<table>
<thead>
<tr>
<th>Start date of project (dur.)</th>
<th>01 April 2015 (24 Months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document due Date</td>
<td>30 September 2016</td>
</tr>
<tr>
<td>Responsible for deliverable</td>
<td>Brno University of Technology</td>
</tr>
<tr>
<td>Reply to</td>
<td><a href="mailto:ihradis@fit.vutbr.cz">ihradis@fit.vutbr.cz</a></td>
</tr>
<tr>
<td>Document status</td>
<td>Final</td>
</tr>
</tbody>
</table>

| Project reference no.         | H2020 644632              |
| Project working name          | MixedEmotions              |
| Project full name             | Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets |
| Security (distribution level) | PU                        |
| Contractual delivery date     | 30th September 2016        |
| Deliverable number            | D4.6                      |
| Deliverable name              | Emotion Recognition from Image and Video Content, final version |
| Type                          | Report                    |
| Version                       | Final                     |
| WP / Task responsible         | WP4                       |
| Contributors                  | Michal Hradiš, Alena Pavelková, Kamil Behúň |
| EC Project Officer            | Martina Eydner            |
Table of Contents
1 Executive Summary .................................................................................................................... 4
2 Introduction ................................................................................................................................ 4
3 Low-level information ................................................................................................................ 6
  3.1 Face detection and tracking ................................................................................................. 6
  3.2 Facial landmark localization ................................................................................................. 7
  3.3 Head pose estimation ......................................................................................................... 10
  3.4 Face alignment .................................................................................................................... 14
  3.5 Body pose estimation ......................................................................................................... 16
4 High-level information ............................................................................................................. 23
  4.1 Identification ....................................................................................................................... 23
  4.2 Age and gender estimation ................................................................................................. 23
  4.3 Expression estimation ......................................................................................................... 24
  4.4 Facial Action Unit detection and Eye gaze tracking ............................................................ 26
  4.5 Video-based Emotion Recognition for AVEC 2016 Challenge ........................................ 26
5 Conclusions ............................................................................................................................... 29
References .................................................................................................................................... 30
1 Executive Summary

Emotion recognition is the keystone of the MixedEmotions project. The project encompasses emotion recognition from multiple modalities (text, speech and video) and multiple languages. This document, Deliverable 4.6, describes tools, methods and datasets used in the MixedEmotions for extracting high-level information from images and videos of people. This includes the analysis of facial expressions and underlying emotions, and of personal attributes such as gender and age.

2 Introduction

This document targets two types of scenarios: i) one person is being recorded while either she passively observes some audio/video content, or participates in a face-to-face or a video-mediated interaction with other people, ii) multiple people appear in an edited video such as a video-interview or a debate. The scenarios with single passive observer include consumer studies of, for example, advertisements, product designs, movie trailers, and TV shows. The dynamic scenarios include various interviews and possibly video-mediated communication and collaboration. In the videos that two or more persons are presented, such as television interviews and debates, we track the identities of the participants through the video by employing CNN-based facial recognition, and we process the individual tracks in the same manner as in the single-person case while handling the parts of video where the person is not present in the same way as tracking failures in the single person scenarios.

This document presents a vision system for estimation of human emotions, facial expressions, age, gender and other attributes. We focus on near-frontal views as methods and datasets for extreme viewpoints either do not exist for many of the required capabilities or the state-of-the-art recognition quality for such views is still not sufficient for practical applications. In addition to extracting information from faces, we employ upper body pose estimation. The parts of the system are illustrated in Figure 1.
In order to obtain any meaningful high-level semantic information such as emotions, basic information about the position and orientation of body parts has to be obtained first. Section 3 presents our approach to face detection, tracking, facial part localization, head orientation estimation, and upper body part localization.

Personal information (e.g. gender, age, and possibly personality) and appearance-related information (e.g. facial hair, glasses) can be estimated from static facial images using existing pattern recognition methods trained on publicly available datasets. Similarly, estimators of facial expressions (neutral, smile, frown) can be learned from existing image and video datasets (Section 4.3). All these methods perform the best on well-aligned facial images. Ideally, the faces should be “frontalized” (transformed as if directly facing a camera). The text briefly presents our approach for face alignment using 2D similarity transformation (Section 3.4).

