SEWA

"Automatic Sentiment Analysis in the Wild"

Innovation Action

Horizon2020

Grant Agreement no. 645094

Deliverable D1.1
SEWA Database

Deliverable Type:	Dataset
Dissemination level:	CO (Consortium)
Month:	M18
Contractual delivery date:	July 31, 2016
Actual delivery date:	August 17, 2016
Version:	1.0
Total number of pages:	16

Document Information

Grant Agreement no.	645094	**Acronym**	SEWA
Full Title	Automatic Sentiment Analysis in the Wild		
Project URL	http://www.sewaproject.eu/		
Document URL	http://www.sewaproject.eu/deliverables/		
EU Project Officer	Philippe Gelin		

Deliverable	**Number**	D1.1	**Title**	SEWA Database
Work Package	**Number**	WP1	**Title**	SEWA DB collection, annotation and release

Authors	Jie Shen, Maja Pantic			
Responsible Author	**Name**	Jie Shen	**E-mail**	js1907@imperial.ac.uk
	Co-ordinator	Jie Shen	**Phone**	+44 7909394377

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
12 August 2016	0.1	Jie Shen	First draft
16 August 2016	0.2	Maja Pantic	Final version

This project has received funding from the European Union's Horizon2020 Programme under grant agreement no 645094.

Table of contents

Executive summary

In this deliverable, we present the SEWA database (SEWA DB), a multilingual dataset of annotated facial, vocal and verbal behaviour recordings made in-the-wild (in naturalistic setting). This database will be not only an extremely valuable resource for researchers both in Europe and internationally but it will also push forward the research in automatic human behavioural analysis and user-centric HCI and FF-HCI in a similar manner as PASCAL pushed forward the field of object detection. SEWA DB will be used for a number of challenges and benchmarking efforts and will have more than 200 active users worldwide by the end of the project. The SEWA DB has been release online at http://db.sewaproject.eu/.

1 Introduction

A database of annotated audio and 2D visual dynamic behaviour (recorded by standard webcams used by the volunteers) has been collected within the SEWA project.

In the data-collection experiments, volunteers have been divided into pairs based on their cultural background, age and gender. Each pair of the subjects participated in two parts of the experiment: watching a total of 4 advertisements, and then discussing about the last advertisement though video-chat software. The entire watching of adverts and the subsequent conversation between the volunteers were recorded using web-cameras and microphones integrated into the laptops/PCs of the volunteers.

In the SEWA project, we recorded 6 groups of volunteers (around 30 persons per group) from six different cultural backgrounds: British, German, Hungarian, Greek, Serbian, and Chinese. The volunteers in each group have a broad distribution in gender and age. Specifically, there are at least three pairs of native speakers in each age group (18~29, 30~39, 40~49, 50~59, and 60+) for each culture. The resulting database contains a total of 204 sessions of experiment recordings: 1525 minutes of audio-visual data of people's reaction to adverts from 408 individuals, and 568 minutes of recorded computer-mediated face-to-face interactions between pairs of subjects.

The SEWA database includes annotations of the recordings in terms of facial landmarks, facial action unit (FAU) intensities, various vocalisations, verbal cues, mirroring, and rapport, continuously valued valence, arousal, liking, and prototypic examples (templates) of (dis)liking and sentiment. The data has been annotated in an iterative fashion, starting with a sufficient amount of examples to be annotated in a semi-automated manner and used to train various feature extraction algorithms developed in SEWA, and ending with a large DB of annotated facial behaviour recorded in the wild.

Accurately labelled/annotated real-world data are the crux in designing audio-visual human behaviour sensing, tracking and interpretation algorithms that will achieve robust performance in-the-wild. The SEWA DB is the very first of that kind to be released for research purposes. This database is not only an extremely valuable resource for researchers both in Europe and internationally but it will also push forward the research in automatic human behavioural analysis and user-centric human-computer interaction and computer-mediated face-to-face interaction (FF-HCI).

2 SEWA Data Acquisition

An important aspect of the SEWA project lies in collecting suitable datasets of enough labeled examples to facilitate the development of robust tools for automatic machine understanding of human behaviours. To create such dataset, a data collection experiment has been conducted, resulting in a large in-the-wild audio-visual corpus containing a wide variety of spontaneous expressions of emotions and sentiment.

In this experiment, participants were divided into pairs based on their cultural background, age and gender. During initial sign-up, participants were asked to complete a questionnaire of demographic measures including gender, age, cultural background, education, personality traits, and familiarity with the other person in the pair. To promote natural interactions, participants within each pair were required to know each other personally in advance of the experiment. Each pair of the participants then took part in two parts of the experiment, resulting in two sets of recordings.

- **Experimental Setup Part 1:** Each participant was asked to watch 4 adverts, each of being around 60 seconds long. These adverts had been chosen to elicit mental states including amusement, empathy, liking and boredom. After watching the advert, the participant was also asked to fill-in a questionnaire to self-report his/her emotional state and sentiment toward the advert.
- **Experimental Setup Part 2:** After watching the 4th advert, the two participants were asked to discuss the advert they had just watched by means of the video-chat function provided by the SEWA data collection website. On average, the conversation was 3 minutes long. The discussion was intended to elicit further reactions and opinions about the advert and the advertised product, such as whether the advertised is to be purchased, whether it is to be recommended to others, what are the best parts of the advert, whether the advert is appropriate, how it can be enhanced, etc.. After the discussion, each participant was asked to fill-in a questionnaire to self-report his/her emotional state and sentiment toward the discussion.

