scene_02_knocking_take_b	66	58	50	150
scene_03_enter_room_fab_dan_light_take_a	433	28	1	126
scene_03_enter_room_fab_dan_take_d	438	25	4	217
scene_04_dan_radio_active_light_take_c	445	216	3	188
scene_04_dan_radio_active_light_take_d	497	170	7	166
scene_04_dan_radio_active_take_a	343	140	16	137
scene_04_dan_radio_active_take_b	299	100	161	148
scene_05_dan_radio_remote_light_take_a	510	59	1	54
scene_05_dan_radio_remote_light_take_b	419	59	1	61
scene_05_dan_radio_remote_take_d	404	53	1	70
scene_10_picksup_dan_light_take_c	84	45	0	63
scene_10_picksup_dan_light_take_d	148	34	2	116
scene_10_picksup_dan_take_a	70	41	1	68

scene_10_picksup_dan_take_b	52	45	0	95
scene_10_picksup_dan_take_e	179	51	0	82
scene_11_fab_picking_up_light_take_a	81	45	6	120
scene_11_fab_picking_up_light_take_b	88	34	5	89
scene_11_fab_picking_up_take_c	70	43	2	125
scene_11_fab_picking_up_take_d	110	48	0	106
scene_13_fab_tie_laces_light_take_a	92	45	3	208
scene_13_fab_tie_laces_light_take_b	144	37	2	249
scene_13_fab_tie_laces_take_c	165	56	4	267
scene_13_fab_tie_laces_take_d	151	73	23	221
scene_14_ste_stumbling_light_take_c	47	54	0	115
scene_14_ste_stumbling_light_take_d	51	55	2	132
scene_14_ste_stumbling_take_a	63	51	0	114
scene_14_ste_stumbling_take_b	36	51	0	117
scene_15_ste_falling_complete_light_take_a	64	81	6	161
scene_15_ste_falling_complete_light_take_b	33	65	1	225
scene_15_ste_falling_complete_take_c	76	63	0	209
scene_16_fab_oov_couch_light_take_c	77	23	0	404
scene_16_fab_oov_couch_take_e	118	31	1	342
scene_16_fab_oov_couch_take_f	99	33	2	346
scene_17_ste_falling_down_light_take_c	30	54	0	180
scene_17_ste_falling_down_light_take_d	10	63	0	263
scene_17_ste_falling_down_take_a	14	48	0	166
scene_17_ste_falling_down_take_b	32	54	0	250
scene_18_fab_picking_up_falling_down_limping_light_take_c	86	48	6	316
scene_18_fab_picking_up_falling_down_limping_take_a	165	38	4	201
scene_18_fab_picking_up_falling_down_limping_take_b	173	48	0	199
scene_20_fab_hits_limping_speech_light_take_a	169	61	2	200
scene_20_fab_hits_limping_speech_light_take_b	186	69	3	198
scene_20_fab_hits_limping_speech_take_c	176	67	8	145
scene_20_woman_hits_limping_speech_light_take_f	101	53	1	253

scene_20_woman_hits_limping_speech_take_d	78	84	3	231
scene_20_woman_hits_limping_speech_take_e	116	69	0	235
scene_21_fab_walk_behind_couch_light_take_c	131	44	4	85
scene_21_fab_walk_behind_couch_light_take_d	109	37	2	104
Sum	13748	2922	341	9174

1.1.5 Tracker Tree – Picking up

29_27Jul2010_KAS_Elementary_Configuration				
	True Positive	True Negative	False Postive	False Negative
scene_23_DH_take_a	9	107	6	18
scene_23_DH_take_b	8	105	2	19
scene_24_DH_take_a	0	112	1	27
scene_24_DH_take_b	0	106	15	27
scene_25_DH_take_a	15	110	13	12
scene_25_DH_take_b	9	104	16	15
scene_26_DH_take_a	0	103	23	24
scene_26_DH_take_b	0	105	17	24
	41	852	93	166
22_17Mar2010_Zurich_living_lab				
scene_10_picksup_dan_light_take_c	13	136	29	14
scene_10_picksup_dan_light_take_d	15	226	41	18
scene_10_picksup_dan_take_a	12	126	24	18
scene_10_picksup_dan_take_b	22	149	13	8
scene_10_picksup_dan_take_e	14	239	40	19
scene_11_fab_picking_up_light_take_a	13	208	20	11
scene_11_fab_picking_up_light_take_b	26	156	30	4
scene_11_fab_picking_up_take_c	6	209	7	18
scene_11_fab_picking_up_take_d	6	214	23	21
scene_17_ste_falling_down_light_take_c	1	239	10	14

scene_17_ste_falling_down_light_take_d	3	321	0	12
scene_17_ste_falling_down_take_a	0	213	0	15
scene_17_ste_falling_down_take_b	0	311	13	12
scene_18_fab_picking_up_falling_down_limping_light_take_c	10	425	4	17
scene_18_fab_picking_up_falling_down_limping_take_a	2	329	58	19
scene_18_fab_picking_up_falling_down_limping_take_b	14	344	55	7
	157	3845	367	227

