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: D6.14
The DIRAC Databases

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.))

D6.14 – The DIRAC Databases

Fraunhofer Institute of Digital Media Technology, Project Group Hearing, Speech and Audio Technology (FRA)

Abstract:

This deliverable gives an overview over the different databases that have been generated within the different research fields of the DIRAC project. The material of the databases was assembled in a collaborative effort by different partners and used to evaluate the development within the different research strands of the project.

Table of Contents

1 Introduction

The one goal of the DIRAC project was to established the paradigm of incongruency as a novel method of information retrieval within a model hierarchy. Research was inspired not only by theoretical questions, but also by the urge to find physiological evidence for special forms of hierarchies and models. With the incongruency reasoning formulated, fostered and the principles tested with first experiments conducted by the partners, a growing need for experimental data to further investigate and evaluate the findings led to the aggregation of different databases within the project. Over time, databases were assembled and used by the partners to drive their development and do verify their findings.

2 The DIRAC Databases

The following sub-sections will give descriptions of the database of audio-visual recordings, the database of multichannel in-ear and behind the ear head related and bin-aural room impulse responses, the database of frequency modulated sweeps for STRF estimation, the database of OOV and OOL recordings and the database of motion captured actor walking on a treadmill.

2.1 *Database of audio-visual recordings*

Within the DIRAC project, two application domains were defined for rare and incongruent event detection, namely the security and surveillance [4] on the one hand, and in-home monitoring of elderly people on the other [5]. In both domains there is a need for 24/7, unobtrusive, autonomous, and therefore intelligent, monitoring systems to assist human observers. The use of sound to augment camera surveillance in the security market plans to use it in in-home monitoring emphazises Spotting and properly responding to unforeseen situations and events as one of the crucial aspects of monitoring systems in both application domains.

Based on these application domains, scenarios have been developed by the project partners to show the potential of the DIRAC theoretical framework and the techniques developed within the project, while attempting to address realistic and interesting situations. These scenarios in turn were recorded by different partners using professional audio and video recording hardware assembled into two recording platforms, the AWEAR II and the OHSU recording setup. First recordings were tested by the partners with the developed detectors and incongruency models, which in turn led to discussions about the detectors as well as the recorded scenes. During the project, the partners refined both the developed detectors and the scene descriptions. That – over time – formed the audio-visual database of the DIRAC project. Several hundred recordings were recorded in more than 50 recording sessions at locations from FRA, ETHZ, OHSU and OL, and were described in detail in the deliverables D6.6, D6.7, D6.8, D6.9, D6.10 and D6.12. The raw data had to be pre-processed (format changes, projections to correct lens distortion, synchronization between video and audio, preview video with reduced resolution etc.) prior to the application of the detectors and model by the partners. The data was categorized using a list of keywords and time stamps to help partners to identify an search for different human actions contained in each recording. For evaluation purposes, the audio and video data of the recordings were annotated on a frame by frame basis, giving the pixel position of different body parts of the actor, e.g. the pixel position of the head, the upper and lower body etc. or the speech/non-speech classification of the recorded audio signal as time positions.

Recording platform

The AWEAR-II recording platform is a portable audio-visual recording platform that was developed in the DIRAC project. It consists of three mini-pcs (Siemens D2703-S mini ITX boards), two cameras (AVT Stingray), four microphones (Tbone EM700 stereo mic set), an audio capturing interface including a triggering device for audio-visual synchronization (FOCUSRITE Saffire PRO 10), a battery pack (Camden 12V gel) and a power distribution box. The system can be controlled using a graphical user interface (GUI) application running on a netbook which in turn communicates with the recording platform via a wireless network connection. The GUI is used as a remote control for the preparation of recording sessions and capturing, for hardware adjustments and controlling. The hardware is mounted on a wearable backpack frame and allows human-centered recordings in both indoor and outdoor environments. Further information on the recording platform can be found in [7], [9]. A second recording setup was used for the recordings done in the living room setups of partner OHSU. The setup consisted of one high-resolution camera with firewire connection and two cardioid condensor microphones mounted on one tripod. Further information about the recording setup can be found in [18].