The SEWA data collection experiment was conducted using a website specifically built for this task (shown in Figure 1). The website (http://videochat.sewaproject.eu) utilises WebRTC/OpenTok to facilitate the playing of adverts, video-chat, and synchronized audio/video recording using the microphone and webcam on the participants' own computer. This setup allowed the participants to be recorded in truly unconstrained "in-the-wild" environments with various lighting conditions, poses, background noise levels, and sensor qualities.

During the SEWA experiment, 204 recording sessions have been successful, with a total of 408 subjects being recorded. The subjects were coming from 6 different cultural backgrounds: British, German, Hungarian, Serbian, Greek, and Chinese. 206 of the participants are male, 202 are female, resulting in a gender ratio (male / female) of 1.020. The participants cover 5 age groups: 18~29, 30~39, 40~49, 50~59 and 60+, with the 18~29 group being most numerous. The detailed participant demographics are shown in Table 1.

Table 1: SEWA Participant Demographics

Cultural Background		**Age Group**		**Years Known the Other Participant**		**Self-Reported Familiarity Rating**	
British	66	**18~29**	203	**<1**	80	**Not Familiar**	9
				1	30		
German	72	**30~39**	94	**2**	39	**Slightly Familiar**	13
				3	44		
Hungarian	70	**40~49**	33	**4**	37	**Somewhat Familiar**	35
				5~9	61		
Serbian	72	**50~59**	48	**10~14**	20	**Moderately Familiar**	120
				15~19	22		
Greek	56	**60+**	30	**20+**	75	**Extremely Familiar**	231
Chinese	72						

A total of 2040 audio-visual recording clips (5 clips per subject: 4 recorded during the advert-watching part and 1 recorded during the video-chat part) were collected during the experiment, comprising of 1525 minutes of audio-visual data of people's reaction to adverts and 568 minutes of video-chat recordings. Due to the wide spread of the participants' computer's hardware capacity, the quality of the video and audio recordings is not constant. Specifically, the spatial resolution of the video recordings ranges from 320x240 to 640x360 pixels and the frame rate is between 20 and 30 fps. The audio recording's sample rate is either 44.1 or 48 kHz.

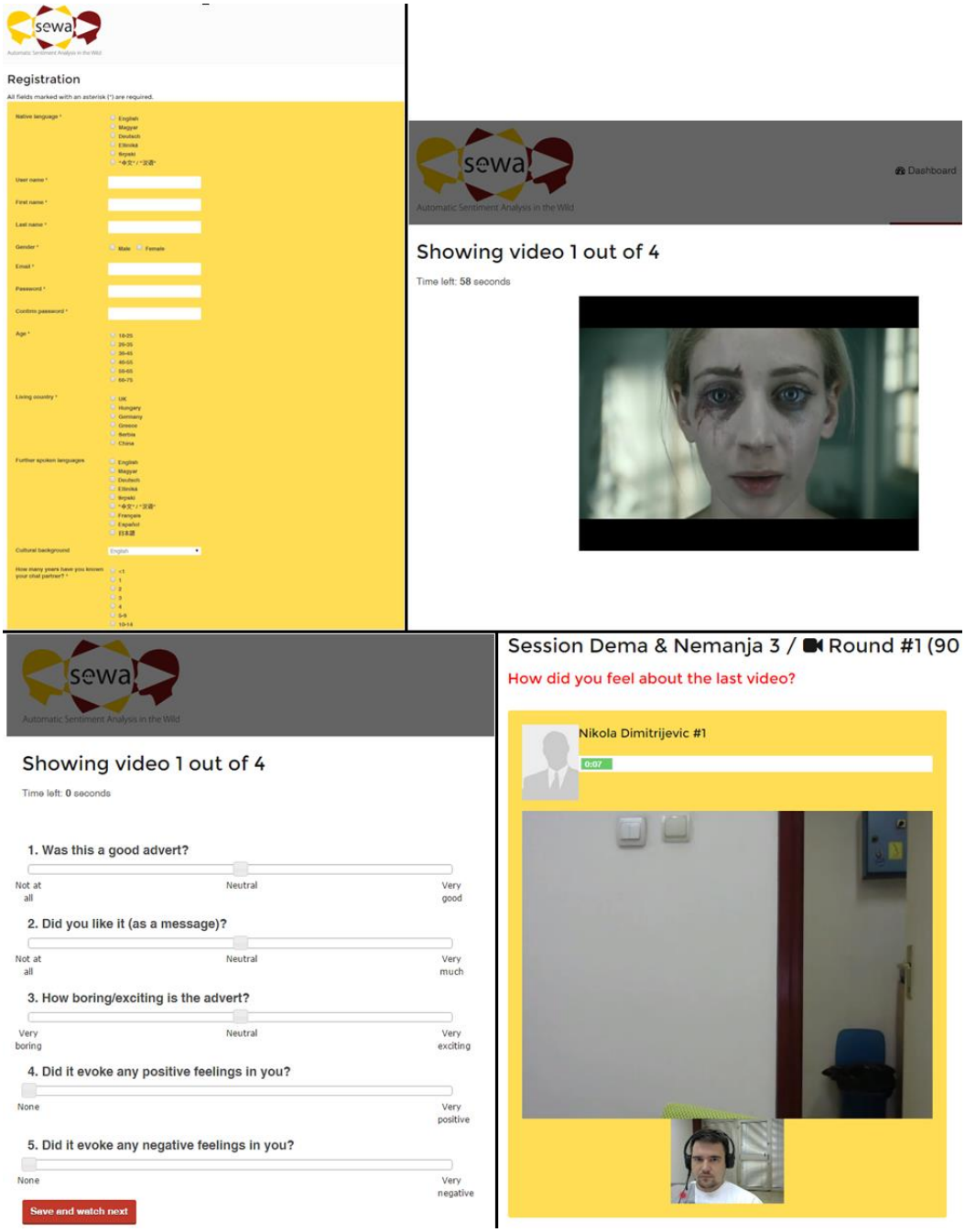


Figure 1: The SEWA collection website.