1.1.6 Tracker Tree – Sitting

29_27Jul2010_KAS_Elementary_Configuration				
	True Positive	True Negative	False Postive	False Negative
scene_39_DH_take_a	59	163	25	13
scene_39_DH_take_b	73	159	36	17
scene_40_DH_take_a	83	156	38	13
scene_40_DH_take_b	80	157	55	13
scene_41_DH_take_b	0	176	50	99
scene_42_DH_take_a	0	181	27	102
scene_43_SG_take_a	23	329	2	49
scene_46_SG_take_a	0	292	1	105
scene_46_SG_take_b	0	308	3	99
scene_47_SG_take_a	54	340	7	63
scene_49_SK_take_a	0	319	0	99
	372	2580	244	672
22_17Mar2010_Zurich_living_lab				
scene_01_woman_telephone_take_c	77	821	9	13
scene_01_woman_telephone_take_d	140	766	17	7
scene_03_enter_room_fab_dan_light_take_b	201	419	2	22

scene_03_enter_room_fab_dan_take_c	180	426	0	24
scene_03_enter_room_fab_dan_take_e	215	431	4	10
scene_07_fab_standup_talkshimself_light_take_d	2	277	238	103
scene_07_fab_standup_talkshimself_light_take_e	20	488	24	91
scene_07_fab_standup_talkshimself_take_b	49	230	353	8
scene_07_fab_standup_talkshimself_take_c	44	345	223	58
scene_02_knocking_light_take_c	0	306	36	78
scene_02_knocking_take_a	50	291	30	25
scene_02_knocking_take_b	0	224	10	90
scene_03_enter_room_fab_dan_light_take_a	13	456	20	99
scene_03_enter_room_fab_dan_take_d	62	213	245	164
scene_04_dan_radio_active_light_take_c	282	466	98	6
scene_04_dan_radio_active_light_take_d	311	363	162	4
scene_04_dan_radio_active_take_a	176	410	40	10
scene_04_dan_radio_active_take_b	105	299	184	120
scene_05_dan_radio_remote_light_take_a	428	156	9	31
scene_05_dan_radio_remote_light_take_b	357	159	18	6
scene_05_dan_radio_remote_take_d	348	156	15	9
scene_10_picksup_dan_light_take_d	74	189	30	7
scene_10_picksup_dan_take_e	67	208	26	11
scene_16_fab_oov_couch_light_take_c	5	416	4	79
scene_16_fab_oov_couch_take_e	4	418	23	47
scene_16_fab_oov_couch_take_f	3	396	15	66
scene_18_fab_picking_up_falling_down_limping_light_take_c	0	415	5	36
	3213	9744	1840	1224

1.1.7 Tracker Tree – Walking/Standing

29_27Jul2010_KAS_Elementary_Configuration				
	True Positive	True Negative	False Postive	False Negative

scene_01_DH_take_a	59	4	10	3
scene_01_DH_take_b	61	24	0	35
scene_02_DH_take_a	83	16	0	25
scene_02_DH_take_b	78	16	0	27
scene_03_DH_take_a	122	24	13	21
scene_03_DH_take_b	125	16	4	25
scene_04_DH_take_a	113	29	0	46
scene_04_DH_take_b	94	15	0	53
scene_05_DH_take_a	77	2	0	39
scene_05_DH_take_b	101	9	0	40
scene_06_DH_take_a	108	14	0	48
scene_06_DH_take_b	108	16	0	51
scene_07_DH_take_a	129	12	4	30
scene_07_DH_take_b	102	12	1	50
scene_08_DH_take_a	101	8	0	46
scene_08_DH_take_b	111	11	0	36
scene_09_DH_take_a	87	9	23	36
scene_09_DH_take_b	103	9	0	38
scene_10_DH_take_a	102	12	0	42
scene_10_DH_take_b	117	9	0	30
scene_11_DH_take_a	114	16	0	33
scene_11_DH_take_b	125	15	0	25
scene_12_DH_take_a	136	13	0	26
scene_12_DH_take_b	130	13	0	32
scene_13_DH_take_a	69	11	0	60
scene_13_DH_take_b	0	5	0	125
scene_14_DH_take_a	72	26	0	42
scene_14_DH_take_b	75	13	0	48
scene_15_DH_take_a	22	51	2	65
scene_15_DH_take_b	31	48	2	59
scene_16_DH_take_a	41	47	0	58
scene_16_DH_take_b	42	48	0	45
scene_19_DH_take_a	32	47	64	22

scene_19_DH_take_b	37	59	44	20
scene_20_DH_take_a	43	43	57	17
scene_20_DH_take_b	45	41	53	24
scene_21_DH_take_a	41	61	56	22
scene_21_DH_take_b	44	66	41	19
scene_22_DH_take_a	43	51	58	8
scene_22_DH_take_b	42	59	52	12
scene_23_DH_take_a	63	34	4	39
scene_23_DH_take_b	57	31	4	42
scene_24_DH_take_a	71	35	3	31
scene_24_DH_take_b	73	36	4	35
scene_25_DH_take_a	81	25	4	40
scene_25_DH_take_b	73	32	4	35
scene_26_DH_take_a	76	38	4	32
scene_26_DH_take_b	80	27	5	34
scene_27_SG_take_a	39	26	1	24
scene_27_SG_take_b	48	29	1	18
scene_28_SG_take_a	50	33	0	37
scene_28_SG_take_b	36	28	0	36
scene_29_SG_take_a	82	27	0	41
scene_29_SG_take_b	92	18	0	50
scene_30_SG_take_a	84	28	1	27
scene_30_SG_take_b	108	24	0	18
scene_31_SG_take_a	8	60	30	22
scene_31_SG_take_b	11	67	27	20
scene_32_SG_take_a	6	79	14	15
scene_32_SG_take_b	17	92	27	4
scene_33_SK_take_a	116	154	16	20
scene_33_SK_take_b	107	149	4	40
scene_34_SK_take_a	96	134	0	30
scene_34_SK_take_b	96	127	0	27
scene_35_SK_take_a	132	169	0	39
scene_35_SK_take_b	66	169	2	65