Data Format

In order to provide the best possible quality to the research community and the DIRAC consortium, all the recordings are stored in an uncompressed data format. For the recorded video signals, the Portable Network Graphics (PNG) [10] format is used, i.e. each recorded frame from each camera is stored as a separate png picture with a resolution of 1920x1080 pixel. All four microphone channels of each recording is stored as a separate file in the PCM-Wave format, with 48kHz sampling rate and a sample resolution of 32 bit floating-point. The audio-visual recordings themselves as well as detailed information on the recording setup, procedures, metadata, ground truth annotations and preview videos are available online at [1] and [2].

Ground Truth Annotation – Audio

The recorded audio material was annotated manually for ground truth generation. For the following audio events, start and stop positions were annotated: speech present, no speech present, person moans, person coughs, key hits table/ground, knocking on table/door/ground, dish noise.

A semi-supervised labeling tool developed by FRA was used to facilitate the annotation process. This tool displays the waveform of an audio file and lets the annotator select multiple segments for instantaneous playback and label creation. The different labels were defined with respect to the context of the audio-visual scene description and the purpose of the recording. Within the project, it was agreed on a common vocabulary [16].

With the audio annotation tool, the labels and their temporal locations, i.e. the start and end sample index of a labeled segment, can be exported in various formats to ensure compatibility with the common office, scientific computation, annotation and machine learning software, such as Microsoft Excel, MATLAB and the HTK Framework [11]. Detailed information on the ground truth annotation procedure can be found in [17].

Ground Truth Annotation - Video

The video data annotation can be used as ground truth for each single model in the so-called "Tracker Tree" developed by project partner ETHZ [12]. For every frame, a 2-D pixel position and a binary visibility information (Yes/No) were annotated for the person, head, upper body and lower body; the following different actions were annotated: sitting, stumbling, lying, walking, limping, standing, picking up something. To comfortably annotate the video data, FRA developed a MATLAB based annotation tool. With this tool, labels are set using the PC mouse for every frame.

Detailed information on the annotation procedure and the ground truth labels are given in [17].

Evaluation

The database of audio-visual recordings was used to evaluate different models, i.e. the models of the tracker tree, the audio localization, the speech/non-speech classification an the models of the conversation detector. For this evaluation, the data was processed, and the model outcomes were compared against the annotation of the recordings. The data involved, the processing, the evaluation procedure and the results are described in detail in Deliverable D6.13.

2.2 *Database of multichannel in-ear and behind the ear head related and binaural room impulse responses*

An eight-channel database of head-related impulse responses (HRIR) and binaural room impulse responses (BRIR) was generated within the DIRAC project. The impulse responses (IR) were measured with three-channel behind-the-ear (BTE) hearing aids and an in-ear microphone at both ears of a human head and torso simulator. The database aims at providing a tool for the evaluation of multichannel hearing aid algorithms in hearing aid research. In addition to the HRIRs derived from measurements in an anechoic chamber, sets of BRIRs for multiple, realistic head and sound-source positions in four natural environments reflecting daily-life communication situations with different reverberation times are provided.

The scenes' natural acoustic background was also recorded in each of the real world environments for all eight channels. Overall, the present database allows for a realistic construction of simulated sound fields for hearing instrument research and, consequently, for a realistic evaluation of hearing instrument algorithms. The database is available online at [1] and is described in more detail in [13].

2.3 *Database of OOV and OOL recordings*

Two data sets of audio recordings have been produced by project partner BUT. The first one is a set of utterances containing Out-Of-Vocabulary (OOV) words and non-speech sounds, and the second one contains English In-Language (IL) spontaneous speech featuring intermittent switches to a foreign language (Out-Of-Language – OOL). The OOV recordings and the OOL recordings as well as detailed information are available online at [1] and [2]. Segmentation and Annotation was on a 10s chunk basis.