3 SEWA Data Annotation

The SEWA database contains annotations for facial landmarks, LLD features, hand gestures, head gestures, facial action units (AUs), verbal and vocal cues, continuously-valued valence, arousal and liking / disliking (toward the advertisement), template behaviours, episodes of agreement / disagreement, and mimicry episodes.

Due to the large amount of raw data acquired from the experiment, the SEWA database has been annotated iteratively, starting with sufficient amount of examples to be annotated in a semi-automated manner and used to train various feature extraction algorithms developed in SEWA. Specifically, 538 short (10~30s) video-chat recording segments were manually selected to form the fully-annotated basic SEWA dataset. These segments were selected based on the subjects' to the subjects' emotional state of low / high valance, low / high arousal, and liking / disliking. All 6 cultures were evenly represented in the basic SEWA dataset, with approximately 90 segments selected from each culture based on the consensus of at least 3 annotators from the same culture.

3.1 Facial Landmarks

The 49 facial landmarks were annotated for all segments included in the basic SEWA dataset. The annotation was performed semi-automatically [1]. We first applied an automatic facial landmark tracker [2] on all video segments and checked the tracking results to identify the frames with tracking errors. We then manually corrected 1/8 of these frames and used the annotation result to train a set of person-specific trackers. These person-specific trackers were applied to the rest of the frames to obtain more accurate tracking results. Afterward, the updated landmark locations were manually verified, and if necessary, corrected, to form the final annotation result. An example of the facial landmark annotation obtained from this process is show in Figure.

Figure 2: An example of facial landmark annotation.

3.2 Audio Low-Level Descriptors

We provide two sets of low-level audio descriptor (LLD) features for all recordings included in the SEWA database: the 65 dimension ComPareELLD and the more compact 18 dimension GeMAPSv01aLLD [3] [4]. The features were extracted automatically in 10ms steps.

3.3 Hand Gestures

We annotated hand gestures for all video-chat recordings in 5 frame steps. Five types of hand gestures were labelled: hand not visible (89.08%), hand touching head (3.32%), hand in static position (0.63%), display of hand gestures (2.39%), and other hand movements (3.68%). Some examples of the labelled frames are shown in Figure 3.

3.4 Head Gestures

For training the head nod / shake detector developed in WP3, we annotated head gestures in terms of nod and shake for all segments in the basic SEWA dataset. The annotation was performed manually on a frame-by-frame basis. To be able to provide good training examples for the head nod/shake detector, we emphasised specifically on high precision during the annotation process. Specifically, only un-ambiguous displays of head nod / shake were labelled. In the end, a total of 282 head nod sequences and 122 head shake sequences were identified. Examples of the labelled head nod / shake sequences are shown in Figure 4.

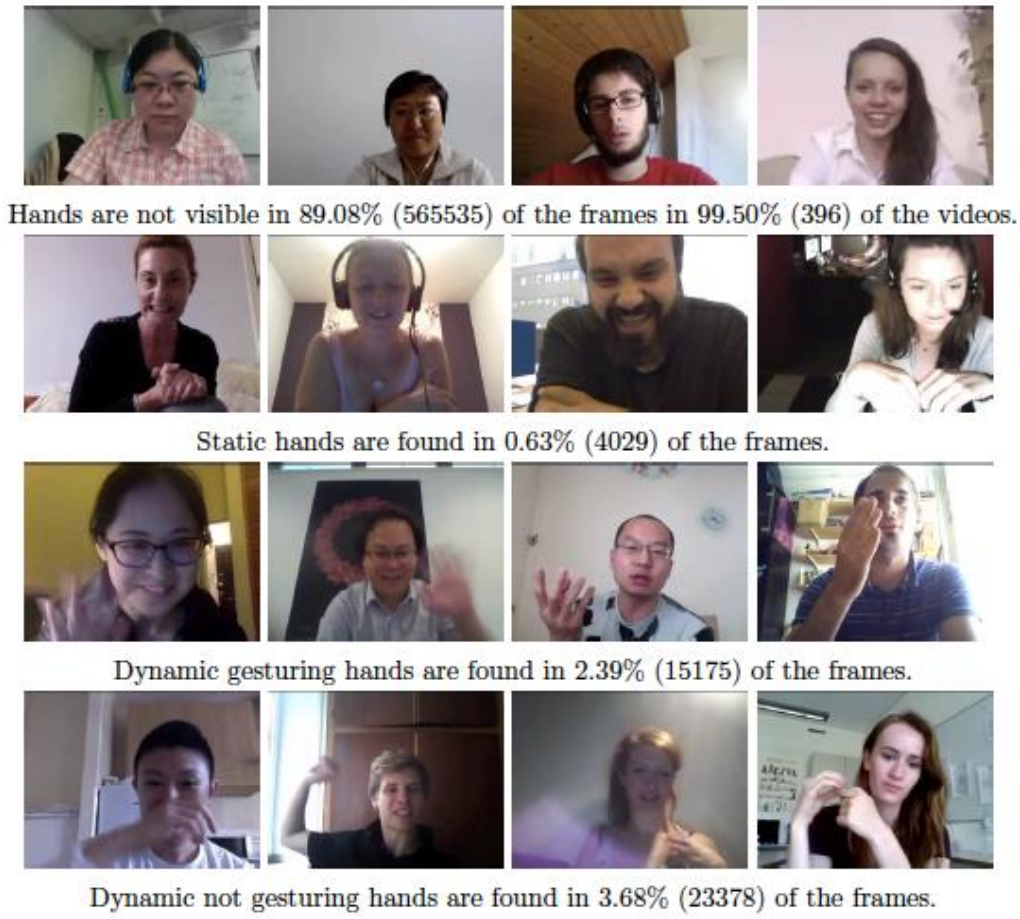


Figure 3: Examples of hand gesture annotation.