scene_36_SK_take_a	148	135	7	20
scene_36_SK_take_b	79	155	7	59
scene_37_SK_take_a	42	215	42	21
scene_37_SK_take_b	34	193	45	28
scene_38_SK_take_b	44	213	43	10
scene_39_DH_take_a	118	103	13	26
scene_39_DH_take_b	130	122	10	23
scene_40_DH_take_a	145	118	7	20
scene_40_DH_take_b	154	118	7	26
scene_41_DH_take_b	78	124	0	123
scene_42_DH_take_a	73	127	0	110
scene_43_SG_take_a	115	261	5	22
scene_46_SG_take_a	57	263	0	78
scene_46_SG_take_b	83	266	0	61
scene_47_SG_take_a	70	329	42	23
scene_49_SK_take_a	52	292	21	53
	6355	5705	888	3007
22_17Mar2010_Zurich_living_lab				
scene_01_woman_telephone_take_c	145	97	5	673
scene_01_woman_telephone_take_d	695	140	13	82
scene_03_enter_room_fab_dan_light_take_b	365	250	0	29
scene_03_enter_room_fab_dan_take_c	370	235	5	20
scene_03_enter_room_fab_dan_take_e	394	254	4	8
scene_07_fab_standup_talkshimself_light_take_d	456	117	4	43
scene_07_fab_standup_talkshimself_light_take_e	352	146	7	118
scene_07_fab_standup_talkshimself_take_b	556	66	0	18
scene_07_fab_standup_talkshimself_take_c	538	111	3	18
scene_02_knocking_light_take_c	99	120	3	198
scene_02_knocking_take_a	82	124	2	188
scene_02_knocking_take_b	30	203	22	69
scene_03_enter_room_fab_dan_light_take_a	181	163	2	242
scene_03_enter_room_fab_dan_take_d	317	273	3	91

scene_04_dan_radio_active_light_take_c	155	359	1	337
scene_04_dan_radio_active_light_take_d	142	374	1	323
scene_04_dan_radio_active_take_a	106	249	0	281
scene_04_dan_radio_active_take_b	53	321	81	253
scene_05_dan_radio_remote_light_take_a	44	519	3	58
scene_05_dan_radio_remote_light_take_b	52	432	0	56
scene_05_dan_radio_remote_take_d	54	419	1	54
scene_10_picksup_dan_light_take_c	77	69	3	43
scene_10_picksup_dan_light_take_d	75	153	9	63
scene_10_picksup_dan_take_a	38	69	3	70
scene_10_picksup_dan_take_b	53	70	5	64
scene_10_picksup_dan_take_e	74	170	7	61
scene_11_fab_picking_up_light_take_a	67	72	3	110
scene_11_fab_picking_up_light_take_b	49	65	4	98
scene_11_fab_picking_up_take_c	13	69	0	158
scene_11_fab_picking_up_take_d	57	75	0	132
scene_13_fab_tie_laces_light_take_a	71	224	1	52
scene_13_fab_tie_laces_light_take_b	57	204	0	171
scene_13_fab_tie_laces_take_c	93	245	1	153
scene_13_fab_tie_laces_take_d	101	228	3	136
scene_14_ste_stumbling_light_take_c	65	74	4	73
scene_14_ste_stumbling_light_take_d	78	73	2	87
scene_14_ste_stumbling_take_a	57	78	0	93
scene_14_ste_stumbling_take_b	23	72	0	109
scene_15_ste_falling_complete_light_take_a	57	177	3	75
scene_15_ste_falling_complete_light_take_b	32	183	0	109
scene_15_ste_falling_complete_take_c	35	201	0	112
scene_16_fab_oov_couch_light_take_c	21	339	0	144
scene_16_fab_oov_couch_take_e	67	372	0	53
scene_16_fab_oov_couch_take_f	63	357	0	60
scene_17_ste_falling_down_light_take_c	1	222	0	41
scene_17_ste_falling_down_light_take_d	0	282	0	54
scene_17_ste_falling_down_take_a	0	198	0	30

scene_17_ste_falling_down_take_b	0	283	2	51
scene_18_fab_picking_up_falling_down_limping_light_take_c	20	300	15	121
scene_18_fab_picking_up_falling_down_limping_take_a	73	246	21	68
scene_18_fab_picking_up_falling_down_limping_take_b	80	253	23	64
scene_20_fab_hits_limping_speech_light_take_a	14	220	68	130
scene_20_fab_hits_limping_speech_light_take_b	40	234	75	107
scene_20_fab_hits_limping_speech_take_c	0	209	79	108
scene_20_woman_hits_limping_speech_light_take_f	34	242	43	89
scene_20_woman_hits_limping_speech_take_d	10	245	22	119
scene_20_woman_hits_limping_speech_take_e	17	261	33	109
scene_21_fab_walk_behind_couch_light_take_c	102	46	2	114
scene_21_fab_walk_behind_couch_light_take_d	61	36	3	152
	6961	11888	594	6742
OHSU_inprogress3				
s39t10	252	592	19	578
s39t3	173	2467	0	167
s39t5	0	893	0	531
s39t6	385	579	0	595
s39t7	359	394	0	741
s39t8	289	523	4	821
s39t9	286	1011	4	594
s40t2	141	1544	0	89
s40t3	128	1725	0	122
s40t5	0	1531	0	230
s40t6	0	1357	0	250
s40t8	391	1492	0	159
s40t9	1010	1263	5	400
s41t1	93	1810	0	677
s41t10/	140	1464	0	450
s41t3/	194	1763	0	396

s41t4/	139	1494	14	531
s41t5/	162	1522	0	418
s41t6/	210	1579	6	270
s41t7/	163	1480	0	367
s41t8/	141	1443	0	389
s41t9/	202	1390	5	438
s42t1	848	543	0	1052
s42t10	774	600	8	1076
s42t3	308	439	0	1442
s42t5	1261	272	7	939
s42t6	1436	345	2	934
s42t7	1131	354	4	1259
s42t8	1067	320	3	1103
s42t9	703	582	3	617
s43t1	1895	1788	0	2175
s43t2	2046	1200	0	1344
s44t10	149	1440	0	141
s44t4	76	1906	0	114
s44t6	71	1632	0	139
s44t7	102	1888	0	158
s44t8	72	1466	0	138
s44t9	63	1937	0	137
	16860	46028	84	21981