Current state-of-the-art in emotion recognition significantly lags behind facial expression recognition. One reason is that emotion recognition is a much harder problem in which the underlying emotion has to be inferred from facial expressions while taking into account contextual and temporal information. On the other hand, facial expressions are directly observed. Additionally, the complexity of emotion recognition is reflected in the small size of available datasets. As an alternative to high-level emotion information, medium-level features can be extracted from videos. These include encoded body, head, eye, and facial movements as action units (FACS encoding). Such information can be later used as features for a number of recognition tasks including emotion recognition. Section 4 presents our approach for extracting high-level
information: identification (Section 4.1), age and gender (Section 4.2), facial expression (Section 4.3), facial action unit detection and eye gaze tracking (Section 4.4).

Finally, high-level emotion estimation based on video data is discussed in Section 4.5 together with the results of the presented system on data from AV+EC 2016 Emotion Recognition Challenge. This document focuses only on the visual part since the whole multi-modal system is presented in the D4.8 (Adaptive Multilingual, Multimodal Fusion for Emotion Recognition, final version).

3 Low-level information

This section describes low-level (non-semantic) video processing parts of the system comprised of face detection and tracking, facial landmark localization, head pose estimation, face alignment, and body pose estimation.

3.1 Face detection and tracking

Face detection is the most basic building block for automatic human behavior understanding. First practical and real-time face detectors appeared after 2001 starting with the frontal face detector (Viola and Jones 2001). State-of-the-art face detectors provide high-quality real-time and multi-view detection, and are considered mature technology. The main competing methods for face detection are boosted cascades inspired by (Viola and Jones 2001), deformable part models (Pedro F Felzenszwalb et al. 2010), and emerging convolutional neural networks. These methods provide similar results and the detection quality is determined more by used training data and implementation details rather than the method family (Li et al. 2015) per se (Benenson et al. 2014; Mathias et al. 2014).

We have several real-time face detectors available, including those implemented in-house at Brno University of Technology (Herout, Hradiš, and Zemčík 2012; Juránek et al. 2015), and other state-of-the-art detectors (Dollár et al. 2009; Mathias et al. 2014).

We are currently using a detector based on a discriminatively trained deformable part model (P.F. Felzenszwalb, Girshick, and McAllester 2010) which is a collection of linear templates on Histograms of Oriented Gradient features with associated relative positions. This detector runs at 16 fps on an Intel Core i5-2500 CPU on 720p video while detecting faces larger than 80px. It is reliable for good quality frontal faces. If higher performance is needed, the detector can be exchanged for a frontal face detector based on
boosted Local Binary Features (Herout et al. 2010) which in our implementation runs at 250 fps on a GTX 980 GPU on 720p video.

Faces are tracked in a video according to standard tracking by detection approach – highly overlapping detections in consecutive frames are expected to belong to the same physical object and are merged into a single track. This approach is robust and it recovers trivially after the object is lost. However, tracks of the same person have to be linked separately. We use convolutional activations for this purpose as described later in Section 4.1.

Figure 2: Examples of facial landmark localization using a Convolutional Network regressor trained on AFLW dataset at Brno University of Technology inspired by (Sun et al., 2013).

3.2 Facial landmark localization

Facial landmarks (eyes, nose, mouth, and exact positions of their parts) are important cues for understanding of faces, and can be further used for head pose estimation, frontalization, identifications and many other tasks.

Two basic families of approaches for facial landmark localization exist; one family of methods uses geometrical and appearance models of faces which are fitted to a particular image. These models can be both 2D and 3D. Examples of such methods are various versions of Active Appearance Models, and methods using local part detectors together with probabilistic shape models (Belhumeur et al. 2013).

The other family of methods is purely appearance-based without an explicit shape model. An example of such methods is the Deep Convolution Network Cascade of Sun, Wang, and Tang (2013).
At the Brno University of Technology, we have replicated the work of Sun, Wang, and Tang (2013) with decent results when training on the Annotated Facial Landmarks in the Wild (AFLW) dataset (Kostinger et al. 2011). Examples of facial landmark localization using this method are shown in Figure 2.