Figure 4: Examples of head nod (top row) and head shake (bottom row) sequences.

3.5 Facial Action Units

In order to train the action unit detectors developed in WP3, we extracted examples of 5 facial action units (AU) from the basic SEWA dataset: inner eyebrow raiser (AU1, 109 examples), outer eyebrow raiser (AU2, 79 examples), eyebrow lowerer (AU4, 94 examples), lip corner puller (AU12, 104 examples), and chin raiser (AU17, 61 examples). Similar to the case of facial landmarks, the AU examples were identified in a semi-automatic manner. Specifically, we first applied automatic AU detectors to the video segments and manually removed all false-positives from the detection results. Consequently, the AU annotation is not exhaustive, meaning that some AU activations may be missed. Examples of the annotated action units are shown in Figure 5.

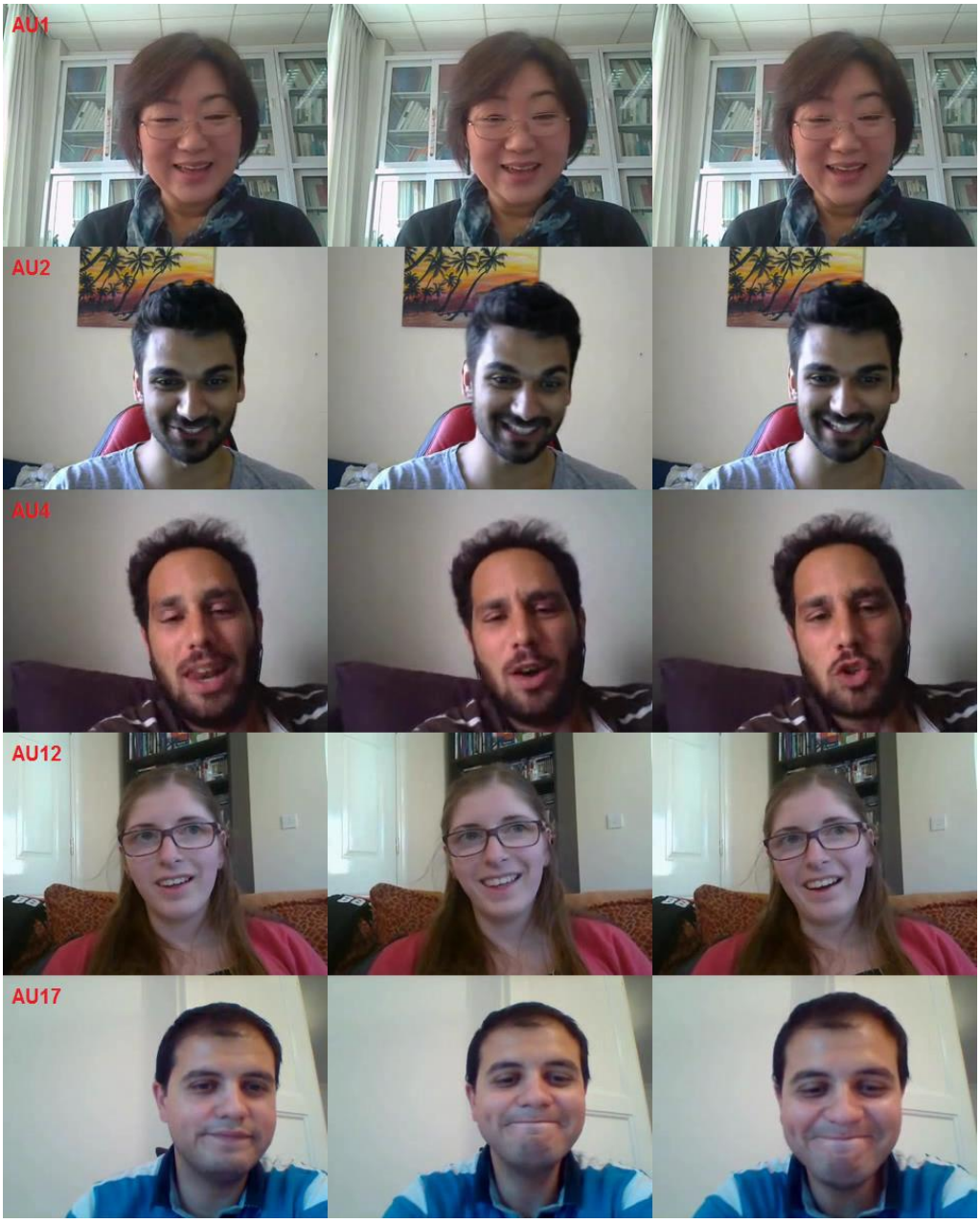


Figure 5: Examples of the facial action units annotation.

3.6 Transcript

We provide the audio transcript of all video-chat recordings. In addition to the verbal content, the transcript also contains labels of certain non-verbal cues, such as sighing, coughing, laughter, and so on. An example of the audio transcript is shown in Figure 6.

```
start_time,end_time,subject,text
6.208465,6.818082,66,"Szevasz "
11.582718,12.416930,65,"Na hogy tetszettek?"
12.850079,14.037227,66,"Én most nem hallak"
16.620077,18.256416,66,"Várjál most hallak, felhangositalak"
17.069268,17.502417,65,"Miért nem hallasz?"
22.283095,24.448839,66,"Szóval mit gondolsz a legutolsó videóról?"
27.464837,27.577135,65,"Hát..."
27.817773,29.277645,65,"A legutolsóóóó..."
29.422028,31.122538,65,"???"
31.459431,31.956750,66,"Micsoda?"
32.502196,36.994109,65,"???"
34.651897,35.630493,65,"?Egyenlő szárú?"
37.010152,37.764151,65,"Nekem az tetszett"
37.892492,40.122405,65,"???"
40.122405,42.400447,66,"Amúgy a csap az nagyon jó ötlet"
```

Figure 6: An example of the audio treanscript.