1.1.8 Tracker Tree – Upper Body

29_27Jul2010_KAS_Elementary_Configuration				
	True Positive	True Negative	False Postive	False Negative
scene_01_DH_take_a	60	0	14	2
scene_01_DH_take_b	78	24	0	18

scene_02_DH_take_a	96	14	2	12
scene_02_DH_take_b	89	16	0	16
scene_03_DH_take_a	126	23	14	17
scene_03_DH_take_b	128	9	11	22
scene_04_DH_take_a	142	28	1	17
scene_04_DH_take_b	128	14	1	19
scene_05_DH_take_a	102	2	0	14
scene_05_DH_take_b	124	7	2	17
scene_06_DH_take_a	138	14	0	18
scene_06_DH_take_b	135	15	1	24
scene_07_DH_take_a	141	4	12	18
scene_07_DH_take_b	131	3	7	24
scene_08_DH_take_a	123	8	0	24
scene_08_DH_take_b	124	11	0	23
scene_09_DH_take_a	121	4	1	29
scene_09_DH_take_b	117	9	0	24
scene_10_DH_take_a	122	12	0	22
scene_10_DH_take_b	127	8	1	20
scene_11_DH_take_a	141	11	5	6
scene_11_DH_take_b	143	10	5	7
scene_12_DH_take_a	157	10	3	5
scene_12_DH_take_b	153	7	6	9
scene_13_DH_take_a	108	6	5	21
scene_13_DH_take_b	0	5	0	125
scene_14_DH_take_a	107	16	1	16
scene_14_DH_take_b	108	11	2	15
scene_15_DH_take_a	50	43	10	37
scene_15_DH_take_b	56	47	3	34
scene_16_DH_take_a	74	46	1	25
scene_16_DH_take_b	63	46	2	24
scene_17_DH_take_a	178	17	5	20
scene_17_DH_take_b	158	11	9	22
scene_18_DH_take_a	172	13	1	14

scene_18_DH_take_b	167	17	4	22
scene_19_DH_take_a	133	3	5	24
scene_19_DH_take_b	126	5	3	26
scene_20_DH_take_a	130	14	2	14
scene_20_DH_take_b	135	10	3	15
scene_21_DH_take_a	144	3	18	15
scene_21_DH_take_b	132	14	12	12
scene_22_DH_take_a	123	14	2	21
scene_22_DH_take_b	126	18	0	21
scene_23_DH_take_a	92	6	5	37
scene_23_DH_take_b	86	5	3	40
scene_24_DH_take_a	82	11	0	47
scene_24_DH_take_b	94	13	0	41
scene_25_DH_take_a	103	2	0	45
scene_25_DH_take_b	87	12	0	45
scene_26_DH_take_a	99	12	6	33
scene_26_DH_take_b	108	8	0	30
scene_27_SG_take_a	49	5	1	35
scene_27_SG_take_b	63	6	0	27
scene_28_SG_take_a	75	10	2	33
scene_28_SG_take_b	49	8	2	41
scene_29_SG_take_a	102	6	0	42
scene_29_SG_take_b	122	0	0	38
scene_30_SG_take_a	95	7	1	37
scene_30_SG_take_b	114	3	0	33
scene_31_SG_take_a	60	6	9	45
scene_31_SG_take_b	81	1	3	40
scene_32_SG_take_a	52	2	10	50
scene_32_SG_take_b	79	12	11	38
scene_33_SK_take_a	144	0	11	151
scene_33_SK_take_b	131	2	10	157
scene_34_SK_take_a	145	4	10	101
scene_34_SK_take_b	143	3	7	97

scene_35_SK_take_a	165	3	10	162
scene_35_SK_take_b	93	0	9	200
scene_36_SK_take_a	199	5	5	101
scene_36_SK_take_b	115	4	11	170
scene_37_SK_take_a	124	0	11	185
scene_37_SK_take_b	117	0	7	176
scene_38_SK_take_b	144	3	19	144
scene_39_DH_take_a	205	11	6	38
scene_39_DH_take_b	227	16	8	34
scene_40_DH_take_a	246	11	0	33
scene_40_DH_take_b	252	14	0	39
scene_41_DH_take_b	192	2	5	126
scene_42_DH_take_a	173	10	0	127
scene_43_SG_take_a	153	2	6	242
scene_46_SG_take_a	125	7	4	262
scene_46_SG_take_b	136	4	10	260
scene_47_SG_take_a	203	0	8	253
scene_49_SK_take_a	133	0	0	285
	10493	858	384	5050
22_17Mar2010_Zurich_living_lab				
scene_01_woman_telephone_take_c	847	0	0	73
scene_01_woman_telephone_take_d	876	0	0	54
scene_03_enter_room_fab_dan_light_take_b	575	0	0	69
scene_03_enter_room_fab_dan_take_c	553	0	0	77
scene_03_enter_room_fab_dan_take_e	603	0	0	57
scene_07_fab_standup_talkshimself_light_take_d	518	0	1	101
scene_07_fab_standup_talkshimself_light_take_e	476	0	0	147
scene_07_fab_standup_talkshimself_take_b	576	0	0	64
scene_07_fab_standup_talkshimself_take_c	565	0	0	105
scene_02_knocking_light_take_c	256	41	4	119
scene_02_knocking_take_a	235	27	3	131
scene_02_knocking_take_b	89	21	87	127