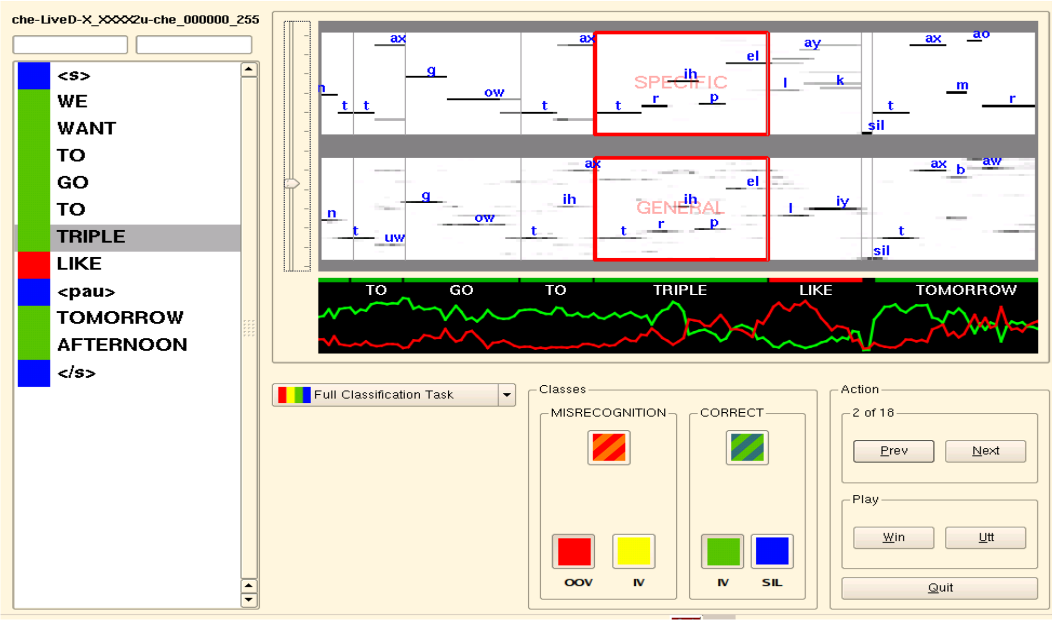

Illustration 1: Visualization of detected OOV word "TRIPOLI", OOV discongruency score (in red)

Furthermore, a NN-based OOV word detection visualization tool bundled with the OOV recordings shown on the review in 2009 has been made available on reviewers and partners request. It can be run under Linux and Windows, and be recompiled, its sources can be found online at [1] and [2].

2.4 *Database of frequency modulated sweeps for STRF estimation*

The spectro-temporal receptive field (STRF) is a common way to describe which features are encoded by auditory and visual neurons. STRFs are estimated by relating stimuli, visual or auditory, to the evoked response ensemble of the neuron. Once an STRF has been estimated it can be used to predict the linear part of the response of the neuron to new stimuli.

A new class of stimuli was generated within the DIRAC project consisting of frequency modulated (FM) sweeps. FM sweeps are an alternative to dynamic moving ripples (DMR) that are often used for STRF estimation. Both stimulus classes are designed such that they sample a large portion of the neuron's input space while having very low autocorrelations which makes them suitable for STRF estimation. However, as there is one main feature to divide the sweep class into two, the up and the down sweep, this characteristic can be used to investigate context dependence of the STRF. In [18] it is shown that neurons as well as local field potentials (LFPs) are sensitive to different sweep orientation. Hence, one could estimate STRFs using a varying number of up and down sweeps to realize different context conditions. The database and example scripts are available online at [1] and [2] and is described in more detail in [18].

It has been shown that there are shortcomings in the STRF and its estimates cannot account for arbitrary stimuli. The impact on the scientific community is a potentially better description of the STRF. Futhermore, it is possible realize different context conditions which may yield new insights into encoding strategies on the level of single neurons. In terms of DIRAC it would give a direct comparison if a stimuli is standard or deviant

2.5 *Database of motion captured actor walking on a treadmill*

Within Work Package 3, a study was set up by KUL to investigate how temporal cortical neurons encode actions differing in direction, both forward versus backward as different facing directions.

The stimuli in this study were locomotory actions, i.e. displacement of a human body, but whereby the translational component is removed, thus resembling an actor as if locomoting on a treadmill. This was done in order to more accurately pinpoint the neural code subserving walking direction (walking left- or rightward either in a forward or backward fashion), for which actual physical displacements of the body are erroneous in determining the neural code.

Computationally, coding between forward and backward versions differs with respect to which aspects of the actions are relevant compared to coding for different facing directions. Therefore, a motion-capture of an actor was generated; the actor walks at a normal pace on a treadmill, from multiple camera angles: in the lateral (0° and 180°), frontal (90° and 270°) and oblique planes (45°, 135°, 225° and 315°). All possible combinations amount to 16 actions in total. Also, for each action, a number of static frames (i.e. snapshots) representative of that action ('static_poses'-postfix) were extracted. Finally, also upper-body and lower-body only action variants ('UppBody' - and 'LowBody' - postfixes) were generated [15]. All the stimuli have been made available as *.avi such that playback is not limited to certain devices.