3.7 Valence, Arousal and Liking / Disliking

Continuously-valued valance, arousal and liking / disliking (toward the advertisement) were annotated for all segments in the basic SEWA dataset. These will be used as the training data for valence / arousal detector to be developed in WP4. In order to identify the subtle changes in the subjects' emotional state, annotators were always hired from the same cultural background of the recorded subjects. In addition, to reduce the effect of the annotator bias, 5 annotators were recruited for each culture. The annotation was performed in real-time using a joystick with a sample rate of 66 Hz. To avoid cognitive overload on the annotators, the three dimensions (valence, arousal and liking / disliking) were annotated separately in three passes. Furthermore, for each dimension, the segments were annotated three times, first based on audio data only, then based on video data only and finally based on audio-visual data. An example of the end result of this annotation process is illustrated in Figure 7.

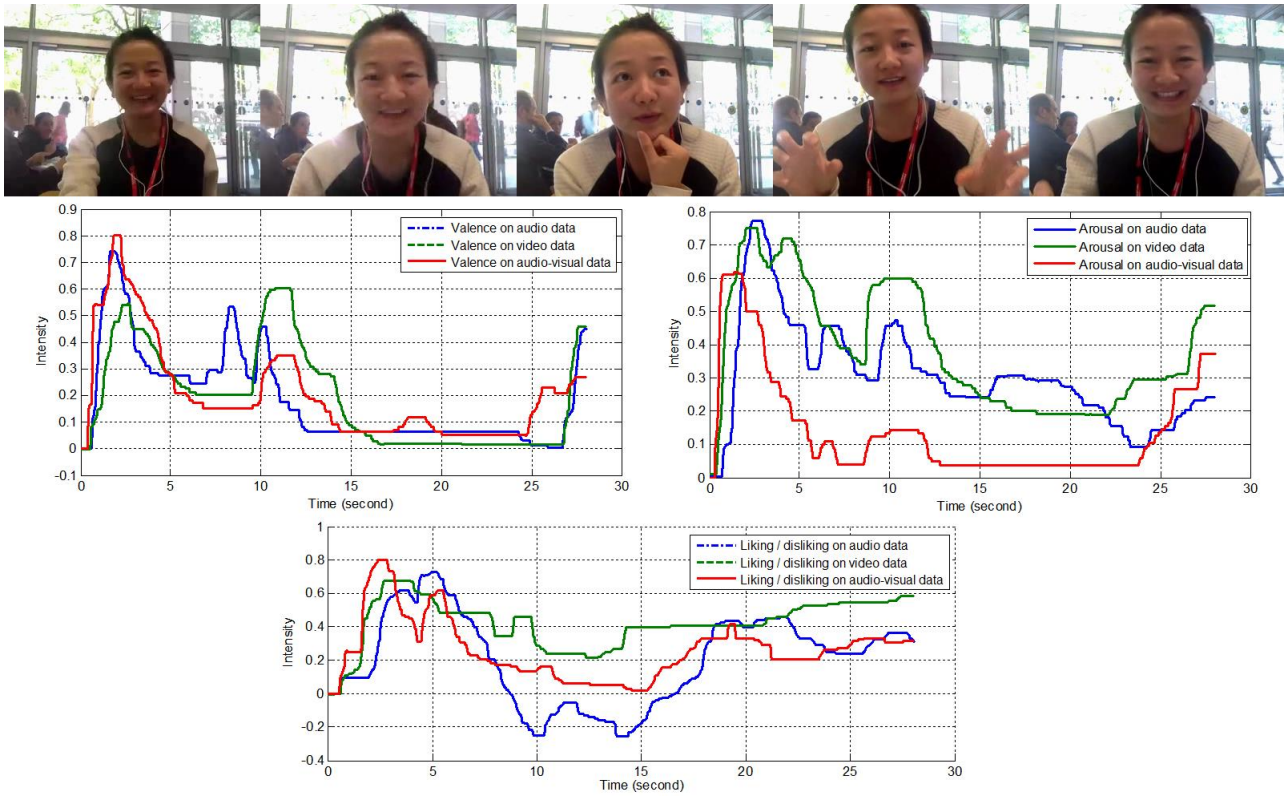


Figure 7: An example of the continuously-valued annotation results on valence, arousal and liking / disliking.

3.8 Behaviour Templates

With the help of the same annotators recruited for valence, arousal and liking / disliking, we identified behaviour templates for each culture when the subjects are in the emotional state of low / high valence, low / high arousal or showing liking / disliking toward the advertisement. For each category, at least two examples were identified. Table 2 shows the exact distribution of the templates found in the basic SEWA dataset. These templates will be used to train and test the behaviour similarity detector, which is to be addressed in WP5. Figure 8 illustrates some examples of these behaviour templates.