scene_03_enter_room_fab_dan_light_take_a	281	29	0	278
scene_03_enter_room_fab_dan_take_d	384	27	2	271
scene_04_dan_radio_active_light_take_c	577	66	0	209
scene_04_dan_radio_active_light_take_d	601	60	0	179
scene_04_dan_radio_active_take_a	418	51	0	167
scene_04_dan_radio_active_take_b	325	64	110	209
scene_05_dan_radio_remote_light_take_a	520	54	6	44
scene_05_dan_radio_remote_light_take_b	411	59	1	69
scene_05_dan_radio_remote_take_d	393	54	0	81
scene_10_picksup_dan_light_take_c	72	42	3	75
scene_10_picksup_dan_light_take_d	120	33	3	144
scene_10_picksup_dan_take_a	55	38	4	83
scene_10_picksup_dan_take_b	60	44	1	87
scene_10_picksup_dan_take_e	139	51	0	122
scene_11_fab_picking_up_light_take_a	147	42	9	54
scene_11_fab_picking_up_light_take_b	104	32	7	73
scene_11_fab_picking_up_take_c	134	42	3	61
scene_11_fab_picking_up_take_d	118	47	1	98
scene_13_fab_tie_laces_light_take_a	76	47	1	224
scene_13_fab_tie_laces_light_take_b	131	38	1	262
scene_13_fab_tie_laces_take_c	167	55	5	265
scene_13_fab_tie_laces_take_d	186	60	0	222
scene_14_ste_stumbling_light_take_c	117	47	7	45
scene_14_ste_stumbling_light_take_d	120	51	6	63
scene_14_ste_stumbling_take_a	109	47	4	68
scene_14_ste_stumbling_take_b	24	49	2	129
scene_15_ste_falling_complete_light_take_a	113	59	10	130
scene_15_ste_falling_complete_light_take_b	69	52	2	201
scene_15_ste_falling_complete_take_c	124	56	4	164
scene_16_fab_oov_couch_light_take_c	100	21	2	381
scene_16_fab_oov_couch_take_e	99	27	5	361
scene_16_fab_oov_couch_take_f	109	35	0	336
scene_17_ste_falling_down_light_take_c	20	48	6	190

scene_17_ste_falling_down_light_take_d	27	60	3	246
scene_17_ste_falling_down_take_a	12	44	4	168
scene_17_ste_falling_down_take_b	33	46	8	249
scene_18_fab_picking_up_falling_down_limping_light_take_c	111	49	5	291
scene_18_fab_picking_up_falling_down_limping_take_a	127	42	0	239
scene_18_fab_picking_up_falling_down_limping_take_b	155	43	5	217
scene_20_fab_hits_limping_speech_light_take_a	185	60	3	184
scene_20_fab_hits_limping_speech_light_take_b	235	65	7	149
scene_20_fab_hits_limping_speech_take_c	168	66	9	153
scene_20_woman_hits_limping_speech_light_take_f	143	49	5	211
scene_20_woman_hits_limping_speech_take_d	105	74	1	216
scene_20_woman_hits_limping_speech_take_e	103	65	4	248
scene_21_fab_walk_behind_couch_light_take_c	214	41	7	2
scene_21_fab_walk_behind_couch_light_take_d	177	34	5	36
	28301	3313	896	17320

DIRAC

Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200 (B.C.))

THE CONVERSATION DETECTOR

FRAUNHOFER INSTITUTE FOR DIGITAL MEDIA TECHNOLOGY,
PROJECT GROUP HEARING, SPEECH AND AUDIO TECHNOLOGY
(FRA)

Abstract:

The multi-modal Conversation Detector uses the DIRAC principle of incongruency detection to discriminate between normal conversation and unusual conversational behavior, e.g. a person talking to himself. Detectors operating in the audio and visual domain are combined into a DIRAC model, thus integrating the efforts of several project partners and work packages. The developed Conversation Detector is evaluated against recorded audio-visual data from the Dirac data base. Input and output signals of the model together with the audio visual data have been rendered in a video for demonstration.

Table of Content

1. Introduction

One of the main objectives in Work Package 6 is to integrate the developments of the other work packages, and to apply these to real-world audio-visual data.

Following this objective, FRA has developed a so-called "conversation detector" to detect unusual behaviour of (aged) people, e.g. a person talking to himself, whether two or more persons talking is considered normal conversation. The multi-modal "conversation detector" operates on audio-visual input data and uses the DIRAC paradigm of incongruency to separate normal from unusual behaviour.

With the development of this detector, FRA has integrated different developments, techniques and methods from all other DIRAC work packages: audio-visual pre-processing (developed in WP1)[1], audio localisation and speech classification (WP2)[2], video classification with the tracker tree (WP3)[3], the Dirac reasoning of incongruence and model description (WP4, WP5)[4,5], and data gathering [6] with the AWEAR-II platform (WP6)[7].

The modelling of the conversation detector uses the DIRAC paradigm on a meta-level, i.e. the input data of the model is generated by other DIRAC models which in turn are driven by detector data.

The conversation detector has been evaluated on recordings from the audio-visual data base presented in Deliverable D6.9 "Database of Recordings of Scenario 1".

2. Modelling

The detector model is composed instantiating a part-whole relationship: three models on the general level are combined to one conjoint model; one model on the specific level uses the fused data input of each model on the general level (see Fig. 1).

2.1 Models

The models on the general level are:

- the Person Localizer (PL); this model is driven by output data of the so-called "Tracker Tree" developed in WP3. The Person Localizer detects one or more persons present in each video frame.
- the Audio object Localizer (AL); this model is driven by output data of the audio localization detector developed in WP2. The Audio object Localizer detects one or more sound sources present.
- the Speech/non-speech Classificator (SC); this model is driven by output data of the speech/non-speech model developed in WP2. The data is classified as speech or non-speech (no localization or sound source separation)
- The conjoint model (&) logically combines the output of the above models.