The Annotated Facial Landmarks in the Wild (AFLW) dataset (Kostinger et al. 2011) contains over 24,000 real-world images of faces in various poses gathered from Flickr. The images are annotated with up to 21 facial landmarks and with roll, pitch, and yaw describing the head pose.

We are currently using face localization based on an ensemble of regression trees (Kazemi and Sullivan 2014) which provides good facial point localization at real-time speed even on a single core CPU. We localize 68 facial points and we use implementation available in dlib\(^1\) library.

\(^1\) http://dlib.net/
Figure 3: Cumulative error histograms of head pose estimation using random regression forests and three different features on a subset of AFLW dataset (5000 images). The y-axis represents the fraction of samples with error (yaw, pitch, and roll angles) less or equal to the value on the x-axis.

Figure 4: Error distribution over the range of yaw angle. The yellow curve represents the distribution of the training samples, blue curve represents the error when the training samples were selected randomly, and the red curve represents the error on samples with more evenly distributed yaw.
3.3 **Head pose estimation**

Head pose estimation from images is considered a mature technology at least for good quality images and frontal or profile views. It can be computed using a geometrical model form facial landmarks or it can be estimated directly from images using appearance-based classifiers. At the Brno University of Technology, we developed two appearance-based methods which are available for the MixedEmotions project. Both methods were trained on the AFLW dataset. One of the methods uses convolutional networks and the other uses Random Regression Forests.

The Random Regression Forests head estimation was originally developed for project A-PiMod\(^2\) (Behúň, Herout, and Pavelková 2015). It was trained on the AFLW dataset (Kostinger et al. 2011). Although, the head poses in the AFLW are estimated from the 21 hand-annotated facial landmarks – which may introduce small errors in the ground truth – the estimated orientations are precise enough for most practical applications.

The head pose estimation accuracy was initially evaluated on a random subset of the AFLW dataset and it is shown in Figure 3. For this experiment, 50 individual trees with depth of 12 were used on three different types of features: pixel intensities, integral channel features (Dollár et al. 2009), and Gabor wavelets (Daugman 1985). Figure 3 also shows errors in yaw estimation for different ground truth yaw angles. Clearly, the estimation is much more precise for near-frontal views and worse for near-profile views. One of the reasons for this behavior is that the AFLW dataset contains mostly near-frontal views (distribution of poses is shown as the yellow curve in Figure 4). When the same random forest regressor was trained on a balanced dataset, the overall error decreased.

The Random Regression Forests head estimation was verified in A-PiMod project on a custom dataset, in which the ground truth pose information is obtained by using OptiTrack motion tracking system\(^3\) with head-mounted TrackClipPRO active infrared positioning accessory. Examples from the dataset are shown in Figure 5. The yaw angle results for one video from the dataset, together with ground truth data from the OptiTrack sensor, are shown in Figure 6, and overall results from multiple video sequences are summarized in Table 1.

---

\(^2\) [http://www.apimod.eu/](http://www.apimod.eu/)
\(^3\) [https://www.naturalpoint.com/optitrack/](https://www.naturalpoint.com/optitrack/)

D4.6 Emotion Recognition from Image and Video Content, final version

Page 10 of 33
Figure 5: Collected simulator dataset for head pose estimation evaluation. In the second row, the head position is drawn into the images together with lines representing the ground truth for yaw and pitch angles (green lines) and the angles estimated by the Random forests (blue, red and yellow lines, for near-frontal, near-profile left and near-profile right detectors, respectively). The TrackClipPRO with three IR LEDs, which was used for orientation tracking to get the ground truth data, is visible on the left side of the subjects' heads.

Figure 6: Quality of head pose estimation on a single video sequence. The ground truth (green ‘+’) and estimated values of yaw angle for one video.

<table>
<thead>
<tr>
<th></th>
<th>Yaw</th>
<th>Pitch</th>
<th>Roll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall pose error (2RF)</td>
<td>0.1418</td>
<td>0.2316</td>
<td>0.0664</td>
</tr>
<tr>
<td>Near-frontal pose error (2RF)</td>
<td>0.0994</td>
<td>0.2309</td>
<td>0.0627</td>
</tr>
<tr>
<td>Near-profile pose error (2RF)</td>
<td>0.3286</td>
<td>0.2374</td>
<td>0.0825</td>
</tr>
<tr>
<td>Overall pose error (1RF)</td>
<td>0.1479</td>
<td>0.2313</td>
<td>0.0667</td>
</tr>
</tbody>
</table>

Table 1: Median of head pose estimation error in radians for video with OptiTrack ground truth data.

