Journal on Multimodal User Interfaces manuscript No.
(will be inserted by the editor)

RAVEL: An Annotated Corpus for Training Robots with Audiovisual Abilities

Xavier Alameda-Pineda · Jordi Sanchez-Riera · Johannes Wienke · Vojtěch Franc · Jan Čech · Kaustubh Kulkarni · Antoine Deleforge · Radu Horaud

Received: February 23, 2012 / Accepted: July 12, 2012

Abstract We introduce RAVEL (Robots with Audiovisual Abilities), a publicly available data set which covers examples of Human Robot Interaction (HRI) scenarios. These scenarios are recorded using the audio-visual robot head POPEYE, equipped with two cameras and four microphones, two of which being plugged into the ears of a dummy head. All the recordings were performed in a standard room with no special equipment, thus providing a challenging indoor scenario. This data set provides a basis to test and benchmark methods and algorithms for audio-visual scene analysis with the ultimate goal of enabling robots to interact with people in the most natural way. The data acquisition setup, sensor calibration, data annotation and data content are fully detailed. Moreover, three examples of using the recorded data are provided, illustrating its appropriateness for carrying out a large variety of HRI experiments. The RAVEL data are publicly available at: `http://ravel.humavips.eu/`

This work was supported by the EU project HUMAVIPS FP7-ICT-2009-247525.

X. Alameda-Pineda, J. Sanchez-Rieira, J. Čech, K. Kulkarni, A. Deleforge, R. Horaud
INRIA Grenoble Rhône-Alpes
655, Avenue de l'Europe
38330, Montobonnot, France
E-mail: first.last@inria.fr

J. Wienke
Universität Bielefeld
Universitätsstraße 25
33615, Bielefeld, Germany
E-mail: jwienke@techfak.uni-bielefeld.de

V. Franc
Czech Technical University
Technická 2, 166 27 Prague
Czech Republic
E-mail: xfrancv@cmp.felk.cvut.cz

Keywords Audio-visual data set · Binocular vision · Binaural hearing · Action/gesture recognition · Human robot interaction · Audio-visual robotics

1 Introduction

In recent years, robots have gradually moved from production and manufacturing environments to populated spaces, such as public spaces, e.g., museums and entertainment parks, offices, hospitals, homes, etc. There is an increasing need to develop robots that are capable of interacting and communicating with people in unstructured, unconstrained and unknown environments in the most natural way. For robots to fulfill interactive tasks, not only they need to recognize humans, human gestures, human intentions, human speech, etc. they equally need to combine data gathered with different sensory modalities, e.g., vision and hearing, as well as to coordinate their perceptive, communicative and motor skills, i.e., *multimodal interaction*.

In this paper we describe a publicly available data set RAVEL (Robots with Audiovisual Abilities). The data set consists of three categories: human activities, robot-commands recognition and verbal communication. A detailed description of the categories and of the scenarios inside the categories is given below. Figure 1 presents some snapshots of the recorded scenarios in all three categories. All scenarios were recorded using an audio-visual (AV) robot head, shown in Figure 2, equipped with two cameras and four microphones, which provide multimodal and multichannel synchronized data recordings.

Researchers working in multimodal human-robot interaction can benefit from RAVEL for several reasons.

(a) *Talk on the phone* **(b)** *Stop!* **(c)** *Where is the kitchen?* **(d)** *Cheers!*

(e) *Cocktail party* **(f)** *Hand-shaking* **(g)** *Let me introduce you!* **(h)** *Someone arrives*

Fig. 1: Scenario examples from the RAVEL data set. (a) Human activity – *talk on the phone*–, (b) Robot command – *stop!*–, (c) Asking the robot for instructions, (d) Human-human interaction, (e) Cocktail party, (f) Human introducing a new person (g) Robot introducing a new person, and (h) New person.

First of all, four microphones are used in order to be able to study the sound source separation problem; robots will face this problem when interacting with humans and/or other robots. Secondly, the simultaneous recording of stereoscopic image pairs and microphone pairs gives an opportunity to test multimodal fusion methods [24] in the particular case of visual and auditory data. Moreover, the fact that a human-like robot head is used, makes the data appropriate to test methods intended to be implemented on humanoid robots. Finally, the scenarios are designed to study action and gesture recognition, localization of auditory and visual events, dialog handling, gender and face detection, and identity recognition. In summary, many different HRI-related applications can be tested and evaluated by means of this data set.

The RAVEL data set is novel since it is the first data set devoted to study the human robot interactions consisting of synchronized binocular image sequences and four channel audio tracks. The stability of the acquisition device ensures the repeatability of recordings and, hence, the significance of the experiments using the data set. In addition, the scenarios were designed to benchmark algorithms aiming at different applications as described later on. To the best of our knowledge, there is no equivalent publicly available data set in terms of data quality and scenario design.

The remainder of the paper is structured as follows. Section 2 delineates the related existing data sets. Section 3 is devoted to describe the acquisition setup: the recording device, the recording environment and the characteristics of the acquired data. A detailed description of the categories and of the scenarios is given in section 4. Afterward, the data set annotation procedure is discussed (section 5). Before drawing the conclusions (section 7), some examples of usage of the RAVEL data set are given (section 6).

2 Related Data Sets

The RAVEL data set is at the cross-roads of several HRI-related research topics, such as robot vision, audio-visual fusion [19], sound source separation, dialog handling, etc. Hence, there are many public data sets related to RAVEL. These data sets are reviewed in this section and the most relevant ones are described.

Accurate recognition of human actions and gestures is of prime importance in HRI. There are two tasks in performing human actions recognition from visual data: classification of actions and segmentation of actions. There are several available data sets for action recognition. KTH [34], Youtube Action Classification [23] and Hollywood1 [21] are data sets devoted to provide a basis for solving the action classification task. For the segmentation task two data sets are available: Hollywood2 [26] and Coffee and Cigarettes [39]. All these data sets provide *monocular* image sequences. In contrast, the INRIA XMAS data set [38] provides 3D visual hulls and it can be used for the classification and localization tasks. In the INRIA XMAS data set, the actors perform actions in a predefined sequence and are

recorded using a complex multiple camera setup that operates in a specially arranged room. The Opportunity data set [32] serves as a data set for the challenge with the same name. The focus of this challenge is benchmarking the different state-of-the-art action recognition methods. Last, but not least, there are three data sets concerning the daily activities on a "kitchen" scenario namely: the KIT Robo-Kitchen Activity Data Set [33], the University of Rochester Activities of Daily Living Data Set [28] and the TUM Kitchen Data Set [36].