The model on the specific level (Cf) utilizes a linear support vector machine and operates on the same (albeit fused) input data as the models on the general level.
An incongruency is detected when the conjoint model accepts the input as conversation, whereas the specific model does not.

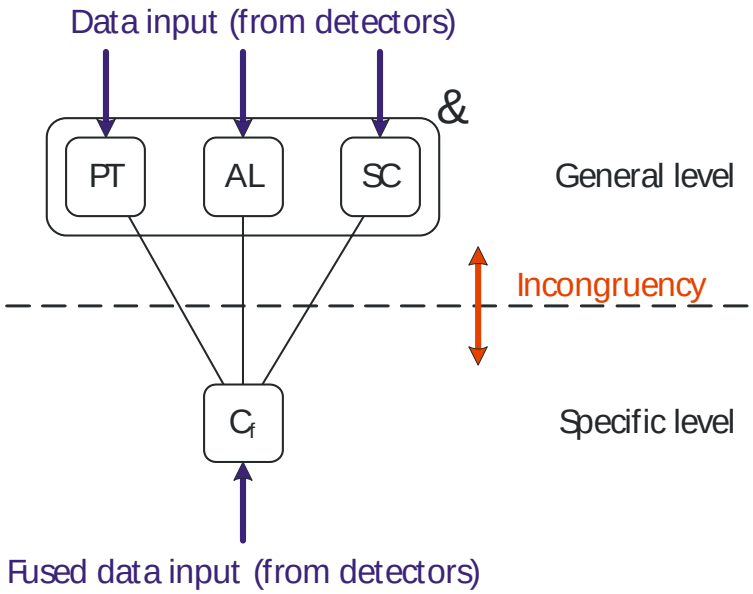

Figure 1: Part-whole relationship model of the "Conversation Detector"

All models generate a binary decision as output:

- the Person Tracker (PT) signals "1" if any person is visible, "0" otherwise.
- the Audio source Localizer (AL) signals "1" if any sound event is detected, "0" otherwise.
- the Speech Classifier (SC) signals "1" if the sound event is classified as speech, "0" otherwise.
- the Conjoint Model (&): uses the output of PT, AL and SC as input. The output of the Conjoint Model is "1" if all three inputs are 1 (logical AND function), "0" otherwise.
- The specific Model (Cf, trained SVM) signals "1" for conversation, "0" otherwise.

As a pre-requisite to data fusion, time and spatial resolution of each of the different detectors in use ("tracker tree", audio localisation and speech classification) have to be interpolated. This Interpolation is done in a pre-processing step.

2.2 Pre-processing of input data

Each data stream from the detectors has a different time and special resolution:

- The "Tracker tree" data is rendered on a time basis of 12 video frames per second (about 83 milliseconds) . For each frame, the position of the position of each detected person is given as box coordinates within a range of 1600x1200 pixel. The coordinate system is distorted due to the cylindrical projection of the fish-eye lens of the AWEAR-II recording device.
- The sound source localization data has a time resolution of 10 milliseconds and an angle resolution of 0.7 degree, i.e. the 180 degree localization area is divided into 257 sectors. The localization accuracy is lower, which means that one sound source is usually smeared over several adjacent bins.
- The speech/non-speech classification data consists of a binary yes/no information on a time basis of 500 milliseconds.

In a first pre-processing step, time resolution for all input data streams is interpolated to 12 frames/second. In a second step, the video localization coordinates are mapped to the angle resolution of the audio localization coordinates using a piecewise linear mapping function derived from calibration data recorded in WP6. In a third step, the position of each person and each sound localization is mapped into an interval of bins with fixed width.

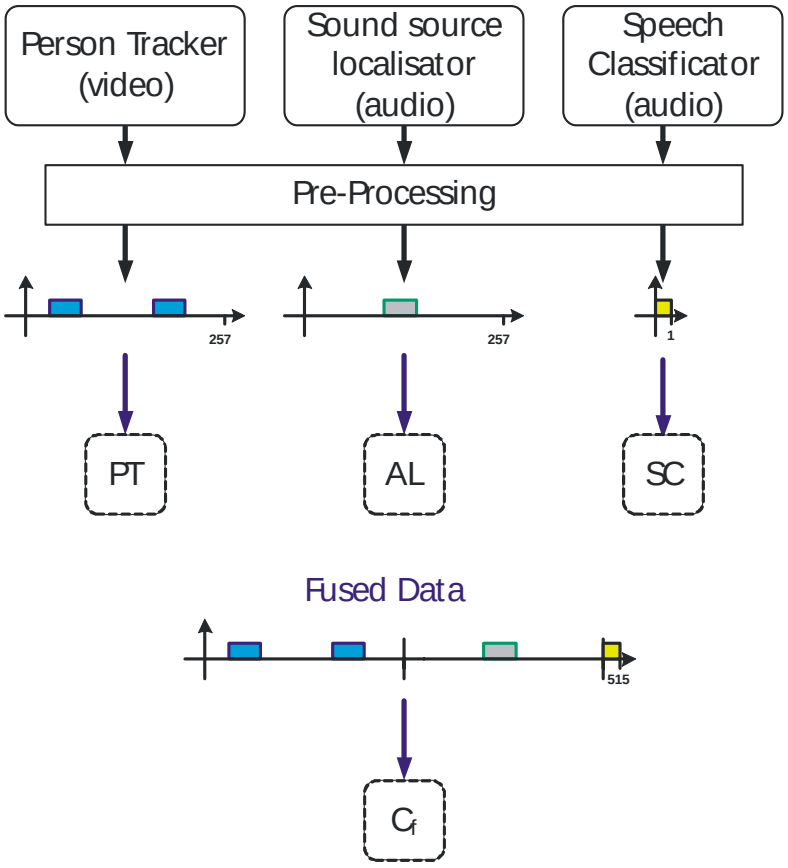

Figure 2: Data format after pre-processing step

After this pre-processing steps, the localization information of both modalities (video and audio) is given as a vector of 257 entries each on a time basis of 12 frames per second. The entries are "1" for object localized and "0" otherwise. The speech/non-speech information is given as a binary information for each frame. The fused data vector has dimension 515 (2x257+1) with a time basis of 12 frames per second (see Fig. 2).