Monkey subjects were instructed to discriminate either forward from backward actions (i.e. forward-backward task) or different facing directions (e.g. facing to the left or to the right; i.e. view task). Numerous versions of each action were generated by varying the starting position. Besides these actions locomoting at a normal pace of 4.2 km/h, which the monkeys subjects were extensively trained on, also similar actions, but at different speeds (2.5, 6, 8, 10 and 12 km/h: '02-prefixes') were generated. These actions at different speeds were created in order to check how general the learned discriminations were. If the monkey subjects rather learned a general abstract concept of 'walking forward to the right' etc., than one would expect quite good discrimination performances for actions of the same category (e.g. 'forward left-2-right') even though the speed is not equal.

A same logic applied when assessing actor-invariance and tests for the generalization of stimulus versions. Also, actions were generated containing just parts (upper or lower body) of the original full body configuration. This was done in order to examine which part of the body configuration contributed to each task separately. Finally, KUL also monitored discrimination performances when gradually reducing information present in the stimulus, i.e. applying spatio-temporal scrambling. The 'percentages' refer to the level of noise added to the walker. Every 5th frame, the scrambling pattern was reshuffled.

Purely at the temporal level, thus without tampering in the spatial domain, KUL developed a few new stimulus types. In one version, frames were removed at certain intervals and replaced by blank frames. This gives the impression of a stroboscopic experience ('08-prefix'). In a similar manner, instead of replacing frames by blank frames, the frame prior for a number of frames was frozen, before jumping to the next frame in the original action ('07-prefix'). The study was concluded by controlling for opponent motion. KUL generated a number of variant locomotions containing no opponent motion ('09-prefix') [14].

3 Summary

In this deliverable, an overview has been given over the different databases collected by the project partners. Intentions of the project partners which led to the assembly of the databases were described as well as the contents of the different databases and the way they were used in the project.

References

[1] IST-027787 project website: Detection and Identification of Rare Audio-visual Cues - DIRAC. http://www.diracproject.org/

[2] DIRAC Recordings: https://dirac.uni-oldenburg.de/DIRAC

[3] Weinshall D., Hermansky H., Zweig A., Luo J., Jimison H., Ohl F.,and Pavel M.: *Beyond Novelty Detection: Incongruent Events, when General and Specific Classifiers Disagree*. Advances in Neural Information Processing Systems (NIPS), Vancouver, December 2008, (2008)

[4] van Hengel P., Andringa T., "Verbal aggression detection in complex social environments", Fourth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2007, 5-7 September, 2007, Queen Mary, University of London, London, United Kingdom

[5] 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)

[6] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.11 “Testing and evaluation plan”, XX. June 2010

[7] Havlena, M., Ess, A., Moreau, W., Torii, A., Janoek, M., Pajdla, T., Van Gool, L.: *AWEAR 2.0 system: Omni-directional audio-visual data acquisition and processing*. In: EGOVIS 2009: First Workshop on Egocentric Vision. pp. 4956 (2009)

[8] Behrens, T.: *Der 'Kommunikationsakustik-Simulator' im Oldenburger Haus des Hörens*. 31. Deutsche Jahrestagung für Akustik: Fortschritte der Akustik DAGA 2005 (1), München, DEGA e.V., pp. 443–445, (2005)

[9] Hollosi D, Wabnik S., Gerlach S., Kortlang S., "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

[10] Wikipedia: Portable Network Graphics

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

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

[13] Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses by H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier

[14] Vangeneugden J, Vancleef K, Jaeggli T, Van Gool L, Vogels R (2010). Discrimination of locomotion direction in impoverished displays of walkers by macaque monkeys. Journal of Vision, 10:22.1-22.19.

[15] Vangeneugden J, De Mazière P, Van Hulle M, Jaeggli T, Van Gool L, Vogels R (2010). Distinct mechanisms for coding of visual actions in macaque temporal cortex. The Journal of Neuroscience (in press).

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

[17] DIRAC – Detection and Identification of Rare Audio-visual Cues, Annex "Generation of Ground Truth Annotations"

[18] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.10 "Data Recordings Scenario 2: Point of Care monitoring for Remote Care"

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.))