To assess how the results on AFLW generalize to data relevant to the MixedEmotions project, we have evaluated the head orientation estimation on Deutsche Welle Conflict Zone videos. We have selected 10...
latest Conflict Zone videos from which we manually sampled 200 frames with diverse head orientations. These frames were manually labeled with facial landmarks and a 3D face model was fitted to estimate head orientation. Examples from the dataset are shown in Figure 8. The results in Figure 7 show that the method generalized to this different type of content quite well.

Figure 7: Cumulative error histograms of head pose estimation using random regression forests and three different features on Deutsche Welle Conflict Zone Videos and AFLW dataset. The y-axis represents the fraction of samples with error less or equal to the value on the x-axis. Left) yaw, right) roll.
3.4 Face alignment

Many publications claim that precise facial alignment significantly improves facial recognition, e.g., the well-known person identification using Convolutional Neural Network (Taigman et al. 2014) largely relies on a complex 3D piece-wise affine transformation of Delaunay triangulation of 67 fiducial points and by replacing invisible triangles with respect to the camera through blending with their symmetrical counterparts. Alignment can be learned in a semi-supervised way without the need of facial point annotations or their detector (Huang et al. 2012); however, supervised alignment tends to provide more stable results. Some recent works skip face alignment and rely rather on large datasets and the modeling power of convolutional neural networks. For example, age estimation based on expectation over a distribution predicted by convolutional neural network (Rothe, Timofte, and Gool 2016) achieves state-of-the-art precision by aligning only the image position with maximal face detector response over position, scale, and rotation. The authors explained this decision by the fact that their facial point localization fails on...
approximately 5% of the images of their dataset, and consequently, it results in larger average error than the simple alignment. Precise alignment improves results but may be unstable and may lead to suboptimal performance due to alignment failures.

We chose an alignment approach from the middle of the available range. We use facial landmark localization as described in the previous section and a simple 2D geometric transformation from the detected landmark positions to an average face template. For example OpenFace (Amos, Ludwiczuk, and Satyanarayanan 2016) uses 2D affine transformation on a small set of facial points (e.g. outer eye corners and nose tip). However, the affine transformation tends to unnaturally deform the faces. We chose the similarity transformation as a reasonable compromise (translation, rotation, and scale). Random examples of aligned faces are shown in Figure 9. This alignment with the chosen facial feature localization method consistently improves results in facial expression recognition, identification, and age estimation as reported in the following sections.
Figure 9: Examples of aligned faces using similarity transformation.

3.5 Body pose estimation

Body pose estimation in the context of this document is a 2D localization of large human body parts in images and in videos. Body pose estimation from images and videos is a hot research topic and the state-of-the-art methods, in general, provide reasonably precise localization of body parts for good quality images and common poses. Most pose estimation methods model appearance of body parts and distribution of their relative positions in 2D. These methods include various part-based model with Pictorial Structure (Dantone et al. 2014). An interesting variant of such approach is the Pose Machine by (Ramakrishna et al. 2014)
which replaces the graphical model of relative part positions with regression trees and applies them recursively on their own body part estimations.

At the Brno University of Technology in the A-PiMod project (Behúň et al. 2015), we have developed a variant of the Pose Machine which is tuned to upper body pose estimation. It is able to localize 10 important upper body joints: head, shoulder center, right/left shoulder, right/left elbow, right/left wrist, and right/left hand. In our experiments, the Pose Machine achieved superior results compared to other state-of-the-art solutions of human pose estimation such DeepPose (Szegedy, Toshev, and Erhan 2013) and Pictorial Structure (Dantone et al. 2014).

