MAHNOB-HCI-TAGGING DATABASE

JEROEN LICHTENAUER, MOHAMMAD SOLEYMANI

Abstract. This is a manual to help the users to use the recorded video, audio, eye-gaze and physiological data in response to emotion-eliciting video clips as well as with respect to perceived appropriateness of media tags.

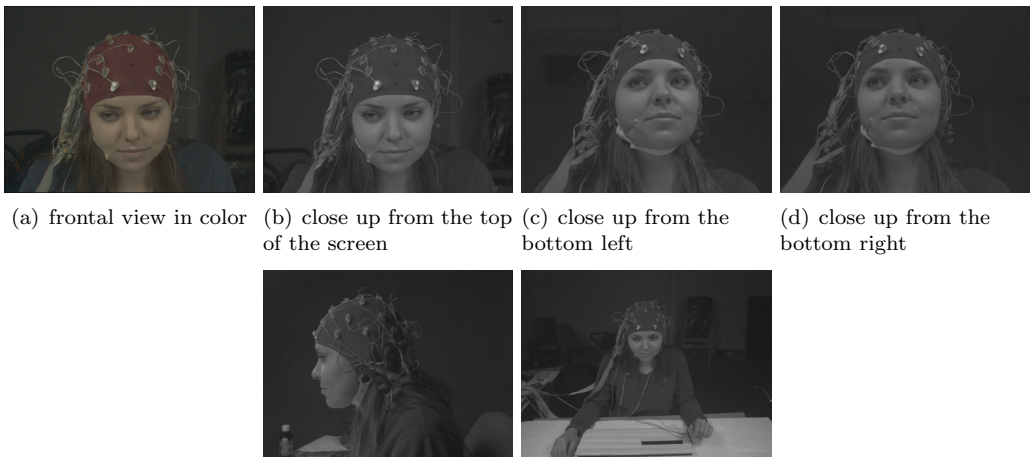
1. Introduction

The experiments from which this database contains the recordings, were conducted with the aim of gaining knowledge about natural behaviour of healthy adults, in interaction with a computer during multimedia watching, designed to elicit affective reactions to the content like amusement or revulsion, and/or the participant's agreement or disagreement with the provided content.

During the experiment, the participant's behaviour is recorded using cameras, microphone, and a gaze tracker. Moreover, the physiological responses of the participant are recorded using a Biosemi Active II system. The Biosemi active II system has been used by many research laboratories around the world. The system is connected to the rest of the system using the fibre optic (galvanic isolation) and the electricity is provided by a battery. Before the experiment, the physiological signals sensors including electro-encephalogram (EEG) sensors using a head-cap, electrocardiogram (ECG) sensors, galvanic skin resistance (GSR) sensors on the fingers, skin temperature sensor, and a respiration belt around chest are attached to the participant's body and the participant is asked to calibrate the gaze tracker by following red circles on the screen. The experiment was controlled by the Tobii studio software (http://www.tobii.com). A photograph of the experimental setup is shown in figure 1.

Figure 1. In the experimental setup, six video cameras were recording facial expressions. The modified keyboard is visible in front of the participant.

London, 2011.



(a) frontal view in color (b) close up from the top of the screen (c) close up from the bottom left (d) close up from the bottom right

(e) profile view (f) wide angle from above the screen

FIGURE 2. Snapshots of videos captured from 6 cameras recording facial expressions and head pose.

30 participants with different cultural and education backgrounds volunteered to participate in response to a campus wide call for volunteers at Imperial College, London. Out of the 30 young healthy adult participants, 17 were female and 13 were male; ages varied between 19 to 40 years old (M = 26.06 SD = 4.39). Participants had different educational background from undergraduate students to post-docs with different English proficiency from intermediate to native speakers. Please, refer to the terms of use, in section 5, for the conditions of using the dataset. Details about the experiment protocol can be found in [3].

2. Description of Recorded Measurement Data

2.1. **Audio Channels.** The audio contains two channels of audio. Channel one (or 'left', if interpreted as a stereo stream) contains the audio signal from a AKG C 1000 S MkIII room microphone, which includes the room noise as well as the sound of the video stimuli. Channel two contains the audio signal from a AKG HC 577 L head-worn microphone. The contents of channel one can be used to reduce the influence of ambient sounds on the processing of any verbal cues. Because of the passive nature of the experiments, the number of verbal cues present in the data is low.

2.2. **Camera Views.** An examples of all camera views are shown in figure 2. The cameras are named as:

- camera 1 = C1 trigger; above the screen, colour, fig. 2 (a)
- camera 2 = BW1; above the screen, monochrome, fig. 2 (b)
- camera 3 = BW2; below the screen, monochrome, fig. 2 (c)
- camera 4 = BW3; below the screen, monochrome, fig. 2 (d)
- camera 5 = BW4; profile view, monochrome, fig. 2 (e)
- camera 6 = BW5; overview from high angle, monochrome, fig. 2 (f)

Two types of cameras have been used in the recordings: One Allied Vision Stingray F-046C, colour camera (C1) and five Allied Vision Stingray F-046B, monochrome cameras (BW1 to BW5). All with a spatial resolution of 780x580 pixels. The cameras were intrinsically and extrinsically calibrated. The extrinsic calibration is shown in figures 3, 4, 5, 6. Calibration parameters are given in tables 1, 2.

2.3. **Physiological Measurements.** The Biosemi active II system (http://www.biosemi.com) with active electrodes was used for acquisition of physiological signals. Physiological signals including ECG, EEG (32 channels), respiration amplitude, and skin temperature were recorded while the videos were shown to the participants. The physiological signals are stored using Biosemi data format (BDF) which is readable by

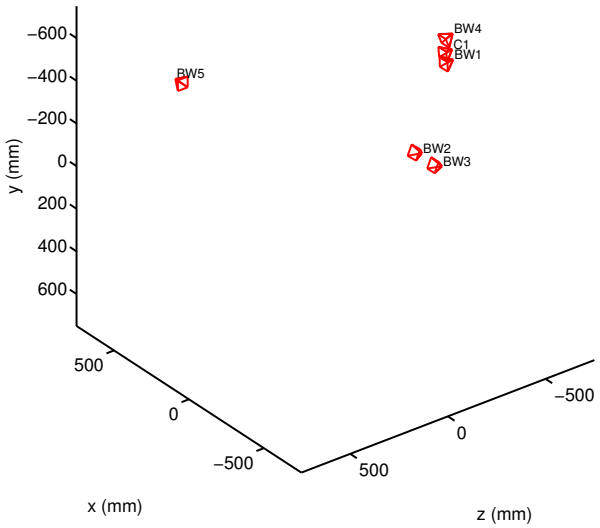


FIGURE 3. Extrinsic camera poses.