Generation of Ground Truth Annotation

Fraunhofer Institute of Digital Media Technology, Project Group Hearing, Speech and Audio Technology (FRA)

Abstract:

In this document, we describe the generation of ground truth annotation for audio and video data of the DIRAC audiovisual data bases assembled within Work Package 6 (WP6). The ground truth labels generated by annotation form the basis for the evaluation of the DIRAC methods. The evaluation itself is not part of this deliverable, but is described in D6.13 “Evaluation Results”.

For the generation of the annotation data, we utilize already existing, custom made, semi-supervised annotation tools that have been modified according to the needs of the DIRAC project. These tools will be presented in this work in a way that they can be used by the project partners as well.

Together with the audio-visual scenes in the DIRAC Database that have been selected using the quality criteria described in [10], the ground truth labels will be used for evaluation of the DIRAC methods as it is presented in [12].

Table of Content

1 Introduction

Ground truth annotation is one of the pre-requisites for the evaluation of the methods and procedures that have been developed within the DIRAC project. Besides this project-internal aspect, ground truth annotation is also of great value for the scientific community. Together with the audio-visual recordings, the community will have the possibility to perform their own experiments based on the DIRAC data bases and evaluate them against ground truth annotation.

In this document we describe how we generate the ground truth labels for the DIRAC audiovisual recordings, what tools we use for annotation, how they work and how they should be used. Furthermore, we describe the text file format for the ground truth annotation in more detail and list for which audiovisual recordings from of the DIRAC database ground truth annotation is provided.

The information given in this deliverable is used in Deliverable D6.13 “Evaluation Results” to prepare the evaluation of the DIRAC detectors and methods for rare and incongruent event detection.

2 Ground truth label annotation

2.1 Audio

Manual labeling for ground truth generation is a very time-consuming process especially for large files and databases. To minimize the working effort, a semi-supervised labeling tool was developed at FRA. This tool provides the possibility to visualize the waveform of an audio file and to select multiple segments within the file via mouse clicks for instantaneous playback and individual label creation. An example can be seen in Figure 1.

The labels can be defined freely, however, the context of the audio-visual scene, the scene descriptions and the purpose of the recording should be taken into account, too. For this purpose, we refer to [7] where a common vocabulary for the application scenarios *in-home care* and *security surveillance* is described.

Once a scene-based annotation session is complete, the labels and their location in time, i.e. the start and end sample index of a labeled segment, can be exported in various formats to ensure compatibility with the common office, scientific computation, annotation and machine learning software, such as Microsoft Excel, MATLAB and the HTK Framework [1].

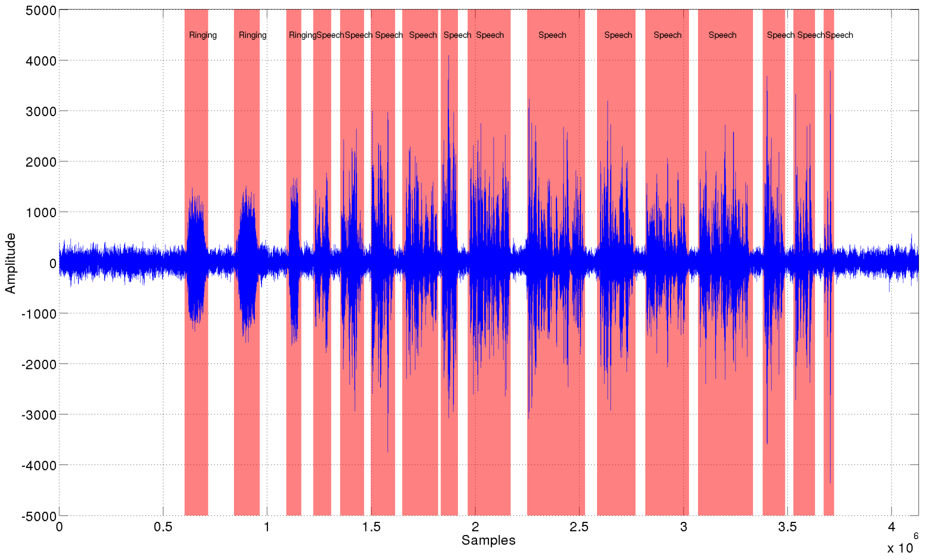

Figure 1: Example of the labeling tool developed by FRA for manual annotation of audio recordings and ground truth generation