2.3 Training of the specific Model (Cf)

The SVM used for the specific Model (Cf) was trained on synthetic data simulating eight different situations:

1. one person visible, one audio event, no localization overlap, no speech -> no conversation
2. two persons visible, one audio event, localization overlap, speech -> conversation
3. two persons visible, one audio event, localization overlap, no speech -> no conversation
4. one person visible, two audio events, localization overlap, speech -> no conversation
5. two persons visible, no audio event, no speech -> no conversation
6. no person visible, no audio event, no speech -> no conversation
7. one person visible, no audio event, no speech -> no conversation
8. one person visible, one audio event, no overlap, speech -> no conversation

The training set for the SVM consisted of 11200 generated vectors with random person/sound localization positions. The vectors were chosen from all situations (5600 for situation 2, 5600 for situation 1, 3-8 for a chance level of 50%). An example vector for each situation is depicted in Figure 3.

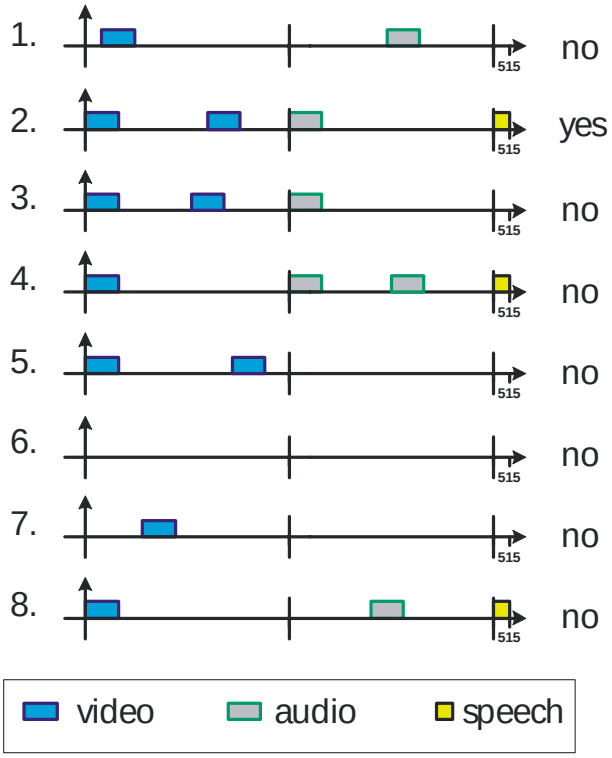

Figure 3: Example representations of training data vectors

Labels for the training set were generated automatically using the following rule:

> Whenever more than one person is detected and at least one audio event is detected
> AND the localization derived from video and audio data at least partly overlap
> AND the audio is classified as speech,
> THEN the data vector is labeled "1" for conversation;
> OTHERWIZE the data vector is labeled "0".

The trained linear SVM was tested against a synthetic test set of 11200 vectors drawn from each of the eight situations, again with chance level 50%. The test set was classified 100% correct, no false positives or false negatives occurred.

3. Results

In a first test, he trained specific model was evaluated not only against generated test data, but also against data derived from audio-visual recordings taken from the indoor data base (D6.9). In a second test, the DIRAC model producing the incongruency between the general and specific model was tested against annotation of the expected incongruency, i.e. the monologue of a single person.

The test set drawn from the audio-visual recordings consisted of 5484 frames taken from eight recordings, four dialog recordings containing typical conversation (2724 frames), and four monologue recordings (2760 frames). The fused data derived from the recordings, i.e. the input data of the specific model, was again labeled automatically. Additionally, all frames were labeled w.r.t. the expected incongruency, i.e. the monologue of a single person. This second labeling formed the ground truth for the evaluation of the DIRAC model producing the incongruency between the general and specific model.

The evaluation of the specific model (the trained linear SVM), produced the following results:

- 95,45% correctly classified frames for the dialog recordings (4,55% false negatives), 100% correctly classified frames for the monologue scenes.
- Altogether performance: 97,73% correctly classified frames (2,27% false negatives).

The evaluation of the DIRAC model produced the following results:

- Overall performance: 98,45% correctly classified frames (1,55% false positives),
- Dialogue performance: 98,86% correctly classified frames (1,44% false positives),
- Monologue performance: 98,04% correctly classified frames (1,96% false positives).

More detailed evaluation results are published in deliverable D6.13 “Evaluation Results”.