Audio-visual perception [19,18] is an useful skill for any entity willing to interact with human beings, since it provides for a spatio-temporal representation of an event. There are several existing data sets for the AV research community. In particular, a strong effort has been made to produce a variety of multimodal data sets focusing on faces and speech, like the AV-TIMIT [15], GRID [9], M2VTS [31], XM2VTSDB [27], Banca [3], CUAVE [30] or MOBIO [25] data sets. These data sets include individual speakers (AV-TIMIT, GRID, M2VTS, MOBIO, XM2VTSDB, Banca) or both individual speakers and speaker pairs (CUAVE). All have been acquired with one close-range fixed camera and one close-range fixed microphone. Two corpora more closely related to RAVEL are the AV16.3 data set [22] and the CAVA data set [2]. Both include a range of situations. From meeting situations where speakers are seated most of the time, to motion situations, where speakers are moving most of the time. The number of speakers may vary over time. Whilst for the AV16.3 data set three fixed cameras and two fixed 8-microphone circular arrays were used, for the CAVA data set two cameras and two microphones were mounted in a person's head. Instead, RAVEL uses an active robot head equipped with far-range cameras and microphones.

Concerning human robot interaction data sets, [40] provides typical robot sensors' data of a "home tour" scenario annotated using human spatial concepts; this allows to evaluate methods trying to semantically describe the geometry of an indoor scene. In [29], the authors present a new audio-visual corpus containing information of two of the modalities used by humans to communicate their emotional states, namely speech and facial expression in the form of dense dynamic 3D face geometries.

Different data sets used different devices to acquire the data, depending on the purpose. In the next section, the acquisition setup used in RAVEL, which includes the recording environment and device, is fully detailed. Furthermore, the type of recorded data is specified as well as its main properties in terms of synchronization and calibration.

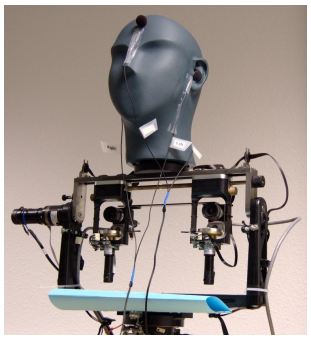


Fig. 2: The POPEYE robot head was used to collect the RAVEL data set. The color-camera pair as well as two (front and left) out of four microphones are shown in the image. Four motors provide the rotational degrees of freedom and ensure the stability of the device and the repeatability of the recordings.

Fig. 3: Two views of the recording environment. The POPEYE robot is in one side of the room. As shown, the sequences were shot with and without daylight providing for lighting variations. Whilst two diffuse lights were included in the setup to provide for good illumination, no devices were used to modify neither the illumination changes nor the sound characteristics of the room. Hence, the recordings are affected by all kind of audio and visual interferences and artifacts present in natural indoor scenes.

3 Acquisition Setup

Since the purpose of the RAVEL data set is to provide data for benchmarking methods and techniques for solving HRI challenges, two requirements have to be addressed by the setup: a robocentric collection of accurate data and a realistic recording environment. In this section the details of this setup are given, showing that the two requisites are satisfied to a large extent. In a first stage the recording device is described. Afterward, the acquisition environment is delineated. Finally, the properties of the acquired data in terms of quality, synchrony and calibration are detailed and discussed.

The POPEYE robot was designed in the framework of the POP project[1]. This robot is equipped with four microphones and two cameras providing for auditory and visual sensory faculties. The four microphones are mounted on a dummy-head, as shown in Figure 2, designed to imitate the filtering properties associated with a real human head. Both cameras and the dummy head are mounted on a four-motor structure that provides for accurate moving capabilities: pan motion, tilt motion and camera vergence.

The POPEYE robot has several remarkable properties. First of all, since the device is alike the human being, it is possible to carry out psycho-physical studies using the data acquired with this device. Secondly, the use of the dummy head and the four microphones, allows for the comparison between using two microphones and the Head Related Transfer Function (HRTF) against using four microphones without HRTF. Also, the stability and accuracy of the motors ensure the repeatability of the experiments. Finally, the use of cameras and microphones gives to the POPEYE robot head audio-visual sensory capabilities in one device that geometrically links all six sensors.

All sequences from the data set were recorded in a regular meeting room, shown in Figure 3. Whilst two diffuse lights were included in the setup to provide for good illumination, no devices were used to modify neither the effects of the sunlight nor the accoustics characteristics of the room. Hence, the recordings are affected by exterior illumination changes, acoustic reverberations, outside noise, and all kind of audio and visual interferences and artifacts present in unconstrained indoor scenes.

For each sequence, we acquired several streams of data distributed in two groups: the *primary* data and the *secondary* data. While the first group is the data acquired using the POPEYE robot's sensors, the second group was acquired by means of devices external to the robot. The *primary* data consists of the audio and video streams captured using POPEYE. Both, left and right, cameras have a resolution of 1024×768 and two operating modes: 8-bit gray-scale images at 30 frames per second (FPS) or 16-bit YUV-color images at 15 FPS. The four Soundman OKM II Classic Solo microphones mounted on the Sennheiser MKE 2002 dummy-head were linked to the computer via the Behringer ADA8000 Ultragain Pro-8 digital external sound card sampling at 48 kHz. The *secondary* data are meant to ease the task of manual annotation for ground-truth. These data consist of one flock of birds (FoB) stream (by Ascension technology) to provide the absolute position of the actor in the scene and up to four wireless close-range microphones PYLE PRO PDWM4400 to capture the audio track of each individual actor.

Both cameras were synchronized by an external trigger controlled by software. The audio-visual synchronization was done by means of a clapping device. This device provides an event that is sharp – and hence, easy to detect – in both audio and video signals. The FoB was synchronized to the visual stream in a similar way: with a sharp event in both FoB and video signals. Regarding the visual calibration, the state-of-the-art method described in [4] uses several image-pairs to provide an accurate calibration. The audio-visual calibration is manually done by annotating the position of the microphones with respect to the cyclopean coordinate frame [12].

Following the arguments presented in the previous paragraphs it can be concluded that the setup suffices conceptual and technical validation. Hence, the sequences have an intrinsic value when used to benchmark algorithm targeting HRI applications. The next section is devoted to fully detail the recorded scenarios forming the Ravel data set.

4 Data Set Description

The Ravel data set has three different categories of scenarios. The first one is devoted to study the recognition of actions performed by a human being. With the second category we aim to study the audio-visual recognition of gestures addressed to the robot. Finally, the third category consists of several scenarios; they are examples of human-human interaction and human-robot interaction. Table 1 summarizes the amount of trials and actors per scenario as well as the size of the visual and auditory data. Figure 1 (a)-(h) shows a snapshot of the different scenarios in the Ravel data set. The categories of scenarios are described in detail in the following subsections.