For our purpose, the labels are stored in the HTK format. This format has the advantage of providing an audio file length independent representation of the sample indexes of each label in an audio-visual recording. The HTK representation of the sample index can be calculated using Eq. 1, where S denote the sample index and f_S is the sampling frequency of the actual audio recording.

$$S_{HTK}=\frac{S}{f_S}\cdot 10^7 \qquad (1)$$

The original representation can be restored using Eq. 2.

$$S=\frac{S_{HTK}}{10^7}\cdot f_S \qquad (2)$$

For further processing of the data, it is important to note that label-specific filters can be applied to the annotations during import. Thus, obtaining speech or segments labeled “speech-like” out of the huge DIRAC dataset for example is an easy to do task.

2.2 Video

We annotate the video data according to the output of the video-based DIRAC method for human movement and action detection developed by ETHZ [11]. Additionally, we provide ground truth labels that are currently not in use to cover future developments within the DIRAC project and to ensure a re-usability of the DIRAC audio-visual database for research apart from the DIRAC project.

Table 1: Ground truth annotation labels and format conventions for text-file export

class	members	format
actions	Person	logical, [0,1]
	Head	logical, [0,1]
	Upper Body	logical, [0,1]
	Lower Body	logical, [0,1]
	Sitting	logical, [0,1]
	Stumbling	logical, [0,1]
	Lying	logical, [0,1]
	Walking	logical, [0,1]
	Limping	logical, [0,1]
	Standing	logical, [0,1]
	Pick Up	logical, [0,1]
locations	Head Location	pos. integer [x,y]
	Upper Body Location	pos. integer [x,y]
	Lower Body Location	pos. integer [x,y]
	Person Location	pos. integer [x,y]

We define two different classes of labels for annotation. The first class consists of binary labels and is used as an indicator whether a particular action is performed by a single person in the field of view or not. These actions were defined within the DIRAC project for scene descriptions based on a common vocabulary and as representative case studies for the DIRAC application scenarios *in-home care* and *security surveillance* [5-7]. The second class are locations of body parts of a person in view, represented as pairs of x- and y- coordinates in pixels. Both classes, their underlying members as well as their format are summarized in Table 1.