Table 2: Templates Behaviours Identified in the Basic SEWA Dataset

Culture	Low Valence	High Valence	Low Arousal	High Arousal	Liking	Disliking
British	2	2	2	2	2	2
German	4	4	3	3	4	4
Hungarian	2	2	2	2	2	2
Serbian	6	5	2	6	6	6
Greek	2	2	2	2	2	2

Chinese	3	4	2	4	5	4

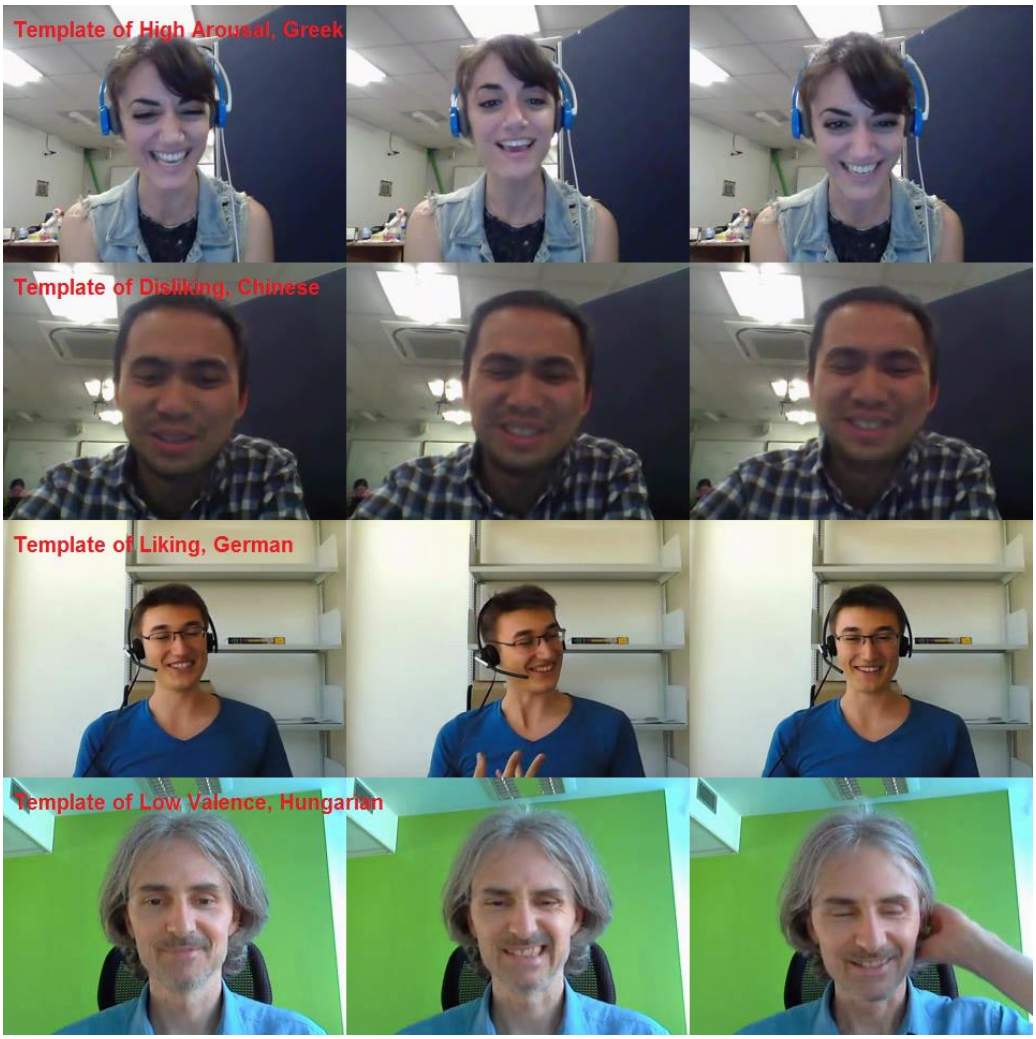


Figure 8: Some behaviour templates found in the basic SEWA dataset.

3.9 Agreement / Disagreement Episodes

To prepare the training data for the sentiment analysis algorithm to be developed in WP6, we extract a number of episodes from the video-chat recordings in which the pair of subjects were in low, mid or high level of agreement / disagreement with each other. The selections were based on the consensus of at least 3 annotators from the same culture of the recorded subjects. The exact numbers of agreement / disagreement episodes are shown in Table 3. Two examples of the agreement / disagreement episodes are shown in Figure 9.

Table 3: Agreement / Disagreement Episodes Identified in the Video-Chat Recordings

Culture	Strong Agreement	Moderate Agreement	Weak Agreement	Weak Disagreement	Moderate Disagreement	Strong Disagreement
British	12	26	29	7	3	3
German	7	7	7	6	9	6
Hungarian	7	6	6	5	5	5
Serbian	7	7	7	4	6	4
Greek	5	5	5	5	5	5