4.1 Action Recognition [AR]

The task of recognizing human-solo actions is the motivation behind this category; it consists of only one scenario. Twelve actors perform a set of nine actions alone and in front of the robot. There are eight male actors and four female actors. Each actor repeats the set of actions six times in different – random – order,

[1] `http://perception.inrialpes.fr/POP/`

Table 1: Summary of the recorded data size per scenario.

Scenario	Trials	Actors	Video in MB	Audio in MB
AR	12	12	4,899	2,317
RG	11	11	4,825	1,898
AD	6	6	222	173
C	5	4	118	152
CPP	1	1	440	200
MS	7	6	319	361
IP	5	7	327	204
Total	–	–	11,141	5,305

which was prompted in two screens to guide the actor. This provides for various co-articulation effects between subsequent actions. The following is a detailed list of the set of actions: (i) *stand still*, (ii) *walk*, (iii) *turn around*, (iv) *clap*, (v) *talk on the phone*, (vi) *drink*, (vii) *check watch* (analogy in [38]), (viii) *scratch head* (analogy in [38]) and (ix) *cross arms* (analogy in [38]).

4.2 Robot Gestures [RG]

Learning to identify different gestures addressed to the robot is another challenge in HRI. Examples of such gestures are: waving, pointing, approaching the robot, etc. This category consists of one scenario in which the actor performs six times the following set of nine gestures: (i) *wave*, (ii) *walk towards the robot*, (iii) *walk away from the robot*, (iv) *gesture for 'stop'*, (v) *gesture to 'turn around'*, (vi) *gesture for 'come here'*, (vii) *point action*, (viii) *head motion for 'yes'* and (ix) *head motion for 'no'*. In all cases, the action is accompanied by some speech corresponding to the gesture. In total, eleven actors (nine male and two female) participated in the recordings. Different English accents are present in the audio tracks which makes the speech processing challenging.

4.3 Interaction

This category contains the most interactive part of the data set, i.e. human-human as well as human-robot interaction. Each scenario consists of a natural scene in which several human beings interact with each other and with the robot. In some cases one of the actors and/or the robot act as a passive observer. This category contains six different scenarios detailed in the following. In all cases, a person emulated the robot's behavior.

Actor	(enters the scene)
Actor	Excuse me, where are the toilets?
Robot	Gentleman/Ladies are to the left/right and straight on 10 meters.
Actor	(leaves the scene)

Script 1: The script encloses the text spoken by the actor as well as by the robot in the "*Asking for directions*" scenario.

Asking for Directions [AD]

In this scenario an actor asks the robot for directions to the toilets. The robot recognizes the question, performs gender identification and gives the actor the right directions to the appropriate toilets. Six different trials (four male and two female) were performed. The transcript of this scenario is in Script 1.

Chatting [C]

We designed this scenario to study the robot as a passive observer in a dialog. The scenario consists of two people coming into the scene and chatting for some undetermined time, before leaving. There is no fixed script – occasionally two actors speak simultaneously – and the sequences contain several actions, e.g. hand shaking, cheering, etc. Five different trials were recorded.

Cocktail Party Problem [CPP]

Reviewed in [14], the Cocktail Party Problem has been matter of study for more than fifty years (see [8]). In this scenario we simulated the cocktail party effect: five actors freely interact with each other, move around, appear/disappear from the camera field of view, occlude each other and speak. There is also background music and outdoor noise. In summary, this is one of the most challenging scenarios in terms of audio-visual scene analysis, action recognition, speech recognition, dialog engaging and annotation. In the second half of the sequence the robot performs some movements. Figure 4 is a frame of the (left camera of the) CPP scenario. Notice the complexity of the scene in terms of number of people involved, dialog engagement, etc.

Where Is Mr. Smith? [MS]

The scenario was designed to test skills such as face recognition, speech recognition and continuous dialog. An actor comes into the scene and asks for Mr. Smith. The robot forwards the actor to Mr. Smith's office.

Fig. 4: A frame of the CPP sequence representative of the complexity of this scenario.

Actor	(enters and positions him in front of the robot)
Actor	I am looking for Mr. Smith?
Robot	Yes Sir, Mr. Smith is in Room No. 22
Actor	(leaves the scene)
Mr. Smith	(enters the scene)
Mr. Smith	Hello Robot.
Robot	Hello Mr. Smith.
Robot	How can I help you?
Mr. Smith	Haven't you seen somebody looking for me?
Robot	Yes, there was a gentleman looking for you 10 minutes ago.
Mr. Smith	Thank you Bye.
Robot	You are welcome.
Mr. Smith	(leaves the scene)

Script 2: Detail of the text spoken by both actors (Actor and Mr. Smith) as well as the Robot in the "*Where is Mr. Smith?*" scenario.

However, he is not there and when he arrives, he asks the robot if someone was looking for him. The robot replies according to what happened. The transcript for the scenario is in Script 2. Seven trials (five male and two female) were recorded to provide for gender variability.

Introducing People [IP]

This scenario involves a robot interacting with three people in the scene. There are two versions of this scenario: passive and active. In the passive version the camera is static, while in the active version the camera is moving to look directly at speakers' face. Together with the *Cocktail Party Problem* scenario, they are the only exception where the robot is not static in this data set.

Actor 1	(enters room, positions himself in front of robot and looks at robot)
Actor 1	Hello, I'm Actor 1.
Robot	Hello, I'm Nao. Nice to meet you.
Actor 2	(enters room, positions himself next to Actor 1 and looks at robot)
Robot	Excuse me for a moment.
Robot	Hello, I'm currently talking to Actor 1. Do you know Actor 1?
Actor 2	No, I don't know him.
Robot	Then let me introduce you two. What is your name?
Actor 2	Actor 2
Robot	Actor 2, this is Actor 1. Actor 1 this is Actor 2.
Actor 3	(enters room, positions himself next to Actor 1, looks at Actor 1and Actor 2)
Actor 3	Actor 1 and Actor 2, have you seen Actor 4?
Actor 2	No I'm sorry, we haven't seen her.
Actor 3	Ok, thanks. I'll have to find her myself then. Bye.
Actor 3	(leaves)
Actor 2	Actor 1, (turn heads towards robot)
Actor 1	We have to go too. Bye
Robot	Ok. See you later.

Script 3: Detail of the script of the scenario "*Introducing people - Passive*". The three people interact with the robot. The robot is static in this scenario.