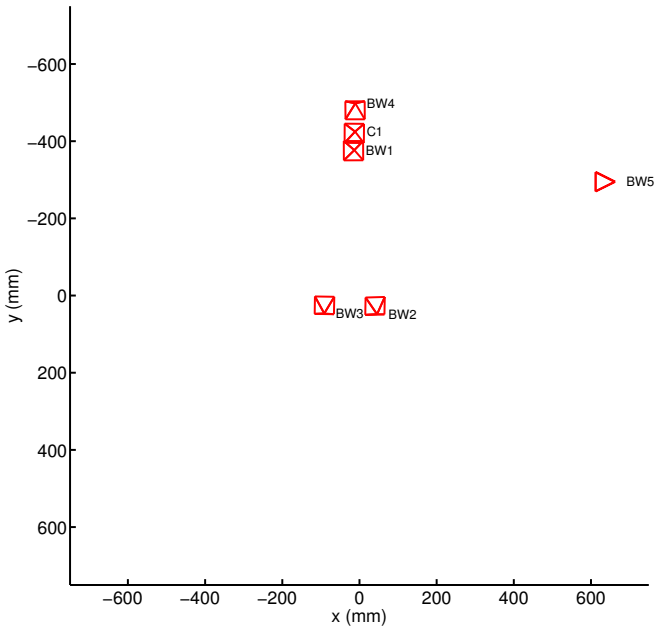


FIGURE 4. Extrinsic camera poses.

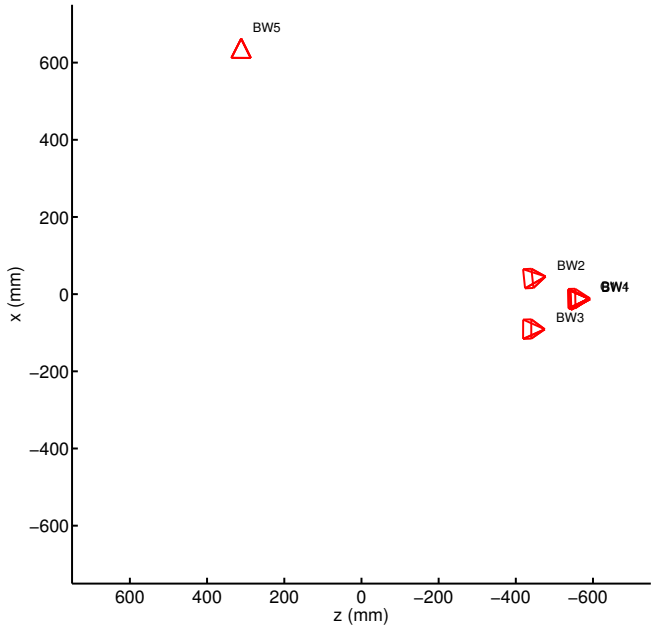


FIGURE 5. Extrinsic camera poses.

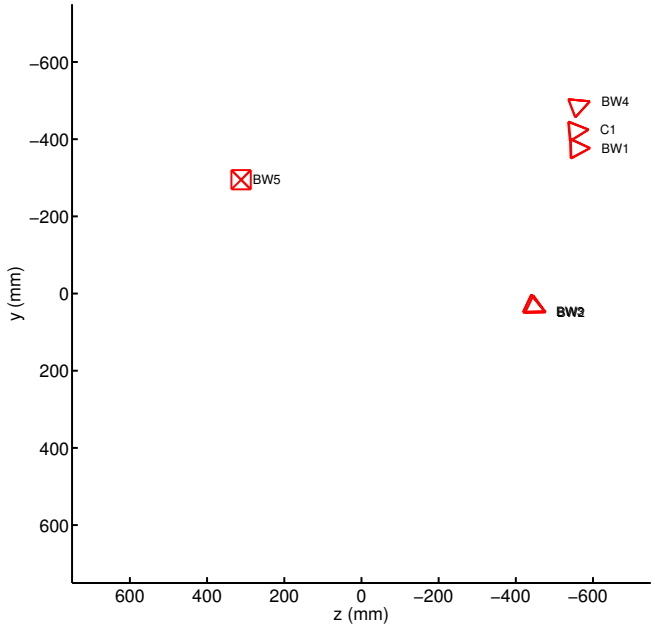


FIGURE 6. Extrinsic camera poses.

TABLE 1. Intrinsic camera parameters.

view#	fc	cc	alpha	kc
1	1415.70	353.558	-3.63209e-4	-0.144767
C1	1416.01	248.382		0.381532
2	1416.22	359.392	-1.73959e-3	-0.131226
BW1	1415.63	248.736		0.383097
3	1396.76	348.343	-1.43128e-2	-0.114650
BW2	1397.48	314.578		0.450306
4	1392.32	313.479	-1.88152e-2	-0.107240
BW3	1393.00	330.995		0.180247
5	610.940	387.146	-2.48452e-3	-0.135013
BW4	610.642	288.418		0.178274
6	982.937	397.317	-2.45798e-3	-0.158361
BW5	982.423	289.636		0.496054

TABLE 2. Extrinsic camera parameters.

view#	R			T
1	0.9991	-0.0064	-0.0416	-11.2566
C1	0.0102	0.9958	0.0913	-424.8976
	0.0408	-0.0916	0.9950	-588.6065
2	0.9990	0.0037	-0.0443	-13.1951
BW1	-0.0024	0.9996	0.0285	-376.8314
	0.0443	-0.0284	0.9986	-592.3195
3	0.9942	-0.0193	-0.1054	45.6233
BW2	-0.0281	0.9021	-0.4306	48.4747
	0.1034	0.4311	0.8964	-477.1215
4	0.9999	-0.0149	0.0084	-91.0789
BW3	0.0171	0.9060	-0.4229	46.5286
	-0.0013	0.4229	0.9062	-474.9482
5	0.9999	0.0065	-0.0144	-10.6171
BW4	-0.0009	0.9333	0.3592	-497.9704
	0.0158	-0.3591	0.9331	-592.0355
6	0.0032	-0.0038	-1.0000	662.1822
BW5	0.0001	1.0000	-0.0038	-294.7766
	1.0000	-0.0001	0.0032	311.6760

EEGLAB, Matlab, EDFBrowser, etc. The sensor attachment positions and protocol details are available in [3].

The bdf files include 47 channels. The list of channels, their labels and physical units are given in Tables 3 and 4. EEG electrodes were placed on a cap using international 10-20 system (see Fig. 7)

All the responses' files contain 30 seconds of before and after. If 30 seconds before or after of each trial was not available, the signals are zero-padded in all channels.

The last channel (channel 47) is the experiment status channel, and contains encodings of events in the experiments, such as the showing of stimuli data and the inputs from the participants. The value of the status channel is manipulated in a way to encode the starting and ending time of the stimuli. The rising edge of status channel from 0 to 16 indicates the moment playing video/displaying image started and stopped. These pulses are used to discriminate between the response signal and the 30 seconds intervals before and after any stimuli. More details on how events are encoded in the status channel are giving in section 3

2.4. **Eye Gaze Data.** The Tobii eye gaze data is stored in .tsv files (tab separated values), and supplied as an annotation to each data track (named "Gaze"), except for the audio tracks. The display resolution is set to 1280 x 800 pixels, on a display size of 51.9 x 32.45cm. The eye gaze direction is given as coordinates on this screen.