4. Visualization

Audio-visual videos were rendered to demonstrate the "conversation detector" operating on the audio-visual data. Together with the original audio and video data (after cylindrical projection), the pre-processed model input and output data have been rendered in a video.
Figure 4 shows a frame of the monologue recording. Below the picture, all input signals (after the pre-processing step) are visualized:

- (V) shows the video localization (blue) of the person,
- shows the audio localization (magenta)
- (S) show the binary Speech (yellow) /no Speech (black) classification
- (K) shows the binary result of a Knock detector (not relevant for this demonstrator)

For the depicted frame, the video and audio localization overlap, and the audio event is classified as speech.
To the right of the picture in Figure 4, all model output signals are displayed; the columns indicate the output signals of all general and specific models and the Incongruency indicator (colored bars indicating "1", not colored bars indicating "0"):

- (V) shows the output of the Person Tracker (PT),
- (A) shows the output of the Audio source Localizer (AL),
- (S) shows the output of the Speech Classifier (SC),
- (&) shows the output of the Combined Model (&),
- (F) shows the output of the Specific Model (Cf)
- (I) shows the detected incongruency

For this frame, the presence of a speaking person detected by the general models (columns 'V','A' and 'S' rendered with color) produces a "1" for the conjoint model (column '&' rendered with color), whereas the specific model indicates "0" for "no conversation" (column 'F' rendered black). The different output of the general and the specific model indicates incongruency for this frame ('I' rendered with color).

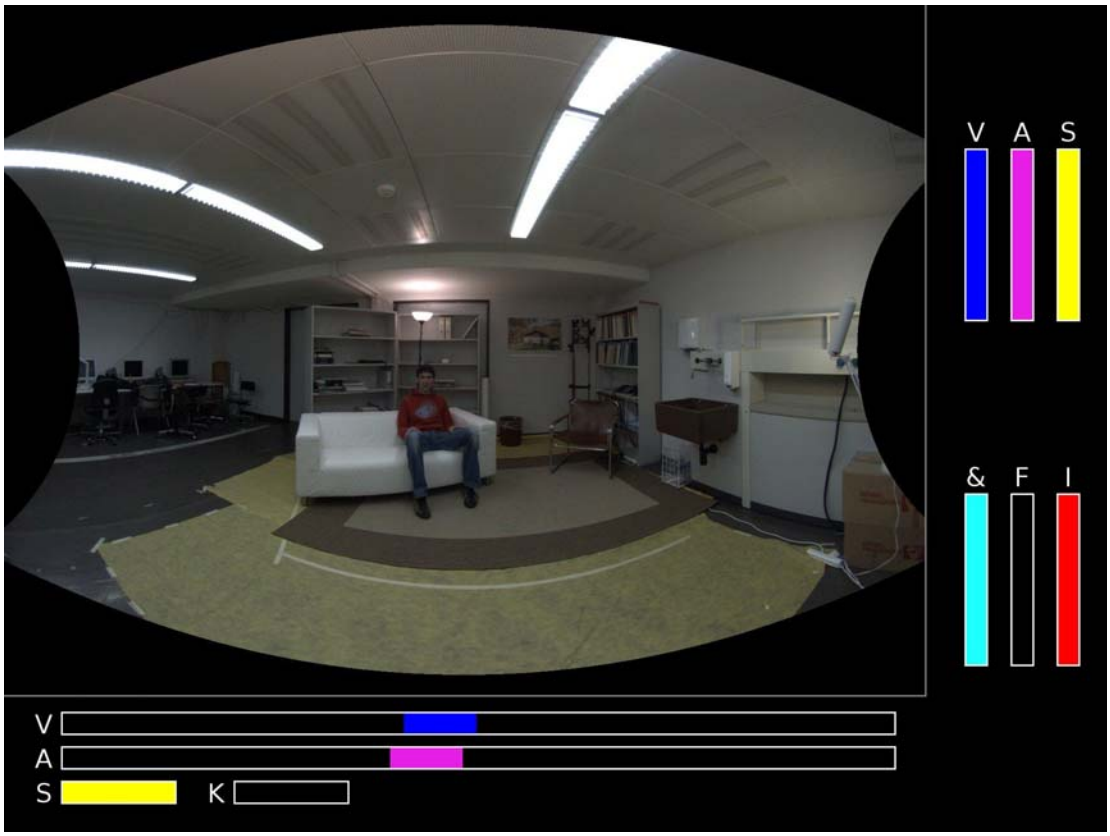

Figure 4: Video demonstration of the "conversation detector" operating on a recorded scene

5. Conclusion

With the "conversation detector", this deliverable presented a successful implementation of the DIRAC incongruency modelling using the part-whole relationship. Combining the results of all work packages of the DIRAC project, a typical home-care situation was addressed and analysed using DIRAC reasoning. The modelling was evaluated against audio-visual recordings taken from the DIRAC database (D6.9), showing excellent overall performance. The visualization of two recordings together with all model input and output data was rendered and presented at the Year 4 review meeting.

6. Reference

[1] D1.7 "Low level processing modules for AWEAR demonstrator" (CTU) DIRAC, December 2008.
[2] D2.11 "Hierarchical acoustic scene classification scheme with detection of unexpected scenes" (OL) *DIRAC,* June 2009.

[3] D3.9 "Incongruences detected between trackers working with weaker and stronger expectations about the world" (ETHZ) *DIRAC,* December 2009.

[4] M. Pavel, H. Jimison, D. Weinshall, A. Zweig, F. Ohl, H. Hermansky "Detection and Identification of Rare Incongruent Events in Cognitive and Engineering Systems" D*IRAC White Paper*. DIRAC 2 April 2008.

[5] D. Weinshall et al. "Beyond Novelty Detection: Incongruent Events - when General and Specific Classifiers Disagree" *NIPS* 2008.

[6] D6.9 ""Database of Recordings of Scenario 1" (FRA), *DIRAC,* December 2009.

[7] Anemüller, J., Bach, J.-H., Caputo, B., Havlena, M. Jie, L., Kayser, H., Leibe, B., Motlicek, P., Pajdla, T., Pavel, M., Torii, A., Gool, L. v., Zweig, A. and Hermansky, H. „The DIRAC AWEAR Audio-Visual Platform for Detection of Unexpected and Incongruent Events" *Proc. International Conference on Multimodal Interaction (ICMI)* 2008, pp. 289-293.