In the passive version of the scenario, Actor 1 and Actor 2 interact together with the Robot and each other; Actor 3: only interacts with Actor 1 and Actor 2. The transcript of the passive version is in Script 3. In the active version, Actor 1 and Actor 2 interact with the Robot and each other; Actor 3 enters and leaves room, walking somewhere behind Actor 1 and Actor 2, not looking at the Robot. The transcript of the active version is detailed in Script 4

4.4 Background Clutter

Since the RAVEL data set aims to be useful for benchmarking methods working in populated spaces, the first two categories of the data set, action recognition and robot gestures, were collected with two levels of background clutter. The first level corresponds to a controlled scenario in which there are no other actors in the scene and the outdoor and indoor acoustic noise is very limited. During the recording of the scenarios under the second level of background clutter, other actors were allowed to walk around, always behind the main actor. In addition, the extra actors occasionally talked to each other; the amount of outdoor noise was not limited in this case.

Actor 1	(enters room, positions himself in front of robot and looks at robot)
Actor 1	Hello, I'm Actor 1.
Robot	Hello, I'm Nao. Nice to meet you.
Actor 2	(enters room, positions himself next to Actor 1 and looks at robot)
Robot	Excuse me for a moment.
Robot	(turns head towards Actor 2)
Actor 1	(turns head towards Actor 2)
Robot	Hello, I'm currently talking to Actor 1. Do you know Actor 1?
Actor 2	No, I don't know him.
Robot	Then let me introduce you two. What is your name?
Actor 2	Actor 2
Robot	Actor 2 this is Actor 1. (turns head towards Actor 1) Actor 1 this is Actor 2.
Actor 3	(enters room, walks somewhere behind Actor 1 and Actor 2, leaves room)
Actor 1	We have to go now. Bye
Robot	(turns head towards Actor 1)
Robot	Ok. See you later.

Script 4: Detail of the script of the scenario "*Introducing people - Active*". Two out of the three people interact with the robot. The latter is a moving robot.

4.5 Data Download

The RAVEL data set is publicly available at `http://ravel.humavips.eu/` where a general description of the acquisition setup, of the data, and of the scenarios can be found. In addition to the links to the data files, we provide previews for all the recorded sequences for easy browsing previous to data downloading.

5 Data Set Annotation

Providing the ground truth is an important task when delivering a new data set; this allows to quantitatively compare the algorithms and techniques using the data. In this section we present two types of annotation data provided together with the data set.

5.1 Action Performed

The first kind of annotation we provided is related to the action and robot gesture scenarios of the data set. This annotation is done using a classical convention, that each frame is assigned a label of the particular action. Since the played action is known only one label is assigned to each frame. Because the annotation we need is not complex a simple annotation tool was designed for this purpose in which a user labels each start and end of each action/gesture in the recordings. The output of that tool is written in the standard ELAN [6] annotation format. A screen shot of the annotation tool is shown in Figure 5.

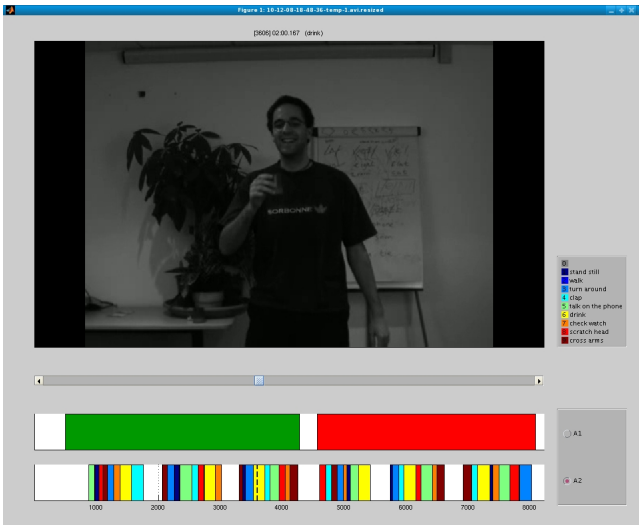


Fig. 5: The annotation tool screen shot. Two time lines are shown below the image. The first one (top) is used to annotate the level of background clutter. The second one (bottom) details which action is performed at each frame.

5.2 Position and Speaking State

The second kind of annotations concern the interaction part of the data set and consists on the position of the actors (both in the images and in the 3D space) and on the speaking state of the actors. In both cases the annotator uses a semi-automatic tool that outputs an ELAN-readable output file. The semi-automatic procedures used are described in the following.

Regarding the annotation of the actors' position, the tracking algorithm described in [17] is used to semi-automatize the process. The annotator is asked for the object's bounding box, which is then tracked along time. At any point, the annotator can reinitialize the tracker to correct its mistakes. Once the object is tracked along the entire left camera image sequence, the correspondent trajectory in the other image is automatically estimated. To do that, the classical approach of maximizing the normalized cross-correlation across the epipolar constraint is used [13]. From these correspondence pairs, the 3D location is computed at every frame using the DLT reconstruction procedure [13]. The location of the speaker in the images is given in pixels and the position in the 3D space are given in millimeters with respect to the cyclopean coordinate reference frame [12].

Concerning the speaking state, the state-of-the-art voice activity detector described in [5] is used on the per-actor close range microphones. In a second step, the annotator is in charge of discarding all false positives generated by the VAD, leading to a clean speaking state annotation per each actor.

6 Data Exploitation Examples

In order to prove the importance of the RAVEL data set, a few data exploitation examples are provided. Three different examples, showing how diverse applications can use the presented data set, are explained in this section: a scene flow extraction method, an event-detection algorithm based on statistical audio-visual fusion techniques and two machine learning-based action recognition methods.

6.1 Scene Flow

Since the entire data set is captured by synchronized and fully calibrated cameras, it is possible to compute a 3D scene flow [37], which is a classical low-level computer vision problem. The 3D scene flow is defined as a motion field such that each reconstructed pixel for a frame has assigned a 3D position and a 3D velocity. It leads to an image correspondence problem, where one has to simultaneously find corresponding pixels between images of a stereo pair and corresponding pixels between subsequent frames.

After the 3D reconstruction using the known camera calibration, these correspondences fully determine the 3D scene flow. A projection of a scene flow is shown in Figure 6, as a disparity (or depth) map and horizontal and vertical optical flow maps. These results are computed using a recent seed growing algorithm [7]. The scene flow results can be used for further processing towards the understanding of a dynamic scene.

6.2 Audio-Visual Event Detection

How to detect audio-visual events, i.e. events that are both heard and seen, is a topic of interest for researchers working in multimodal fusion. An entire pipeline – from the raw data to the concept of AV event – is exemplified in this section. This pipeline consists of three modules: visual processing, auditory processing and audio-visual fusion. In the following, the method is roughly described; interested readers can find a more detailed explanation in [1].

The task is to extract audio-visual events from the synchronized raw data; these are a sequence of stereo-image pairs and the corresponding binaural audio stream. The method looks for events of interest in the visual and auditory domain, that is events that can be both seen and heard at the same time. Thus, there is a need for (i) choosing the right features to extract from the raw data, (ii) linking the visual and auditory modalities, (iii) accounting for feature noise and outliers and (iv) selecting the right number of events.