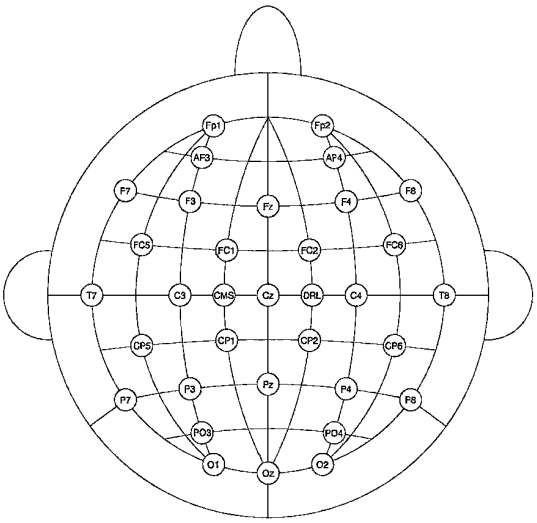


FIGURE 7. The EEG cap layout for 32 EEG electrodes in addition to two reference electrodes. Retrieved from the Biosemi website (http://www.biosemi.com).

TABLE 3. 32 electrodes were placed on a participants' scalp using a head cap. The physical unit of the measured EEG signals is uV and the positions of the electrodes are given in Figure 7. Their measurements were recorded in the following order in the bdf files.

Ch. no.	Ch. name	Ch. no.	Ch. name
1	Fp1	17	Fp2
2	AF3	18	AF4
3	F3	19	Fz
4	F7	20	F4
5	FC5	21	F8
6	FC1	22	FC6
7	C3	23	FC2
8	T7	24	Cz
9	CP5	25	C4
10	CP1	26	T8
11	P3	27	CP6
12	P7	28	CP2
13	PO3	29	P4
14	O1	30	P8
15	Oz	31	PO4
16	Pz	32	O2

Note that all the videos shown are resized and centered to touch the borders of the screen without changing the video aspect ratio. The remaining screen areas above and below, or left and right of the movie content, were filled with black.

TABLE 4. The peripheral nervous system physiological signals were recorded in bdf files in the following order and physical dimensions.

Ch. no.	Ch. name	position	physical unit
33	EXG1	ECG1 (upper right corner of chest, under clavicle bone)	uV
34	EXG2	ECG2 (upper left corner of chest, under clavicle bone)	uV
35	EXG3	ECG3 (left side of abdomen)	uV
36	EXG4	Unused	-
37	EXG5	Unused	-
38	EXG6	Unused	-
39	EXG7	Unused	-
40	EXG8	Unused	-
41	GSR1	Galvanic skin response, left middle and ring finger	Ohm
42	GSR2	Unused	-
43	Erg1	Unused	-
44	Erg2	Unused	-
45	Resp	Respiration belt	uV
46	Temp	Temperature, left pinky	Celsius
47	Status	Status channel containing markers	Boolean

The last column of the gaze data file, titled 'AudioSampleNumber' contains the corresponding location in the audio, with '1' corresponding to the first audio sample of the entire recording. Note that each session in the database is a fragment of a continuous recording of one experiment. The AudioSampleNumber corresponds to the position in the entire recording. To get the relative location in the fragment, the start time of the fragment has to be subtracted. The fragment start time can be found in the session.xml file. See section 2.5 for details. The gaze data file for each session is pruned to only contain the events related to the respective session.

Besides gaze data themselves, the Tobii data file also contains the events in the experiment program. These do not have an AudioSampleNumber related to them. However, the Tobii 'Timestamp' in the first column can easily be used to derive their respective location in the audio.

The starting and stopping of a movie are indicated in the 'Event' column, by event labels "MovieStart" and "MovieEnd", respectively. The starting and stopping of the showing of an image are indicated by event labels "ImageStart" and "ImageEnd", respectively. The 'Descriptor' column contains the respective video of image file names. The key presses of the participant are indicated by a 'KeyPress' event label. For the yes/no questions, special big buttons were used (Slammers). "Yes" was associated with key number "1", and "no" with number "2". Thus, a key press with 'Descriptor' value 'D1' corresponds to "yes", and a descriptor value 'D2' corresponds to "no". Each yes/no key press also generated three other KeyPress events in the data, which contain no information other then that the first event corresponds most closely to the actual moment the key was pressed.

Each track is also accompanied by a "Guide-Cut" annotation file. This file contains modified non-gaze data lines extracted from the gaze data file of the complete recording that a session is a fragment of. It shows how the entire recording is cut into fragments, with the video frame and the audio sample location of each cut (sample numbers are based at "1"). All cuts were made exactly between the middle of two

video frame exposures. The audio sample location of the cut is chosen as the nearest to this moment. The cut locations are always between two samples, in order to prevent confusion about whether a start- or end sample is included or excluded from a fragment. The fragment start- and end locations in the meta data that are described in paragraph 2.5, are derived from this file.

2.4.1. *Eye Gaze Data File Description.* Below, the different columns in the eye gaze data are described. The text is cited from the Tobii Studio Manual [4], except for the column 'AudioSampleNumber', which was added by us, after synchronising the Tobii data with the recorded audio.

- **Timestamp** - Timestamp in milliseconds for when the gaze data was collected
- **DateTimeStamp**- Timestamp recalculated and shown in minutes, seconds and milliseconds (mm:ss:ms)
- **DateTimeStampStartOffset** - Time for the timestamp in hh:mm:ss. The start time for the recording is also shown in the Replay and Visualization views in Tobii Studio if the option Show Date is selected in the Recordings field. The start time is taken from the computer time
- **Number** - Timestamps in numbered order
- **Gaze PointXLeft** - Horizontal screen position of the gaze point for the left eye
- **GazePointYLeft** - Vertical screen position of the gaze point for the left eye
- **CamXLeft** - Horizontal location of the left pupil in the camera image (0 is left edge, 1 is right edge)
- **CamYLeft** - Vertical location of the left pupil in the camera image (0 is top, 1 is bottom)
- **DistanceLeft** - Distance from the eye tracker to the left eye. The distance is given in mm on a straight axis right out from the eye tracker plane
- **PupilLeft** - Size of the pupil (left eye) in mm. The distance and pupil size measures are calculated to be as close to real values as possible. However, individual differences in the eyes of subjects and the strength of glasses/contact lenses will cause errors in these values. The measures still reflect changes in head position and pupil size accurately.
- **ValidityLeft** - Validity of the gaze data. The validity is 0 if the eye is found and the tracking quality good. If the eye cannot be found by the eye tracker the validity code will be 4. Read more under the Validity codes section further down
- **Gaze PointXRight** - The horizontal screen position of the gaze point for the right eye
- **GazePointYRight** - Vertical screen position of the gaze point for the right eye
- **CamXRight** - Horizontal location of the right pupil in the camera image (0 is left edge, 1 is right edge)
- **CamYRight** - Vertical location of the right pupil in the camera image (0 is top, 1 is bottom)
- **DistanceRight** -Distance from the eye tracker to the right eye. The distance is given in mm on a straight axis right out from the eye tracker plane
- **PupilRight** - Size of the pupil (right eye) in mm. The distance and pupil size measures are calculated to be as close to real values as possible. However, individual differences in the eyes of subjects and the strength of glasses/contact lenses will cause errors in these values. However, the measures still reflect changes in head position and pupil size accurately
- **ValidityRight** - Validity of the gaze data. The validity is 0 if the eye is found and the tracking quality good. If the eye cannot be found by the eye tracker the validity code will be 4. The value is for the right eye. Read more under the Validity codes section further down
- **FixationIndex** - Indexes for the fixation points
- **GazePointX** - Gaze PointX can be the horizontal screen position for either eye or the average for both eyes. The value to show here is specified in Tobii Studio under Tools → Settings → Fixation Filters → Eye Selection Filter. This value is also used for the fixation definition
- **GazePointY** - Gaze PointX can be the horizontal screen position for either eye or the average for both eyes. The value to show here is specified in Tobii Studio under Tools → Settings → Fixation Filters → Eye Selection Filter. This value is also used for the fixation definition
- **Event** - Events, automatic and logged, will show up under Events. The automatic events are start and end events for the different media, mouse clicks and key presses. The automatic events are listed in the event table in [4] under Event Key and Data. The logged events are the manually logged events entered either in the replay view or from the remote logger