The principle of Pose Machine is shown in Figure 10. Pose machine is a hierarchical method consisting of multiple stages. Each stage is modeled by a multiclass random forest which produces position probability maps for each body part. The position probability maps produced by one stage of Pose Machine is used as input for the next stage as spatial context features which allow the random forest to learn relationships among body parts, and thus, improve pose estimation accuracy for each body part. In fact, infinitely deep Pose Machine is equivalent to inference a fully-connected graphical model of the body part positions; however, the random forests with context features are able to capture much stronger dependencies among the parts and their appearances while the graphical models used in pictorial structures cannot.
Figure 10: Human pose estimation by hierarchical inference machine named Pose Machine. Input of the stage $t$ includes features are computed from the input image, and context features are computed from the output of the stage $t-1$ for each body part. Each stage is modeled by a random forest.

The human pose estimation based on Pose Machines is trained on a dataset which contains videos of 24 seated people in an office environment (Behúň, Herout, and Páldy 2014) (see Figure 11). The dataset contains 6,213 pose frames at 640x480 resolution with manually verified positions of 10 upper body joints. The same set of features (except for the output of a skin detector) was used as in the work by (Dantone et al. 2014) (resulting in 16 feature channels). Each stage contained 15 trees and the Pose Machine reached the best achievable performance when using 3 stages. The results are shown in Figure 12.

For videos, the per-frame joint position estimations can be smoothed using for example a particle filter. The A-PiMod dataset is fairly small and, as a consequence, the Pose Machines struggle to learn general body part appearance. If needed, the accuracy of this method could be significantly improved by increasing the size of the training dataset.
Figure 11: Examples from the A-PiMod pose estimation dataset with annotations.

Figure 12: Pose estimation results on the A-PiMod dataset. Three methods are compared – Convolutional Neural Network regression (DNN), Pictorial Structure (PS), and Pose Machines (PM). The graph shows what fraction of body parts is correctly located within increasing tolerance. The error tolerance is expressed as fraction of upper body size (distance between opposite shoulder and hip).
To assess how the trained upper body pose estimation generalizes to data relevant to the MixedEmotions project, we have evaluated the head orientation estimation on videos of political debates and videos from the First impression challenge dataset (Lopez et al. 2016). We have selected 5 political debate videos and 30 videos from the First impression challenge, and we randomly sampled and manually annotated 200 frames. Examples from the dataset are shown in Figure 13.

The results in Figure 14 show that the localization precision decreases on this new dataset, but the decrease is surprisingly small. As can be seen in Figure 15, the difference is mainly due to worse localization of joints representing head and neck. This is surprising as these two parts are localized more easily compared to the other joints. Part of this degradation can be explained by different head poses in the new dataset. The original dataset contains only nearly frontal faces while the new dataset contains profile faces and faces with significant pitch.

Figure 13: Examples from the MixedEmotions pose estimation test dataset.
Figure 14: Pose estimation results on the MixedEmotions dataset compared to the A-PiMod (oldData) dataset. The method applied is Pose Machines. The graph shows what fraction of the body parts is correctly located within increasing tolerance. The error tolerance is expressed as fraction of upper body size (distance between opposite shoulder and hip).
Figure 15: Pose estimation results on the MixedEmotions dataset compared to the A-PiMod (oldData) dataset. The method is Pose Machines. The error increased surprisingly and most significantly for head and shoulder center (results for the left body side are similar).
4 High-level information

This chapter reviews high-level processing modules including: assignment of video-wide consistent identities, age and gender estimation, expression estimation, and valence/arousal estimation.

4.1 Identification

To be able to process videos in which multiple people present, especially edited video, it is desirable to maintain identities across discontinuous face tracks. We extract convolutional activation features from individual video frames and compute simple similarities on these feature vectors. A new track is associated with an existing identity if the similarity surpasses a fixed threshold, otherwise a new identity is created. We assign identities only to face tracks containing at least some near-frontal faces as the features are not reliable for profile faces. This simple greedy approach performs well for videos with moderate number of identities (10-20).

The network we use to extract the features is based on (Parkhi, Vedaldi, and Zisserman 2015) and is fine-tuned on the Megaface dataset. We use cosine similarity as it proved to be the most suitable. Table 2 shows the effect of fine-tuning on Megaface and of alignment.

Table 2 Accuracy on Labeled Faces in the Wild.