Chinese	5	6	6	4	5	3

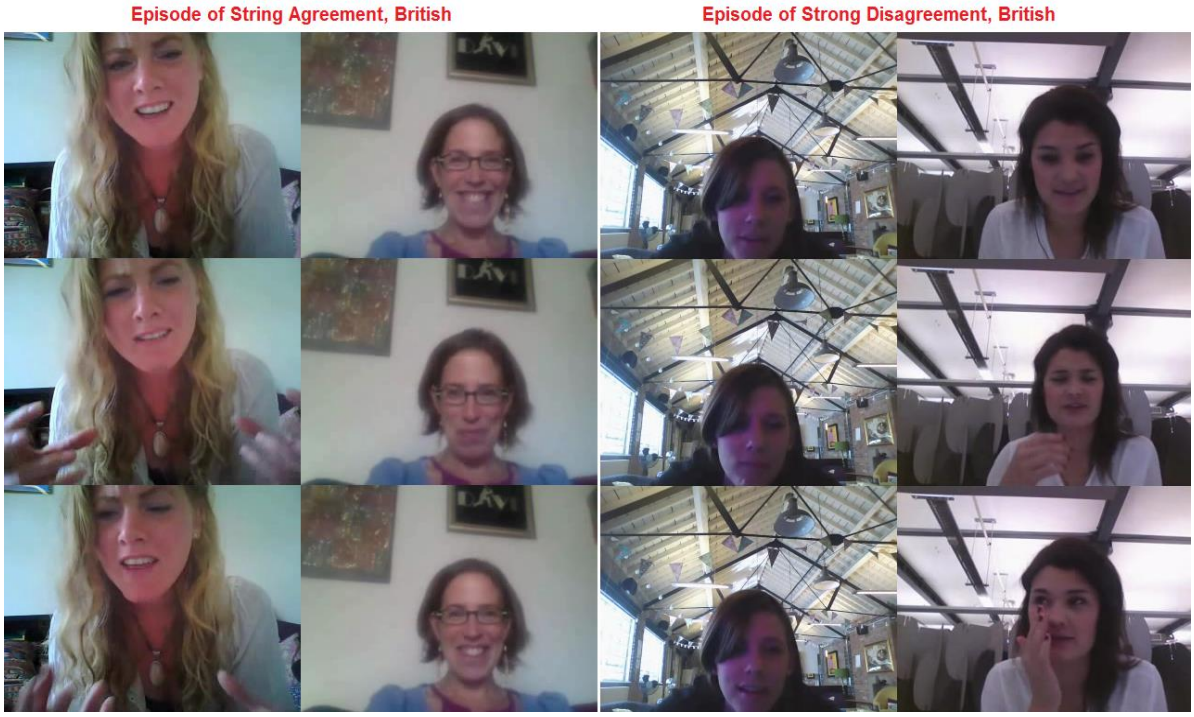


Figure 9: Examples of episodes of agreement / disagreement.

3.10 Mimicry Episodes

Last but not least, a total of 197 mimicry episodes (48 British, 31 German, 39 Hungarian, 20 Serbian, 41 Greek and 17 Chinese), in which one subject mimicked the facial expression and / or head gesture of the other subject, were identified from the video-chat recordings. They will be used as the training data for the mimicry detection algorithm, which is to be addressed in WP6. Two examples of the identified mimicry episodes are shown in Figure 10.

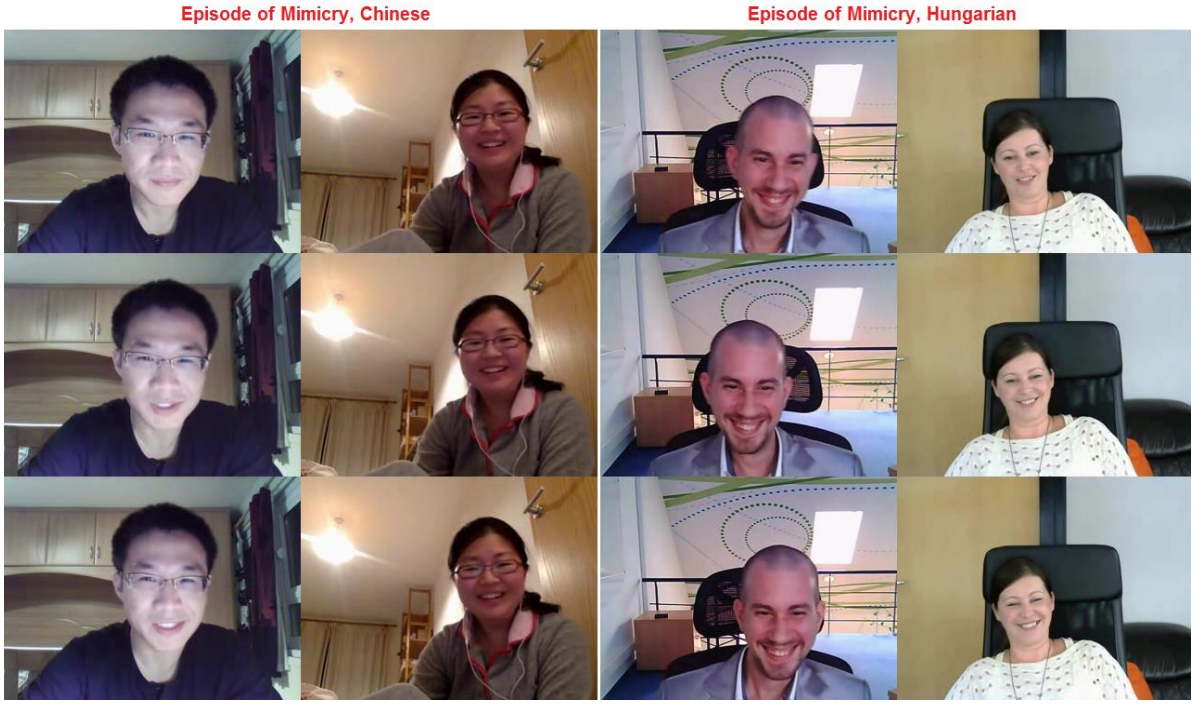
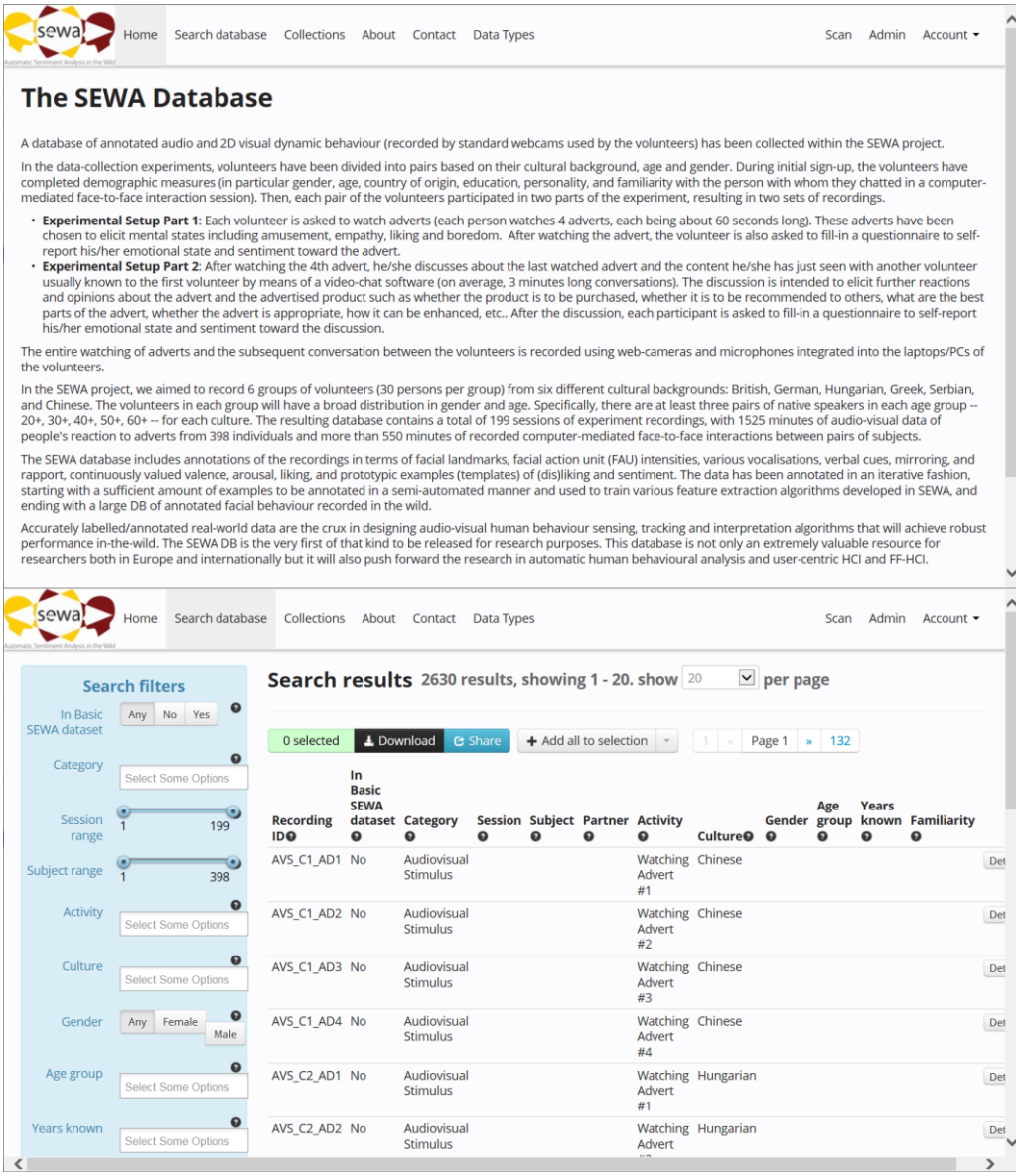