- **EventKey** - Unique event key is shown for different key presses. The different event keys with corresponding events, data and descriptions are listed under Event Key and Data.
- **Data1** - Data field for the event. The contents of this field vary depending on what type of event this is. See the Event key table in [4]
- **Data2** - Data field for the event. The contents of this field vary depending on what type of event this is. See the Event key table in [4]
- **Descriptor** - Description of the event. The contents vary depending on what type of event this is. See the Event key table in [4]
- **StimuliName** - The file name of the media given in the setup in Tobii Studio
- **MediaWidth** - Media size in pixels
- **MediaHeight** - Media size in pixels
- **MediaPosX** - Distance from the left side of the screen to the media on the screen given in pixels
- **MediaPosY** - Distance from the top of the screen to the media on the screen given in pixels
- **MappedFixationPointX** - X coordinate for the fixation point mapped to the media coordinate system, where the origin for the coordinate system is the top left corner of the image/media
- **MappedFixationPointY** - Y coordinate for the fixation point mapped to the media coordinate system, where the origin for the coordinate system is the top left corner of the image/media
- **FixationDuration** - Fixation duration. The time in milliseconds that a fixation lasts
- **AoiIds** - ID number for the AOI, usually a counter starting at zero for the first AOI
- **AoiNames** - Name of the AOI or AOIs if fixations on multiple AOIs are to be written on the same row
- **WebGroupImage** - Filename of the image file that is used to visualize the web group
- **MappedGazeDataPointX** - X coordinate for the raw gaze data point mapped to the media coordinate system where the origin for the coordinate system is the top left corner of the image/media
- **MappedGazeDataPointY** - Y coordinate for the raw gaze data point mapped to the media coordinate system where the origin for the coordinate system is the top left corner of the image/media
- **MicroSecondTimestamp** - Timestamp for this export row in microseconds, relative to gaze recording start.
- **AbsoluteMicroSecondTimestamp** - Timestamp for this export row in microseconds.
- **AudioSampleNumber** - Timestamp for this export row in the corresponding sample in the recorded audio, relative to audio recording start.

2.4.2. *Validity Codes.* Cited from the Tobii Studio Manual [4]:
Validity code ranges from 0 to 4, with the following interpretations for each value:

- **0** - The system is certain that it has recorded all relevant data for the particular eye and that the data recorded belongs to the particular eye (no risk of confusing left eye with right eye by the system).
- **1** - The system has only recorded one eye, and has made some assumptions and estimations regarding if the recorded eye is left or right. However, it is still highly probable that the estimations made are correct. The validity code on the other eye is in this case always set to 3.
- **2** - The system has only recorded one eye and has no way of determining if this is the left or the right eye.
- **3** - The system is fairly confident that the actual gaze data is actually incorrect or corrupted. The other eye will always have validity code 1.
- **4** - The actual gaze data is missing or definitely incorrect. A couple of gaze data with validity code 4 on both eyes, followed by a number of gaze data with validity code 0 on both eyes, are usually a sure sign of a blink.

It is recommended that the validity codes should always be used for data filtering, to remove data points that are obviously incorrect. For most studies, we recommend removing all data points with a validity code of 2 or higher.

2.5. **Meta Data.** In the session.xml files, the session tag contains important meta data about each session. Apart from the experiment-specific labels, all the experiments in the database have a basic set of labels in the session tags. Below is an list of the basic meta data labels with their possible values or value range and a short description.

- hasBeard : {True,False} *Participant's facial hear state*
- hasMoustache : {True,False} *Participant's facial hear state*
- hasGlasses : {True,False} *Whether she/he was wearing glasses*
- expType : {1,2,3,4} *See Section 3.1 for expType="1", and 3.2 for expType={2,3,4}*
- isStim : {0,1} *"0" for 'no stimulation', "1" for 'stimulus present'*
- cutNr : {1,2,...} *Count for adjoining cuts of the original source, split in multiple sessions*
- cutLenSec : $\langle 0,\infty \rangle$ *audio cut length in seconds (audRate used as time base)*
- vidRate : $\langle 0,\infty \rangle$ *video frame/sample rate per second (audRate used as time base)*
- audRate : $\langle 0,\infty \rangle$ *Specified audio sample rate per second*
- audBeginSmp : {1.5,2.5,...} *Begin sample number (counting from "1") of this fragment in its original*
- audEndSmp : {1.5,2.5,...} *End sample number (counting from "1") of this fragment in its original*
- vidBeginSmp : {1.5,2.5,...} *Begin sample number (counting from "1") of this fragment in its original*
- vidEndSmp : {1.5,2.5,...} *End sample number (counting from "1") of this fragment in its original*

The following meta data labels appear only in the sessions of experiment type "1" where a participant gave feedback to a video that was shown to stimulate their emotions (see Section 3.1 for the meaning of the emotion numbers):

- mediaFile : {[file name].avi} *The stimulus video (also given for the neutral videos in between)*
- feltEmo : {0,1,...,12} *Emotion that was felt, see Section 3.1 for the complete list*
- feltArsl : {1,...,9} *Arousal that was felt, 1 for 'none', 9 for 'maximum'*
- feltVlnc : {1,...,9} *Valence that was felt, 1 for most negative, 9 for most positive, 5 for neutral*
- feltCtrl : {1,...,9} *Control that was felt, 1 for no control, 9 for full control*
- feltPred : {1,...,9} *Predictability that was experienced, 1 for unpredictable, 9 for completely predictable*

The following meta data labels appear only in the sessions of experiment type 2,3 and 4 where a tagged image or video was shown:

- mediaFile : {[file name].avi,[file name].jpg} *The presented image or video*
- tagValid : {0,1} *"0" when the shown tag was meant to apply, or "1" for tags that should not apply*
- tagAgree : {0,1} *"0" when the participant found the tag inappropriate, or "1" when (s)he agreed*

Please note that the first video frame is not captured at the same time as the first audio sample number. The difference varies, with the start of the video capture typically around half a second after the start of audio capture. The audio and video streams can be related by knowing that the session start and stop times do correspond to the exact same moment. This has been achieved through analysis of the camera trigger signal that is included in the original audio recordings (not included in the fragments in the database).