<table>
<thead>
<tr>
<th></th>
<th>Original network</th>
<th>Affine alignment</th>
<th>Similarity alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>97,27</td>
<td>97,43</td>
<td>97,81</td>
</tr>
</tbody>
</table>

4.2 Age and gender estimation

The age estimation relies on a rather large convolutional network trained on the IMDB-WIKI dataset (Rothe et al. 2016). The dataset contains over 500 thousand images from pages of actors on IMDB which were automatically annotated. The annotations are based on two assumptions: (1) if there is only one person presented in an image, it is very likely s/he is the actor whose page the image was downloaded from; (2) if image caption contains a date, it is probably the date when the photo was taken. Of course, these assumptions are not always valid and the resulting dataset is very noisy.
To handle the noisy labels, Rothe et al. (2016) proposed to train a classifier predicting probabilities of different ages. They argue that a classifier with soft-max output layer and cross-entropy loss does not suffer as much from noisy labels as, for example, regressor with Euclidean loss layer which penalizes large errors severely.

A deep convolutional network VGG-16 (Simonyan and Zisserman 2014) is fine-tuned on IMDB-WIKI. The average absolute error of the network on MORPH 2 dataset is 3.1 years without alignment and 2.9 with facial alignment.

Gender estimation is also based on the work of Rothe et al. (2016), it is the same VGG-16 network fine-tuned on the IMDB-WIKI dataset, this time as binary (male/female) classifier.

Both the age and gender network can process 80 faces per second on GTX 780.

4.3 Expression estimation

Facial expressions are one of the observable expressions of emotions (most of the time). The expressions are directly observable and quantifiable (Ekman and Friesen 1978), and as such are much easier to automatically detect compared to emotions. Basic facial expressions can be detected similarly to the appearance attributes with reasonable accuracy using existing publicly available datasets. Such expression detectors are constrained by the existing datasets which contain mostly frontal faces and basic facial expressions (neutral, smile, sad, angry, surprise).

We have used data from the Kaggle Facial Expression Recognition Challenge which contains over 30,000 48x48 gray-scale face images which are classified into one of seven facial expression classes (Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral). See Figure 16 for examples from the dataset.

We have trained convolutional neural networks on the facial expressions dataset. The network has two convolutional layers with 128 3x3 filters, a 2x2 max pooling layer, two convolutional layers with 256 3x3 filters, another pooling layer, two fully connected layers with 1500 channels, and a final output layer with seven outputs corresponding to the individual facial expressions. The non-linearities are ReLU in all layers except the output one which is soft-max. The network was learned by optimizing cross-entropy loss.

function by Stochastic Gradient Descent with momentum. The network was regularized by a very small weight decay and dropout with probability 0.5 after the two fully connected layers.

The confusion matrix on a test dataset is shown in Table 3, and the results are compared to other published results in Table 4. Three different alignments were used for our experiments – no alignment, 2D affine alignment (Amos et al. 2016), and 2D similarity transformation alignment as described in Section 3.4. From these, the similarity transformation clearly provides the best results compared to the other transformations.

Figure 16 Examples from Facial Expression Recognition Challenge dataset.

Table 3 Confusion matrix of facial expression recognition. The lines correspond to real expressions and the columns to the predicted expressions.

<table>
<thead>
<tr>
<th></th>
<th>Angry</th>
<th>Disgust</th>
<th>Fear</th>
<th>Happy</th>
<th>Sad</th>
<th>Surprise</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>0.61</td>
<td>0.00</td>
<td>0.08</td>
<td>0.04</td>
<td>0.14</td>
<td>0.03</td>
<td>0.10</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.24</td>
<td>0.51</td>
<td>0.07</td>
<td>0.02</td>
<td>0.10</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>Fear</td>
<td>0.10</td>
<td>0.01</td>
<td>0.55</td>
<td>0.01</td>
<td>0.16</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>Happy</td>
<td>0.01</td>
<td>0.00</td>
<td>0.02</td>
<td>0.86</td>
<td>0.01</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Sad</td>
<td>0.13</td>
<td>0.00</td>
<td>0.12</td>
<td>0.03</td>
<td>0.54</td>
<td>0.02</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Table 4 Accuracy of different alignments on the Facial Expression Recognition Challenge dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>(Tang 2013)</th>
<th>(Ionescu, Popescu, and Grozea 2013)</th>
<th>Original alignment</th>
<th>Affine</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.711</td>
<td>0.675</td>
<td>0.640</td>
<td>0.636</td>
<td>0.705</td>
</tr>
</tbody>
</table>