Figure 10: Examples of the identified mimicry episodes.

4 SEWA Database Release

The SEWA DB is released online at: http://db.sewaproject.eu/. The web-portal provides a comprehensive search filter (shown in Figure 11) allowing users to search for specific recordings based on various criteria, such as demographic data (gender, age, cultural background, etc.), availability of certain types of annotation, and so on. This will facilitate investigations during and beyond the project in the field of machine analysis of facial behaviour as well as in other research fields.

The SEWA database is made available to researchers for academic-use only. To comply with clauses stated in the Informed Consent signed by the recorded participants, all non-academic/commercial uses of the data are prohibited. To enforce this retraction, an end-user licence agreement (EULA) is prepared (see Appendix). 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, the data are protected by SSL (Secure Sockets Layer) with an encryption key. If at any point, the administrators of the SEWA database and/or SEWA 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.



Figrue 11: Front page of the online SEWA database and the search filters.

5 Conclusion

We introduced the SEWA database (SEWA DB), a multilingual dataset of annotated facial, vocal and verbal behaviour recordings made in-the-wild. The primary use of the SEWA database is to provide training data for the technologies developed during the SEWA projects. In addition, the SEWA DB has also been made publicly available to the research community, representing a benchmark for efforts in automatic analysis of audio-visual behaviour in the wild. The SEWA DB contains the recordings of 204 experiment sessions, covering 408 subjects recruited from 6 different cultural backgrounds: British, German, Hungarian, Greek, Serbian, and Chinese. The database includes a total of 1525 minutes of audio-visual recordings of the subjects' reaction to the 4 advertisement stimuli and 568 minutes of video-chat recordings of the subjects discussing the advertisement. In addition to the raw audio and video data, the SEWA DB also contains a wide range of annotations including: low-level audio descriptor (LLD) features, facial landmark locations, hand-gesture, head gesture, facial action units, audio transcript, continuously-valued valence, arousal and liking / disliking (toward the advertisement), template behaviours, agreement / disagreement episodes, and mimicry episodes. The SEWA DB has been released online at: http://db.sewaproject.eu/.

6 References

[1] G. S. Chrysos, E. Antonakos, S. Zafeiriou and P. Snape. Offline deformable face tracking in arbitrary videos. In IEEE International Conference on Computer Vision Workshops (ICCVW), 2015. IEEE, 2015.

[2] A. Asthana, S. Zafeiriou, G. Tzimiropoulos, S. Cheng, and M. Pantic. From pixels to response maps: Discriminative image filtering for face alignment in the wild. IEEE PAMI, 37(6):1941–1954, 2015.

[3] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, S. Kim, The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism, Proc. Interspeech 2013, ISCA, Lyon, France, 2013.

[4] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, K. Truong, The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing, in IEEE Transactions on Affective Computing, vol.PP, no.99, pp.1-1.