2.6. **Synchronized setup.** An overview of the synchronization in the recording setup is shown in Fig. 8. To synchronize between sensors, we centrally monitor the timings of all sensors, using a MOTU 8pre[1] eight-channel audio interface ('c' in Fig. 8). Since the analog inputs of the 8Pre are sampled using a shared device clock, an event in one of the channels can be directly related to a temporal location in all other channels. The external trigger signal of the cameras('b' in Fig. 8) was directly recorded alongside the recorded sound, in a parallel audio track (see the fifth signal in Fig. 9). The camera trigger pulses can be easily detected and matched with all the captured video frames, using their respective frame number and/or time stamp. With the audio sampling rate of 48kHz, the uncertainty of localizing the rising camera trigger edge is around 20μs. When the 30μs latency and jitter of 1.3μs of the camera exposure is taken into account, and the timing of multiple trigger pulses is combined, the resulting synchronization error between audio and video can be kept below 25μs. More details about the data synchronization can be found in [2].

The gaze tracking data and physiological signals were recorded with separated capture systems. Neither of them allowed to connect to the actual sensor trigger signals. This is why an alternative synchronization strategy was required for both. The physiological data was captured with a multi-channel A/D converter ('a' in Fig. 8) that allowed to record one binary input signal alongside the data. This input was used to connect the camera trigger signal. Since the accurate timing of each camera frame is known, this allowed to synchronize the physiological data with all the other modalities.

[1] http://www.motu.com/products/motuaudio/8pre

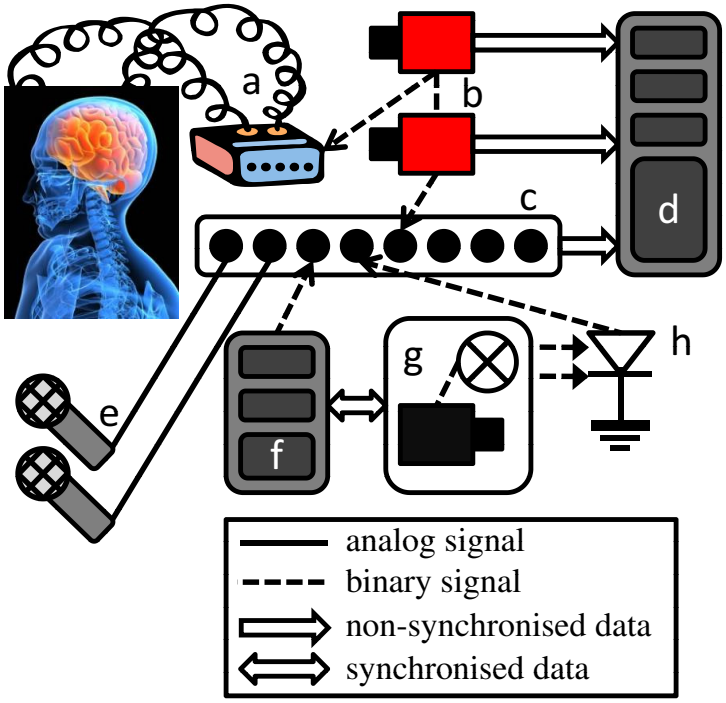


FIGURE 8. Overview of our synchronized multi-sensor data capture system, consisting of a physiological measurement system (a), video cameras (b), a multi-channel A/D converter (c), an A/V capture PC (d), microphones (e), an eye gaze capture PC (f), an eye gaze tracker (g) and a photo diode (e) to capture the pulsed IR-illumination from (g).

The eye gaze tracker ('g' in Fig. 8) synchronizes with the CPU cycle counter of its dedicated capture PC ('f') with an accuracy of approximately one millisecond. To synchronize the respective CPU cycle counter to the audio interface, we developed an application that periodically (twice per second) outputs binary time-stamp signals with the current time, through the serial port output (see the third signal in Fig. 9), with an error below 10 microseconds. To get a more accurate timing accuracy than the 1ms accuracy of the timestamps of the gaze tracking data, the infrared strobe illumination of the gaze tracker was recorded using a photo diode ('h' in Fig. 8 and the fourth signal in Fig. 9). This allows to correct the gaze data timestamps up to 10 microseconds accurate, if necessary.

The start moments of the stimuli data were timestamped using the same synchronized CPU cycle counter as the eye-gaze data. An uncertainty in timing of the stimuli data is introduced by the video player software, as well as the latency of the audio system, graphics card and the screen. Furthermore, the accuracy of the time codes of the fragments may introduce further errors in synchronizing the recorded data with the actual stimuli. The room microphone was placed close to the speaker that produced the stimuli sound. Therefore, the recorded ambient sound provides an implicit synchronization, as it includes the sound of the stimuli.

3. EXPERIMENTS

For each participants, four recordings are made subsequently. The first is the Emotion Elicitation Experiment, explained in section 3.1. The three other recordings belong to the implicit tagging experiment, explained in section 3.2.

3.1. **Emotion Elicitation Experiment.** The emotion elicitation experiment (expType="1"), which includes the responses to emotional videos, is the first recording. Each volunteer is asked to watch a sequence

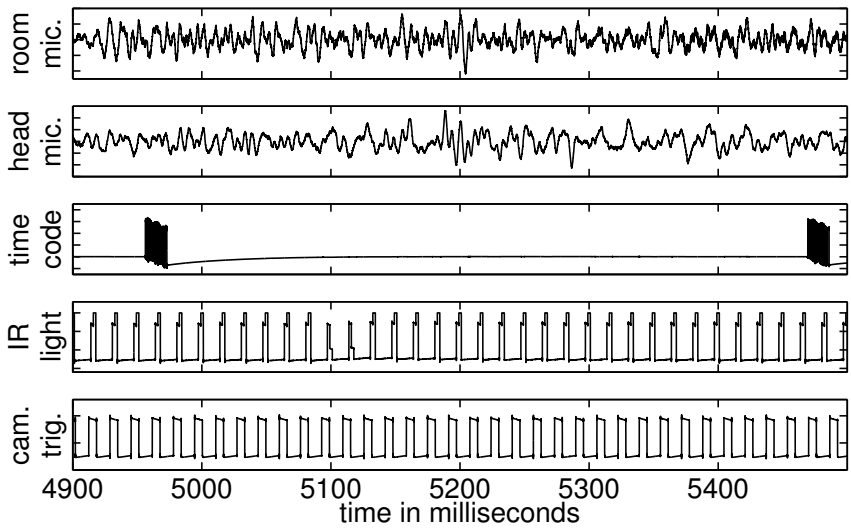


FIGURE 9. 5 tracks recorded in parallel by MOTU 8pre audio interface. From top to bottom: (1) room microphone; (2) head microphone; (3) serial port time stamp output (transmitted at 9600bps), showing 2 time stamp signals; (4) measured infrared light in front of eye tracker; (5) camera trigger.

of video clips. The clips are selected to elicit emotions such as happiness, sadness, disgust, amusement. The total duration of the experiment is approximately 40 minutes.