To extract visual features, Harris interest points are computed and filtered to keep those image locations related to motion. Stereo-matching is performed to later on reconstruct the points in the 3D space. This provides us for a set of points related to motion, $\{\boldsymbol{f}_m\}_m$, which are assumed to be around the locations of interest. The audio features are the so called Interaural Time Differences (ITD), measuring the different of time arrival between the two microphones. These values approximate the direction of the active sound sources, and they are denoted by $\{g_k\}_k$.

In order to link the two modalities, the geometrical properties of the recording device, i.e. the microphone positions in the 3D space ($\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^3$), are used to map the 3D points into the ITD space. More precisely, this model assumes that a sound source placed at $\boldsymbol{S} \in \mathbb{R}^3$ produces ITD values around the point given by:

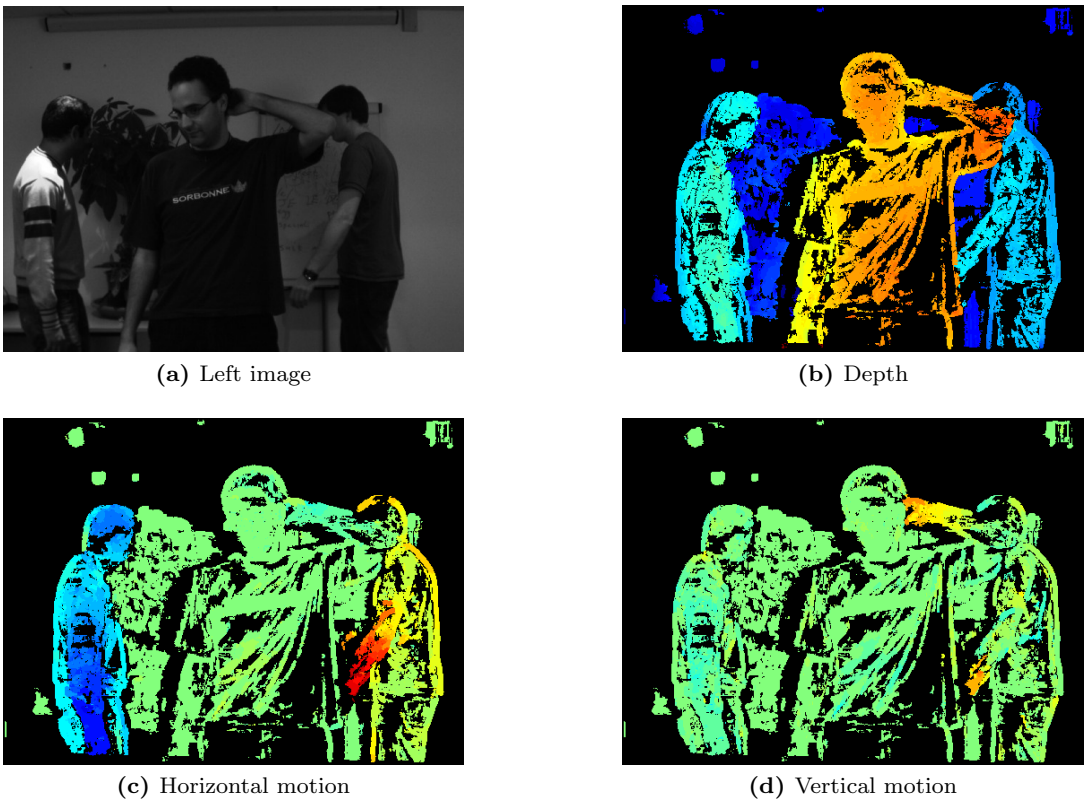
$$\mathrm{ITD}(\boldsymbol{S}) = \frac{\|\boldsymbol{S} - \boldsymbol{M}_1\| - \|\boldsymbol{S} - \boldsymbol{M}_2\|}{c},$$

where c is the sound speed. Since this is defined for any position in the space, we can project the visual features to the ITD space, hence defining: $\tilde{f}_m = \mathrm{ITD}(\boldsymbol{f}_m)$. Thus linking the audio and visual modalities.

Once all the feature points lie in the same space, we run a modified version of the EM algorithm to fuse the two types of data. This EM algorithm uses the mapped visual features to supervise the fusion of audio features into visual clusters. The probabilistic model fit by the EM is a mixture model with the following probability density:

$$\mathrm{p}(x) = \sum_{n=1}^{N} \pi_n \mathcal{N}(x|\mu_n, \sigma_n) + \pi_{N+1}\mathcal{U}(x).$$

In this mixture, each of the Gaussian components represents an audio-visual event. The variance of the components will account for the noise and the uniform component accounts for the outliers due both, the ITD computation and the projection of the 3D features. Given



(a) Left image **(b)** Depth

(c) Horizontal motion **(d)** Vertical motion

Fig. 6: Results of the scene flow algorithm. The original left image is shown in (a). The actual results are color coded. For depth map (b), warmer colors are closer to the camera. For horizontal (c) and vertical (d) motion maps, green color stands for zero motion, while colder colors correspond to right and up motion respectively, warmer colors the opposite direction. Black color stands for unassigned disparity or optical flow.

the number of AV events N, the EM algorithm will estimate the optimal set of parameters $\{\pi_n, \mu_n, \sigma_n\}_n$ in the maximum likelihood sense.

The last step of the algorithm consists on choose the right N. In order to do that, the authors rely on the statistically consistent model selection criterion called Bayesian Information Criterion (BIC). The main idea is that different models are penalized depending on the number of free parameters; the higher the number, the higher the penalization. The optimal model in the BIC sense is chosen. Finally, these corresponding clusters are back-projected to the 3D space providing localization of audio-visual events.

The algorithm was applied onto the *CPP* sequence of the RAVEL data set. Figure 7 shows the results of the method in nine frames of the sequence. In this sequence the AV events are people in an informal social gathering. Although the method has some false positives, it correctly detects and localizes 26 objects out of 33 (78.8%).

6.3 Action Recognition

To demonstrate some of the potentialities of the RAVEL data set we establish a baseline for the Action Recognition subset of RAVEL. In this section we show the performance of the state-of-the-art methods when applied to the RAVEL data set. The results are split depending on the application: either "isolated" recognition or "continuous" recognition.

6.3.1 Isolated Recognition

Among all the different methods to perform isolated action recognition, we decide to use the one described in [20]. Its performance is comparable to the state-of-the-art methods and binaries can be easily found and downloaded from here[2].