4.4 Facial Action Unit detection and Eye gaze tracking

For rough eye gaze tracking and Facial Action Unit detection we use methods implemented in the OpenFace\(^5\) toolkit (Baltrušaitis, Robinson, and Morency 2016). The gaze tracking (Wood et al. 2015) uses Constrained Local Neural Field (CLNF) deformable model trained on a synthetically generated images of eyes. The Facial Action Unit detection (Baltrušaitis, Mahmoud, and Robinson 2015) relies on appearance features (HOG) and geometric features (CLNF facial landmark localization) together with Support Vector Machines and Support Vector Regression trained on merged SEMAINE, DISFA and BP4D datasets.

4.5 Video-based Emotion Recognition for AVEC 2016 Challenge

Automatic recognition of emotions from video has wide applications for example in human-computer interaction (Brown 2014), and recommendation systems (Berkovsky 2015). However the research in automatic emotion recognition is hampered by the available datasets which are usually focused on a specific scenario or situation, and therefore, existing methods trained on such datasets are hard pressed to generalize to other applications (e.g. RECOLA database (D4.7 Section 3) (Ringevaland et al. 2013)). Further, the task of emotion recognition is inherently hard since reliable ground truth is very hard to create due to the hidden nature and complexity of emotions. The available approaches and their performance on RECOLA database is summarized in document D4.7 (Emotion Recognition from Multilingual Audio Content, initial version) and D4.8 (Emotion Recognition from Multilingual Audio Content, final version).

---

\(^5\) https://github.com/TadasBaltrusaitis/OpenFace
D4.6 Emotion Recognition from Image and Video Content, final version
Page 26 of 33
RECOLA database contains spontaneous and naturalistic interactions collected during the resolution of a collaborative task that was performed in dyads and remotely through video conference. Multimodal signals, i.e., audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA), were synchronously recorded from 27 French-speaking subjects. The recordings were annotated with time-continuous ratings (40ms binned frames) of emotional arousal and valence by six gender-balanced French-speaking assistants for the first five minutes of all recordings. In the initial five minutes, participants discussed more about their strategy – hence showing emotions – at the beginning of their interaction.

Audio-Visual + Emotion Recognition Challenge (AV+EC 2016) (Valstar et al. 2016). AV+EC is an annual challenge held since 2011. Its main purpose is emotion recognition from multimodal data — audio, video and physiological data. Emotion is understood as a value in two-dimensional arousal-valence continuous space. The data comes with three sets of features for audio, video and physiological signals. This section presents alternative visual features and evaluates their performance.

We have complemented the baseline video features by activations of several convolutional neural networks (CNN) pre-trained for other tasks. The network CNN-lm was trained to localize facial landmarks on the AFLW dataset. This regression network has 4 convolutional layers followed by a fully connected layer with absolute hyperbolic tangent activation. A final fully connected layer outputs x and y coordinates of 5 facial landmarks. It is necessary to use a pre-trained network due to the very small size of the AVEC dataset, and the facial landmark localization task should be suitable for emotion recognition considering the good performance of the baseline geometric features on the valence task. We have extracted the activations of the last convolutional layer (CNN-lm-L4) and the first fully connected layer (CNN-lm-L5) from the provided baseline facial regions enlarged by a factor of 1.3 and rescaled to 40 × 40 pixels. The CNN-lm-L4 features should contain more appearance information while the CNN-lm-L5 should encode more geometric information.

Another network CNN-id is the identification network from Section 4.1. Using this network, we extracted activations of layer fc6 (CNN-id-fc6). Although this network is trained to identify people and it should discard facial expression information, it is still interesting to see how the network performs. These feature vectors are rather larger (4,096).
The third network was trained to recognize facial expressions (Section 4.3) and should be therefore very suitable for valence and arousal. We extracted activations of both fully connected layers of the network (CNN-em-fc5 and CNN-em-fc6).