For the emotional experiment for each trial there is a 15 seconds neutral clip which contains the baseline before each emotional video. The emotion experiment is the first recording for all participants. Five multiple choice questions were asked during the self report for each video. The five questions were 1. emotional label/tag; 2. arousal; 3. valence; 4. dominance; 5. predictability (Fontaine et al, 2007). The emotional labels and their associated key were: 1. Sadness 2. Joy, happiness 3. disgust 4. neutral 5. amusement 6. anger 7. fear 8. surprise 9. anxiety. To simplify the interface a keyboard was provided with only nine numerical keys and the participant could answer each question by pressing one of the nine keys. Questions 2 to 5 were on a nine points scale.

The 'session' tag in the session XML file contains the file name of the video shown and the participants' responses to all five questions. For example:

- feltEmo="2"
- feltArsl="4"
- feltVlnc="3"
- feltCtrl="'1"
- feltPred="5"
- mediaFile="69.avi"

Note that, for database compatibility, the corresponding assignment of emotion numbers to 'feltEmo' are different from the key numbers used for the participant's feedback. The correct assignments are given in table 5.

Unfortunately, we do not have the rights for sharing the videos that we used in this experiment. Please refer to [3] for the sources of the videos.

3.2. **Implicit Media Tagging Experiments.** In the Implicit Media Tagging experiments, each volunteer is asked to watch a sequence of photographs (expType=2,3) or video clips (expType="4"). First without any tag, and then with a tag underneath that can be related or unrelated to the content ('correct' or 'incorrect', respectively). The clips and photographs are selected to observe the participant's reactions when agreeing or disagreeing with the tag associated with the displayed material. After each trial, the participants were

TABLE 5. Emotion numbers assigned to the session variable 'feltEmo', together with the corresponding keyboard numbers that were used for giving the feedback. "n.a." is indicated for emotions that were not included as a choice.

feltEmo#	Emotion name	Feedback Key#
0	Neutral	4
1	Anger	6
2	Disgust	3
3	Fear	7
4	Joy, Happiness	2
5	Sadness	1
6	Surprise	8
7	Scream	n.a.
8	Bored	n.a.
9	Sleepy	n.a.
10	Unknown	n.a
11	Amusement	5
12	Anxiety	9

TABLE 6. Video fragments used for having a neutral affect.

Cut#	file name
1,9,17,25,33	colorbars_Final.avi
3,11,19,27,35	seagulls_Final.avi
5,13,21,29,37	sticks_Final.avi
7,15,23,31,39	waves_Final.avi

asked whether the tag was correct or incorrect. Using this protocol, it is possible to study agreement and disagreement on displayed tags.

The 'session' tag in the session XML file contains the file name of the image of video shown and the correctness of the tag according to the participant. For example:

- mediaFile="1-421615509_7637215ddd_b-Y.jpg"
- tagValid="1"
- tagAgree="1"

The videos shown in the experiment are from the Hollywood Human Action dataset [1], and are included in the supplemental data. Unfortunately, since we did not have copyright on the images, we could not provide them in the way they were shown. Instead, the images included in the supplemental data only contain extracted edges.

The tables 8, 9 and 10 list the order in which the tagged images and videos were presented to the subject. The file name extension '-T' corresponds to media with appropriate tags, while the extension '-N' corresponds to media with tags that are considered to not apply.

There are two ways to find the timings of events in the experiment. The ".tsv" files that contain the eye gaze data also contains the moments when the media fragments are shown, as well as the time-stamped key inputs from the participant. Alternatively, the status channel in the physiological data can be used:

Again, the rising edge of square shaped pulses (from 0 to 16) on the status channel indicates the moment playing video/displaying image started and stopped. These pulses in the status channel are used to discriminate between the response signal and the 30 seconds intervals before and after the stimuli.

In the tagging experiments, the moment that the tag was shown is also indicated using a pulse in the status channel. Therefore there are three pulses in the status channel of tagging experiment (see Fig. 10). the rising edge of the second square shaped pulse indicates the moment the image or video was displayed with a tag. The correctness of the tag and the participant's response are coded in the pulse amplitude (This information is also available in the XML files). The following table shows the four conditions and their corresponding pulse amplitude in the channel status. The answers (yes/no) were given in response to a question which was asked after each image which was whether the tag was correct.

TABLE 7. Video fragments shown as stimuli in the affective tagging experiments (Experiment Type 1).

			source		
Cut#	**file name**	**emotion**	**movie name**	**start time**	**end time**
1	colorbars_Final.avi		neutral clip included in supplementary material		
2	69.avi	disgust	*Hannibal*	1:44:50.7	1:45:49.9
3	seagulls_Final.avi		neutral clip included in supplementary material		
4	55.avi	anger/sadness	*The pianist*	0:54:33.3	0:55:50.4
5	sticks_Final.avi		neutral clip included in supplementary material		
6	58.avi	amusement	*Mr Bean's Holiday*	1:17:19	1:18:18
7	waves_Final.avi		neutral clip included in supplementary material		
8	earworm_f.avi	disgust	http://blip.tv/file/1335283/		
9	colorbars_Final.avi		neutral clip included in supplementary material		
10	53.avi	amusement	*Kill Bill VOL I*	1:12:12.2	1:13:57.2
11	seagulls_Final.avi		neutral clip included in supplementary material		
12	80.avi	joy	*Love actually*	0:09:45.76	0:11:22.96
13	sticks_Final.avi		neutral clip included in supplementary material		
14	52.avi	amusement	*Mr Bean's Holiday*	1:05:53.2	1:07:30.6
15	waves_Final.avi		neutral clip included in supplementary material		
16	79.avi	joy	*The thin red line*	0:07:37.96	0:08:21.68
17	colorbars_Final.avi		neutral clip included in supplementary material		
18	73.avi	fear	*The shining*	2:16:42.3	2:17:55.2
19	seagulls_Final.avi		neutral clip included in supplementary material		
20	90.avi	joy	*Love actually*	0:33:59.6	0:35:25.8
21	sticks_Final.avi		neutral clip included in supplementary material		
22	107.avi	fear	*The shining*	2:07:02.8	2:07:38.2
23	waves_Final.avi		neutral clip included in supplementary material		
24	146.avi	sadness	*Gangs of New York*	2:34:41.1	2:36:10
25	colorbars_Final.avi		neutral clip included in supplementary material		
26	30.avi	fear	*Silent Hill*	1:22:27.6	1:23:39.5
27	seagulls_Final.avi		neutral clip included in supplementary material		
28	138.avi	sadness	*The thin red line*	1:06:32	1:08:29.8
29	sticks_Final.avi		neutral clip included in supplementary material		
30	newyork_f.avi	neutral	http://accuweather.com/ n.a. (please refer to audio ch. 1)		
31	waves_Final.avi		neutral clip included in supplementary material		
32	111.avi	sadness	*American History X*	1:52:05.9	1:54:00
33	colorbars_Final.avi		neutral clip included in supplementary material		
34	detroit_f.avi	neutral	http://accuweather.com/ n.a. (please refer to audio ch. 1)		
35	seagulls_Final.avi		neutral clip included in supplementary material		
36	cats_f.avi	joy	http://www.youtube.com/watch?v=E6h1KsWNU-A		
37	sticks_Final.avi		neutral clip included in supplementary material		
38	dallas_f.avi	neutral	http://accuweather.com/ n.a. (please refer to audio ch. 1)		
39	waves_Final.avi		neutral clip included in supplementary material		
40	funny_f.avi	joy	http://blip.tv/file/1854578/		

Again, the physiological signals are stored using Biosemi data format (BDF) which is readable by EEGLAB, Matlab, EDFBrowser, etc. The files are named using the following syntax:

`Part_[participant's code]_Trial[trial number]_tagging[experiment id*].bdf`

The video or image file names and participants' responses to express their agreement with the tag are given in a xml file where the videos/images and the responses are listed in the order in which they were played.