This method represents an action as a histogram of visual words. Once all the actions are represented, a Support Vector Machine (SVM) is used to learn each

[2] http://www.irisa.fr/vista/Equipe/People/Ivan.Laptev.html

Fig. 7: A sample of the AV events detected in the *CPP* sequence of the RAVEL data set. The ellipses correspond to the localization of the events in the image plane. The method correctly detects and localizes 26 objects out of 33 (78.8%).

class (action) to afterwards determine if an unknown action belongs to one of the classes previously trained. This methodology is known as a Bag-of-Words (BoW) and it can be summarized into four steps.

1. Collect set of features for all actions/actors for each video clip.
2. Apply clustering algorithm to these features, for instance, k-means.
3. Apply 1-NN to classify the features of each action into the centroids found by k-means, and obtain an histogram of k bins.
4. Train a classifier with these histograms, for instance, SVM.

In this experiment the *Laptev* features are used. These features correspond to a set of spatiotemporal Harris detected points described by a concatenation of Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) descriptors [10]. The clustering method to select the k most representative features is k-means. Such features allow us to represent each action as histograms of visual words.

Table 2: Results of the experiments on isolated action recognition. Recognition rates of the *Laptev* features using different number of clusters for the k-means algorithm for the controlled and normal levels of background clutter.

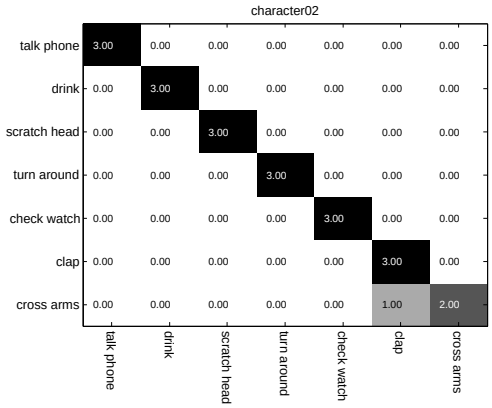
k	500	1000	2000	3000
Controlled	0.6320	0.6883	0.6926	0.6797
Normal	0.4892	0.4675	0.4762	0.5281

Finally a linear multiclass SVM is trained with the histograms. Contrary to the paradigm the multiclass SVM was designed for, we do not have a huge amount of positive and negative examples, just 12 actors. To overcome this issue, a leave-one-out cross-validation strategy is applied.

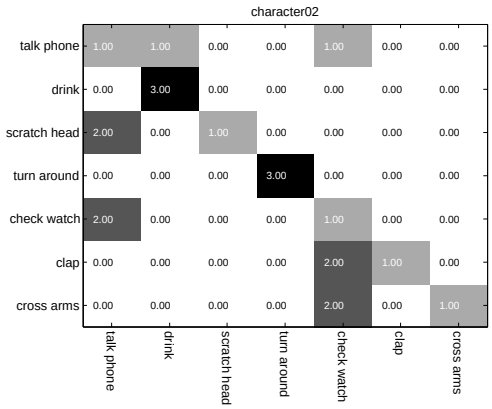
Due to the large amount of data, the computation of the k-means algorithm becomes prohibitive. That is why the algorithm is applied to the set of features corresponding to one actor. Of course, these features are not used neither for training nor for testing.

To have some quantitative evaluation, we perform different experiments varying the number of clusters, k. Table 2 shows the recognition rate, that is defined as the number of times that the actions have been classified correctly over the total number of attempts that have been done to classify the actions.

In addition to the recognition rates we show several confusion matrices. The ij position of a confusion matrix represents the amount of instance of the i category classified as the j category. Figures 8, 9 and 10 show the confusion matrices when the characters 2, 3 and 11 were tested. The matrices on the top correspond to the scenarios under controlled background clutter and the matrices on the bottom to the scenarios under normal background clutter. We can observe the expected behavior: the matrices under the controlled conditions report much better results than those under normal conditions. In addition, we observe some variation across different actors on where are the wrong detections. This is caused by two effects: the different ways of performing the actions and the various co-articulations. All together justifies the use of a cross-validation evaluation strategy. Finally, Figure 11 reports the global confusion matrices, from which we can observe that the expected behavior regarding the performance on controlled vs. normal clutter level, observed before is extensible to the entire data set.

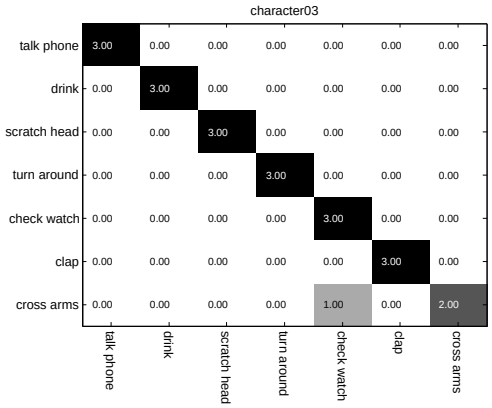


(a) Character 2 - Controlled

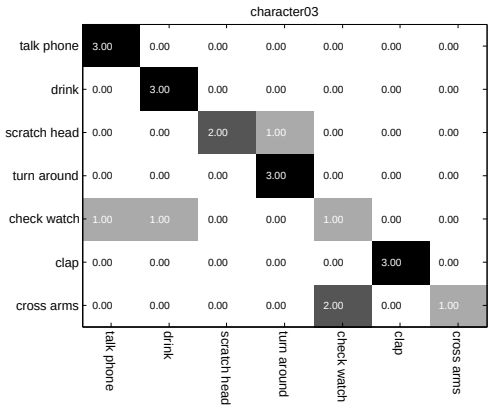


(b) Character 2 - Normal

Fig. 8: Confusion matrices for the 2-nd actor with $k = 2000$ clusters and *Laptev* features. (a) Controlled background clutter. (b) Normal background clutter.



(a) Character 3 - Controlled



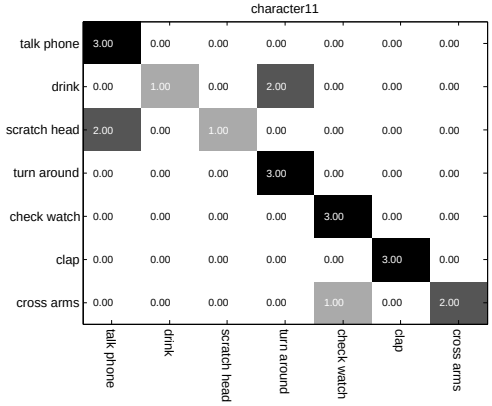
(b) Character 3 - Normal

Fig. 9: Confusion matrices for the 3-rd actor under the same conditions as Figure 8.