The extracted per-frame activation features were further processed by a pipeline consisting of several operations whose parameters were selected by cross-validated grid-search for each feature type and separately for valence and arousal. At first, Principal Component Analysis (PCA) may be used for dimensionality reduction, and the resulting features are normalized to have zero mean and unit variance. In our experiments, we trained regression models for valence and arousal values for each frame (every 40ms). In many other classification and recognition tasks, we have seen the need of adding larger temporal context to make a good prediction. This context is different for each modality. We chose to provide the context primarily by stacking together features from a temporal neighborhood. The features themselves have quite smooth trajectories, so we do not need to take every frame but rather skip some frames, in order to keep the size of frame feature vectors manageable - we call it subsampling. Further context is provided by computing local statistics for each feature. We compute mean, variance, maximum, and minimum for each feature from a temporal window. Finally, we apply PCA again to reduce the size of feature vectors and we optionally normalize the features again to zero mean and unit variance. Finally, the features were shifted in time to account for annotation delays. All parameters of all the operations were obtained by grid search with the performance measured as CCC on the development partition of the AVEC 2016 database.

The regressor used in the experiments is linear with CCC objective function and weight decay. The weight decay strength is one of the parameters optimized by the grid search and the classifiers were trained using stochastic gradient descent with momentum.

The results on the training and development sets of AV+EC 2016 are show in Table 5. From the evaluated neural networks, best results are clearly achieved by the facial expression recognition network. Both fc5 and fc6 layers improve performance compared to the baseline features in the arousal task; however, the best result for valence are provided by the purely geometric baseline features.

Table 5 Performance on AV+EC training and development sets when trained on the other partition.

Features video-appearance and video-geometric were provided as baselines by the challenge organizers.
### 5 Conclusions

MixedEmotions focuses on the extraction of emotion from audio, video and text. This deliverable describes the available tools and methods for emotion recognition from video content, their performance, and how they are used within MixedEmotions video processing pipeline. The low level processing (face detection, facial landmark localization, pose estimation, and face alignment) provides a necessary foundation for the methods extracting higher level information. Together the system provides a wealth of information which can be used in fusion with other modalities in practical applications, including behaviour analysis and recommendation. The whole processing is able to run in real-time provided that suitable GPU is available.

<table>
<thead>
<tr>
<th></th>
<th>valence</th>
<th></th>
<th>arousal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>video-appearance train</td>
<td>0.368</td>
<td>0.365</td>
<td>0.165</td>
<td>0.395</td>
</tr>
<tr>
<td>video-appearance dev</td>
<td>0.365</td>
<td></td>
<td>0.395</td>
<td></td>
</tr>
<tr>
<td>video-geometric</td>
<td>0.522</td>
<td>0.579</td>
<td>0.516</td>
<td>0.388</td>
</tr>
<tr>
<td>CNN-lm-L4</td>
<td>0.185</td>
<td>0.335</td>
<td>0.485</td>
<td>0.498</td>
</tr>
<tr>
<td>CNN-lm-L5</td>
<td>0.476</td>
<td>0.419</td>
<td>0.495</td>
<td>0.559</td>
</tr>
<tr>
<td>CNN-id-fc6</td>
<td>0.311</td>
<td>0.382</td>
<td>0.457</td>
<td>0.560</td>
</tr>
<tr>
<td>CNN-em-fc5</td>
<td>0.495</td>
<td>0.512</td>
<td>0.510</td>
<td>0.532</td>
</tr>
<tr>
<td>CNN-em-fc6</td>
<td>0.482</td>
<td>0.498</td>
<td>0.533</td>
<td>0.585</td>
</tr>
</tbody>
</table>
References


Brown, Stephen. 2014. “Meet Pepper, The Emotion Reading Robot.” *TECHNOLOGY.*


Juránek, Roman, Adam Herout, Markéta Dubská, and Pavel Zemčík. 2015. “Real-Time Pose Estimation Piggybacked on Object Detection.” in *ICCV.*

Kazemi, Vahid and Josephine Sullivan. 2014. “One Millisecond Face Alignment with an Ensemble of Regression Trees.” in *CVPR.*


(CVPR).