*experiment id can be:

- "Images1": the first image tagging experiment with images (expType="2")

TABLE 8. Tagged Images shown in Experiment Type 2.

Cut#	file name
1	introduction
2	1-421615509_7637215ddd_b-Y.jpg
3	2-1336550827_d2a841d3ec_b-Y.jpg
4	3-2358624529_b3aac64037_b-N.jpg
5	4-2517011622_a5f5740fa6_b-Y.jpg
6	5-155258820_87853679c0_b-N.jpg
7	6-2318937925_a14631d93f_b-Y.jpg
8	7-2414609572_be4b7d4288_o-Y.jpg
9	8-2983347275_884c79bd49_o-Y.jpg
10	9-2435839690_f22c20ec01_o-Y.jpg
11	10-2358624529_b3aac64037_b-Y.jpg
12	11-2851771094_7876a96c6d_b-Y.jpg
13	12-2318937925_a14631d93f_b-N.jpg
14	13-2932458839_f7baef980d_b-N.jpg
15	14-1184434206_e2a5d115a2_b-Y.jpg
16	15-2517011622_a5f5740fa6_b-N.jpg
17	16-2414609572_be4b7d4288_o-N.jpg
18	17-155258820_87853679c0_b-Y.jpg
19	18-2435839690_f22c20ec01_o-N.jpg
20	19-2932458839_f7baef980d_b-Y.jpg
21	20-2959652616_1d6d4067cf_b-N.jpg
22	21-2983347275_884c79bd49_o-N.jpg
23	22-3192358088_c6664d8fde_b-Y.jpg
24	23-421615509_7637215ddd_b-N.jpg
25	23-1184434206_e2a5d115a2_b-N.jpg
26	25-3192358088_c6664d8fde_b-N.jpg
27	26-1336550827_d2a841d3ec_b-N.jpg
28	27-2959652616_1d6d4067cf_b-Y.jpg
29	28-2851771094_7876a96c6d_b-N.jpg
30	end of experiment

- "Images2": the second image tagging experiment with images (expType="3")
- "Videos": the video tagging experiment with images (expType="4")

If any of the trials is missing due to technical difficulties, its bdf file is not included. Please see section 4 for details on missing data.

4. Missing and Incomplete Recordings

Table 12 lists al the inconsistencies in the data.

5. Terms of Use

To protect the data from unauthorized access, the data has to be stored on firewall-protected data servers which are not directly connected to Internet. Each participant has declared that his/her audiovisual and gaze recordings may be used for academic research publication in documents. In addition, some participants have declared that his/her audiovisual and gaze recordings may be used for multimedia presentations for academic purposes. Table 13 shows the permissions each participant has agreed to.

In the above, 'academic research' refers to a non-commercial research conducted at academic institutions. This rules out any research by commercial companies as well as non-academic governmental research institutions.

Publications in documents for academic purposes include articles submitted to scientific conferences or journals and posters used to present research at scientific conferences. Multimedia presentations are presentations where audio and/or video features of the recordings of the signers may be used. Again, the

TABLE 9. Tagged Images shown in Experiment Type 3.

Cut#	file name
1	introduction
2	1-155899204_f8454cd229_b-N.jpg
3	2-2096233113_ce4a0bfb50_b-Y.jpg
4	3-2173973284_3363fb9aae_b-Y.jpg
5	4-157764976_e46a9c376b_o-N.jpg
6	5-2376828121_d3f2f5d819_b-Y.jpg
7	6-245796052_3b1e3ce595_b-Y.jpg
8	7-2565809652_0affbb6393_o-N.jpg
9	8-2827062969_951d6cf19b_b-N.jpg
10	9-3265128695_45df003e22_b-N.jpg
11	10-3554356369_d5f7014735_o-Y.jpg
12	11-3635250988_9a5712f44b_o-Y.jpg
13	12-39571035_924e6f24e0_o-N.jpg
14	13-2376828121_d3f2f5d819_b-N.jpg
15	14-155899204_f8454cd229_b-Y.jpg
16	15-2173973284_3363fb9aae_b-N.jpg
17	16-245796052_3b1e3ce595_b-N.jpg
18	17-2565809652_0affbb6393_o-Y.jpg
19	18-3669149316_89d7833ab1_b-Y.jpg
20	19-499214049_8ac816897f_b-N.jpg
21	20-2096233113_ce4a0bfb50_b-N.jpg
22	21-157764976_e46a9c376b_o-Y.jpg
23	22-2827062969_951d6cf19b_b-Y.jpg
24	23-3635250988_9a5712f44b_o-N.jpg
25	24-3669149316_89d7833ab1_b-N.jpg
26	25-3554356369_d5f7014735_o-N.jpg
27	26-3265128695_45df003e22_b-Y.jpg
28	27-499214049_8ac816897f_b-Y.jpg
29	28-39571035_924e6f24e0_o-Y.jpg
30	end of experiment

presentations will be given for non-commercial, academic purposes, which might include presentations for conferences and course lectures.

All researchers that wish to use the database for their research are required to sign the End User License Agreement (EULA). Only researchers who signed the EULA will be granted access to the database. In order to ensure secure transfer of data from the database to an authorised user's PC, data will be protected by SSL (Secure Sockets Layer) with an encryption key. If at any point, the administrators of MAHNOB database and/or MAHNOB researchers have a reasonable doubt that an authorised user does not act in accordance to the signed EULA, he/she will be declined the access to the database.

6. Acknowledgement

The recording of this dataset was not possible without the financial support from Swiss national science foundation. The work also recieved support in part from the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement Petamedia no. 216444. I would like to thank Jozef Doboš, Prof. Didier. Grandjean and Dr. Guillaume Chanel (University of Geneva) for their valuable scientific comments and technical support during the experiments.

References

1. Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld, *Learning realistic human actions from movies*, IEEE Conference on Computer Vision & Pattern Recognition, 2008.

TABLE 10. Tagged Videos from Hollywood Human Actions (HOHA) database, shown in Experiment Type 4.