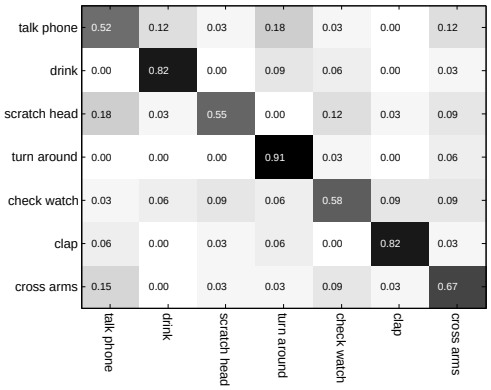
6.3.2 Continuous Recognition

Continuous action recognition, or joint segmentation and classification, refers to the case where a video to be analyzed contains a *sequence* of actions played by one actor or by different actors. The order of the actions in the video is not known. Most of the earlier methods assume that the segmentation and classification tasks may be carried out completely independently of each other, i.e., they consider an isolated recognition scenario where the boundaries of action in videos are known a priori. In continuous recognition scenarios the objective is to find the best label sequence for each video frame. The isolated recognition framework of representing an action as a single histogram of visual words can be modified to perform continuous action recognition. In [35], the authors propose a method which uses SVM for classification of each action and the temporal segmentation is done efficiently using dynamic programming. Multiclass SVMs are trained on a segmented training set. In the classification stage, actions are searched over several temporal segments at different time scales. Each temporal segment is represented by a single histogram. The search over the time scale is restricted by the maximum and minimum lengths of actions computed from the training set. Each of these segments are classified by SVM trained on the action classes. This classification yields ordered sets of labels for the segments. To find the best set of labels for the whole video one needs an optimization criteria. In [35] the optimization criteria is to maximize the sum of the SVM classifier scores computed by concatenating segments over different temporal scales. This optimization is efficiently cast in the dynamic programming framework.

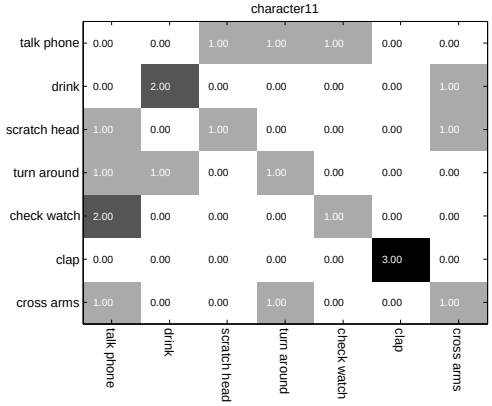
Both [35] and [16] are similar in the way they perform continuous action recognition, i.e., the classification is done at different temporal scales using SVMs,



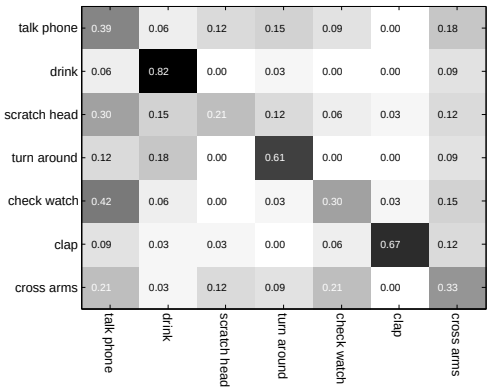
(a) Character 11 - Controlled



(a) All characters - Controlled



(b) Character 11 - Normal



(b) All characters - Normal

Fig. 10: Confusion matrices for the 11-th actor under the same conditions as Figure 8.

Fig. 11: Global confusion matrices under the same conditions as Figure 8.

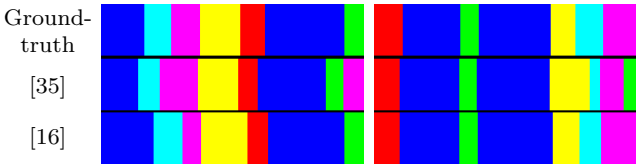
while the segmentation is efficiently done using dynamic programming. The crucial difference between these two methods is the optimization criteria used for dynamic programming. In [35], the sum of the SVM scores for the concatenated segments is maximized. This ensures the best sequence of labels for the whole video but does not ensure that the best label is assigned to each segment. This problem is overcome in [16] where a difference between the SVM score of the winning class label for a segment and the next best label is computed. The sum of these differences computed for each segment is then maximized over concatenated segments at different time scales over the whole video. This optimization is also cast in the dynamic programming framework.

Results on the Ravel dataset using the state-of-art continuous recognition algorithms [16,35] are shown in Table 3. The accuracy of the algorithms were measured by percentage of correctly labeled frames. The recognition accuracy is also computed on Weizmann [11] and Hollywood datasets [21]. Since these dataset contain isolated actions only, we created a sequence of multiple actions by concatenating single-action clips following the protocol of [16]. This concatenation creates abrupt artificial inter-action transitions. In contrast, the Ravel dataset is recorded continuously in one shot per actor. The actions are played by the actors in random order (given by a hidden prompt) and moreover we did not instruct the actors to come to a rest position after every action. Therefore this dataset is well suitable for the the continuous action recognition.

In figure 12 we show the estimated labels of two video sequences by [16,35] in a comparison with the ground-truth segmentation.

Table 3: Accuracy of the continuous recognition methods using artificially merged actions (Weizmann and Hollywood) and actions involving smooth transitions (RAVEL).

Dataset:	Weizmann	Hollywood	RAVEL
Shi et. al. [35]	69.7	34.2	55.4
Hoai et. al. [16]	87.7	42.2	59.9



(a) actor 1 **(b)** actor 4

Fig. 12: Continuous action recognition results on the RAVEL datasets. Colors encodes action labels of frames of the video sequences. Top row shows ground-truth labeling, while two rows below show results of two state-of-the-art algorithms [35, 16]. The results are shown for two selected actors.

7 Conclusions

The RAVEL data set consists of multimodal (visual and audio) multichannel (two cameras and four microphones) synchronized data sequences. The data set embodies several scenarios designed to study different HRI applications. This new audiovisual corpus is important for two main reasons. On one hand, the stability and characteristics of the acquisition device ensure the quality of the recorded data and the repeatability of the experiments. On the other hand, the amount of data is enough to evaluate the relevance of the contents in order to improve the design of future HRI systems.

The acquisition setup (environment and device) was fully detailed. Technical specifications of the recorded streams (data) were provided. The calibration and synchronization procedures, both visual and audio-visual, were described. The scenarios were detailed; their scripts were provided when applicable. The recorded scenarios fall in three categories representing different groups of applications: action recognition, robot gesture and interaction. Furthermore, the data set annotation method was also described. Finally, three examples of data exploitation were provided: scene flow extraction, audio-visual event detection and action recognition. These prove the usability of the RAVEL data set.