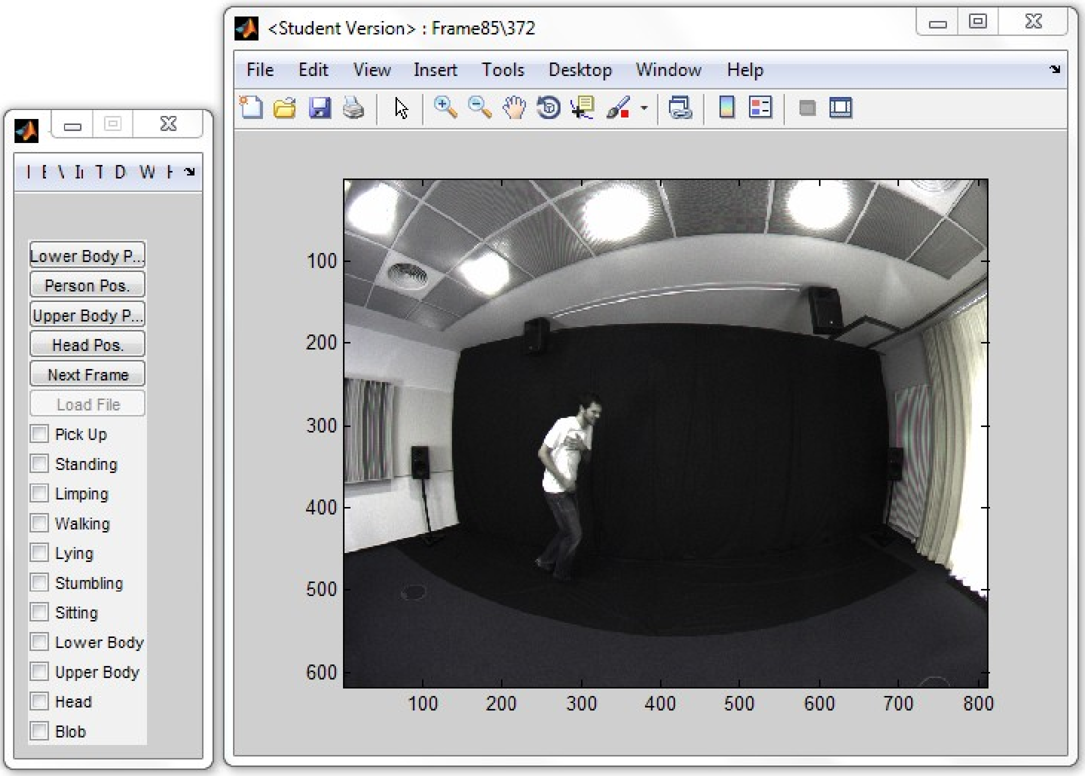

Figure 2: Example of the frame tool for manual annotation of video scenes and ground truth generation

To annotate the video data comfortably, project partner FRA developed a MATLAB based annotation tool. For each individual video frame, labels can be defined by simply using the PC mouse. The tool, as it can be seen from the screenshot in Figure 2, consists of two individual windows: a visualization window and a control window. The latter one contains a number of checkboxes for binary label annotation of specific human actions and movements, their incongruent equivalent and a number of buttons for manual annotation of the video data according to the location of body parts of a person in view. In order to perform the ground truth annotation in a more efficient way, only every third video frame was annotated. The missing labels and coordinates were generated using linear interpolation using the customized MATLAB function *annotation_interp.m*. This is feasible since movements have continuous and/or steady character, such that the error that is introduced to the annotations stays negligible small.

The so-derived ground truth labels are then stored in an annotation text file. The file contains the full path to the current frame from which the ground truth annotations were derived from, the binary labels according to the demands given by the underlying DIRAC method as well as the location and presence of body parts as the absolute position in pixels within a video frame. It is important to note that only single coordinate pairs in x- and y-directions are annotated instead of a whole bounding box. This is due to the fact that the actual size of an object can change within a scene because of the camera setup and the chosen fish-eye lens such that a bounding box approach would not provide the desired information. Additionally, the bounding boxes generated by the DIRAC methods were not developed for accurate tracking of body parts, but to roughly approximate their locations within the scenes and to provide a visual anchor for the presence of body parts.

Each frame path, its binary labels for actions and locations of body parts are separated by a comma. Consecutive frames are separated by a line feed and a semicolon. This has the advantage that the labels can be easily imported for further processing without putting too much effort into text parsing. A text file format was chosen in order to stay independent from any special platform, operation system, computational framework and software development tools such as MATLAB. An example of a ground truth annotation file can be found in Fig. 3.

```
<path>\000001.png, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
<path>\000004.png, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
<path>\000007.png, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
<path>\000010.png, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
<path>\000013.png, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0;
...
```

Figure 3: Example of a ground truth annotation text file.

Ground truth annotation was done on uncalibrated video frames, i.e. directly on the raw data. Despite that the DIRAC methods developed by the project partners require pre-processed video frames, evaluation tasks can now be performed directly on raw data, too. Of course, the ground truth annotations can be transformed in such a way that their projections match the camera properties and parameters for camera calibration. Therefore, the MATLAB function *annotation2setup.m* can be used. This function takes a ground truth annotation text file as an input and transforms coordinates according to the chosen camera setup in a very efficient way. Both the function *annotation2setup.m* and *annotation_interp.m* can be found on the internal DIRAC web page [8] and on the official DIRAC web page [9] .

3 Summary

In this document, we describe methods and procedures to generate and store ground truth annotations of audio-visual material recorded within the DIRAC project. Although generating audio and video ground truth labels is facilitated by this methods, it nevertheless remains a time consuming process, especially since the annotation needs to be accurate enough to meet the temporal requirements given by the DIRAC methods for evaluation as well as the high standards of the scientific community. At the moment, more than 180 audiovisual sequences have been completely ground truth annotated, i.e. both the audio and video modality. For almost every sequence in the DIRAC database, ground truth labels for the video parts are available. Again, it is important to note that ground truth annotation is still a very time consuming process, especially for the video part. The results of the evaluation are presented in [12].

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, B. Kollmeier, 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", XX. 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 D6.11 "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] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.13 "Evaluation"