Cut#	**file name**
1	introduction
2	1-ID-02774-N.avi
3	2-BF-02479-N.avi
4	3-BJM-01293-N.avi
5	4-BJM-01877-Y.avi
6	5-BJM-02001-Y.avi
7	6-BJM-02669-N.avi
8	7-BOTD-01740-Y.avi
9	8-DPS-00205-N.avi
10	9-ID-00172-N.avi
11	10-BOTD-00239-Y.avi
12	11-DPS-02590-N.avi
13	12-ID-01698-Y.avi
14	13-ID-02241-Y.avi
15	14-BF-00077-N.avi
16	15-BOTD-01740-N.avi
17	16-BF-02479-Y.avi
18	17-DPS-00205-Y.avi
19	18-BOTD-00239-N.avi
20	19-DPS-02590-Y.avi
21	20-BJM-01293-Y.avi
22	21-ID-00172-Y.avi
23	22-ID-02241-N.avi
24	23-BF-00077-Y.avi
25	24-ID-02774-Y.avi
26	25-BJM-01877-N.avi
27	26-BJM-02669-Y.avi
28	27-ID-01698-N.avi
29	28-BJM-02001-N.avi
30	end of experiment

TABLE 11. The status channel second pulse amplitude for the starting time of displaying tags on videos and images and their meanings.

	Correct displayed tag	Incorrect displayed tag
Positive response (yes)	32(agr.)	48 (disagr.)
Negative response (no)	64(disagr.)	80(agr.)

2. Jeroen Lichtenauer, Jie Shen, Michel Valstar, and Maja Pantic, *Cost-effective solution to synchronised audio-visual data capture using multiple sensors*, Tech. report, Imperial College London, 180 Queen's Gate, London, UK, 2010.
3. M. Soleymani, J. lichtenauer, T. Pun, and M. Pantic, *A multi-modal affective database for affect recognition and implicit tagging*, IEEE Transactions on Affective Computing (under review).
4. Tobii Technology AB, *User manual: Tobii x60 & x120 eye trackers, revision 3*, November 2008.

IMPERIAL COLLEGE LONDON, DEPARTMENT OF COMPUTING
E-mail address: j.lichtenauer@imperial.ac.uk

TABLE 12. Missing data. A "+" means that a media stream is available, "-" means that it is missing.

subj.	exp.#	cut#	audio	video	body	details
all	2,3,4	1,30	+	+	-	use cut 2 and 29 (data include 30 seconds before and after each cut)
3	1	35-40	-	-	-	experiment ended prematurely due to technical failure
9	1	29	+	+	-	experiment ended prematurely due to technical failure
		30-40	-	-	-	
12	1	all	+	+	-	recording error
15	1	33	+	+	-	experiment ended prematurely due to physical discomfort
		34	-	+	-	
		35-40	-	-	-	
	2,3,4	all	-	-	-	experiments cancelled due to physical discomfort
16	1	33	+	+	-	experiment ended prematurely due to technical failure
		34	-	+	-	
		35-40	-	-	-	
26	1	all	+	cam 2,3	+	loss of data from cameras 1,4,5,6

TABLE 13. Permissions given by the recorded subjects. 'Research' is defined as non-commercial research conducted at academic institutions. 'Publication' is defined as publication in documents and/or multimedia presentations for academic purposes.

subj.#	research	publication
1	yes	yes
2	yes	yes
3	yes	yes
4	yes	yes
5	yes	yes
6	yes	no
7	yes	yes
8	yes	yes
9	yes	yes
10	yes	yes
11	yes	no
12	yes	yes
13	yes	yes
14	yes	yes
15	yes	no
16	yes	yes
17	yes	yes
18	yes	yes
19	yes	yes
20	yes	yes
21	yes	yes
22	yes	yes
23	yes	yes
24	yes	yes
25	yes	no
26	yes	yes
27	yes	yes
28	yes	yes
29	yes	yes
30	yes	yes

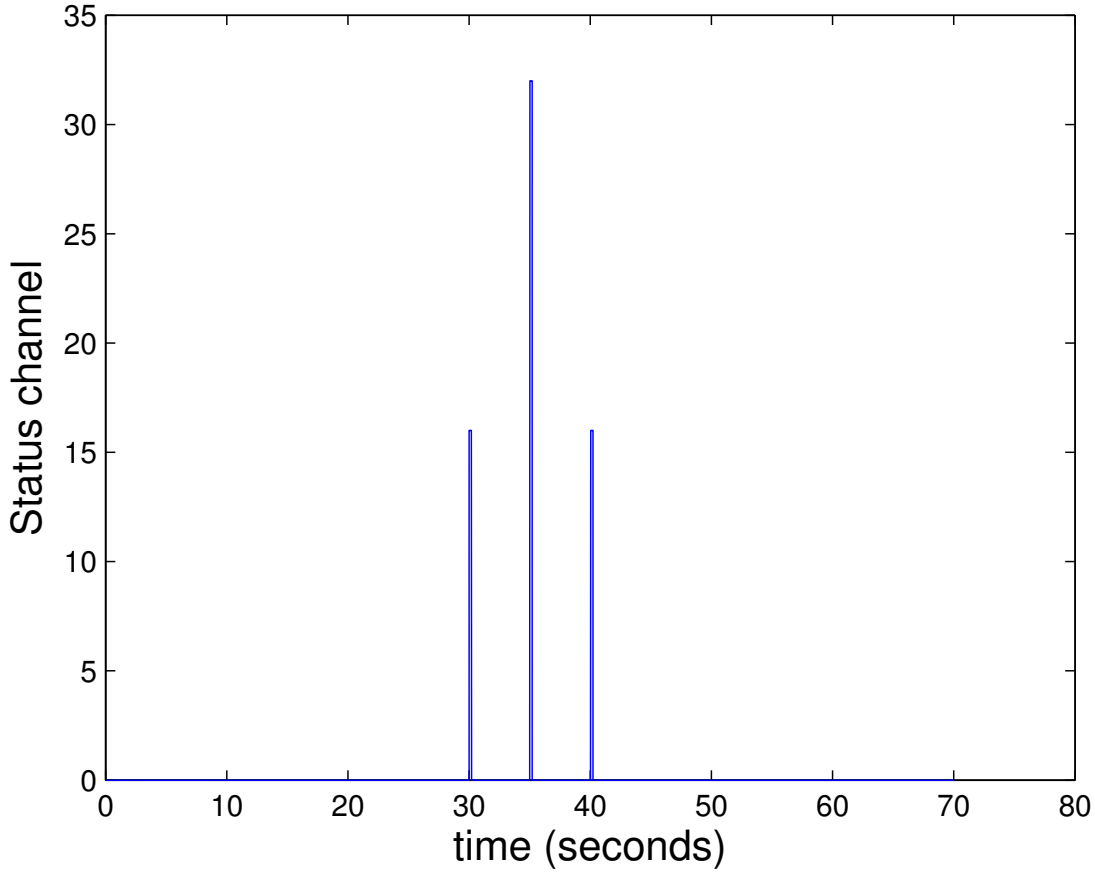


FIGURE 10. An example of the status channel for one image tagging experiment. The stimulus started exactly at 30s and around 35s the image with a corect tag shown which receieved a "yes" response from the participant (status channel = 32). The stimulus ended around 40s.