References

1. X Alameda-Pineda, V Khalidov, R Horaud, F Forbes: Finding audio-visual events in informal social gatherings. In: Proceedings of the ACM/IEEE International Conference on Multimodal Interaction (2011)
2. E Arnaud, H Christensen, Y.-C Lu, J Barker, V Khalidov, M Hansard, B Holveck, H Mathieu, R Narasimha, E Taillant, F Forbes, R. P Horaud: The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements. In: Proceedings of the ACM/IEEE International Conference on Multimodal Interfaces (2008). `http://perception.inrialpes.fr/CAVA_Dataset/`
3. E Bailly-Baillire, S Bengio, F Bimbot, M Hamouz, J Kittler, J Mariéthoz, J Matas, K Messer, F Porée, B Ruiz: The BANCA database and evaluation protocol. In: Proceedings of the International Conference on Audio and Video-Based Biometric Person Authentication, pp. 625–638. Springer-Verlag (2003)
4. J.-Y Bouguet: Camera calibration toolbox for Matlab (2008). `http://www.vision.caltech.edu/bouguetj/calib\_doc/`
5. M Brookes: Voicebox: Speech processing toolbox for matlab. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`
6. H Brugman, A Russel, X Nijmegen: Annotating multimedia / multimodal resources with ELAN. In: Proceedings of the International Conference on Language Resources and Evaluation, pp. 2065–2068 (2004)
7. J Cech, J Sanchez-Riera, R. P Horaud: Scene flow estimation by growing correspondence seeds. In: Computer Vision and Pattern Recognition, In the Proceedings of the IEEE International Conference on (2011)
8. E. C Cherry: Some experiments on the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America **25**(5), 975–979 (1953)
9. M Cooke, J Barker, S Cunningham, X Shao: An audio-visual corpus for speech perception and automatic speech recognition (l). Speech Communication **49**(5), 384–401 (2007)
10. N Dalal, B Triggs: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2005)
11. L Gorelick, M Blank, E Shechtman, M Irani, R Basri: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12), 2247–2253 (2007). `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`
12. M Hansard, R. P Horaud: Cyclopean geometry of binocular vision. Journal of the Optical Society of America **25**(9), 23572369 (2008)
13. R. I Hartley, A Zisserman: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)
14. S Haykin, Z Chen: The cocktail party problem. Journal on Neural Compututation **17**, 1875–1902 (2005)
15. T. J Hazen, K Saenko, C.-H La, J. R Glass: A segment-based audio-visual speech recognizer: data collection, development, and initial experiments. In: Proceedings of the ACM International Conference on Multimodal Interfaces, ICMI '04, pp. 235–242. ACM (2004)
16. M Hoai, Z Zhong Lan, F De la Torre: Joint segmentation and classification of human actions in video. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2011)
17. Z Kalal, K Mikolajczyk, J Matas: Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(7), 1409 –1422 (2012)
18. V Khalidov, F Forbes, R Horaud: Conjugate mixture models for clustering multimodal data. Journal on Neural Computation **23**(2), 517–557 (2011)

19. H Don Kim, J suk Choi, M Kim: Human-robot interaction in real environments by audio-visual integration. International Journal of Control, Automation and Systems **5**(1), 61–69 (2007)
20. I Laptev: On space-time interest points. International Journal of Computer Vision **64**(2-3), 107–123 (2005)
21. I Laptev, M Marszalek, C Schmid, B Rozenfeld: Learning realistic human actions from movies. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2008)
22. G Lathoud, J Odobez, D Gatica-Pérez: AV16.3: an audio-visual corpus for speaker localization and tracking. In: Proceedings of the Workshop on Machine Learning and Multimodal Interaction. Springer Verlag (2005)
23. J Liu, J Luo, M Shah: Recognizing realistic actions from videos "in the wild". Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2009)
24. R Luo, M Kay: Multisensor integration and fusion in intelligent systems. IEEE Transactions on Systems, Man and Cybernetics **19**(5), 901 –931 (1989)
25. S Marcel, C McCool, P Matejka, T Ahonen, J Cernocky: Mobile biometry (MOBIO) face and speaker verification evaluation. Idiap-RR Idiap-RR-09-2010, Idiap (2010)
26. M Marszalek, I Laptev, C Schmid: Actions in context. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recogntion (2009)
27. K Messer, J Matas, J Kittler, K Jonsson: XM2VTSDB: The extended M2VTS database. In: Proceedings of the International Conference on Audio and Video-based Biometric Person Authentication, pp. 72–77 (1999)
28. R Messing, C Pal, H Kautz: Activity recognition using the velocity histories of tracked keypoints. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE Computer Society, Washington, DC, USA (2009)
29. Y Mohammad, Y Xu, K Matsumura, T Nishida: The H3R explanation corpus human-human and base human-robot interaction dataset. In: International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp. 201 –206 (2008)
30. E. K Patterson, S Gurbuz, Z Tufekci, J. N Gowdy: CUAVE: A new audio-visual database for multimodal human-computer interface research. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, pp. 2017–2020 (2002)
31. S Pigeon: M2vts database (1996). `http://www.tele.ucl.ac.be/PROJECTS/M2VTS/`
32. T. O Project: (2011). `http://www.opportunity-project.eu/`
33. L Rybok, S Friedberger, U. D Hanebeck, R Stiefelhagen: The KIT Robo-Kitchen Data set for the Evaluation of View-based Activity Recognition Systems. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots (2011)
34. C Schüldt, I Laptev, B Caputo: Recognizing human actions: A local svm approach. In: Proceedings of the International Conference on Pattern Recognition, pp. 32–36 (2004)
35. Q Shi, L Wang, L Cheng, A Smola: Discriminative human action segmentation and recognition using SMMs. International Journal on Computer Vision **93**(1), 22–32 (2011)
36. M Tenorth, J Bandouch, M Beetz: The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In: Proceedings of the IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences in conjunction with the International Conference on Computer Vision (2009)
37. S Vedula, S Baker, P Rander, R Collins, T Kanade: Three-dimensional scene flow. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(3) (2005)
38. D Weinland, R Ronfard, E Boyer: Free viewpoint action recognition using motion history volumes. Journal on Computer Vision and Image Understanding **104**(2), 249–257 (2006). `http://4drepository.inrialpes.fr/public/viewgroup/6`
39. G Willems, J. H Becker, T Tuytelaars: Exemplar-based action recognition in video. In: Proceedings of the British Machine Vision Conference (2009)
40. Z Zivkovic, O Booij, B Krose, E Topp, H Christensen: From sensors to human spatial concepts: An annotated data set. IEEE Transactions on Robotics **24**(2), 501 –505 (2008)