Automatic Recognition of Personality Traits: A Multimodal Approach

Maxim Sidorov, Stefan Ultes, Alexander Schmitt
Institute of Communications Engineering
University of Ulm
Ulm, Germany
{maxim.sidorov, stefan.ultes, alexander.schmitt}@uni-ulm.de

ABSTRACT
A system being capable of recognize personality traits may be utilized in an enormous number of applications. Adding personality-dependency may be useful to build speaker-adaptive models, e.g., to improve Spoken Dialogue Systems (SDSs) or to monitor agents in call-centers. Therefore, the First Audio/Visual Mapping Personality Traits Challenge (MAPTRAITS 2014) focuses on estimating personality traits. In this context, this study presents the results for multimodal recognition of personality traits using support vector machines. As only small portions of the data is used for personality estimation at a time (which are later combined to a final estimate), different segmentation methods (and how to derive a final hypothesis) are analyzed regarding the task as both a regression and a classification problem.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Metrics—complexity measures, performance measures

Keywords
Personality traits, audio-visual features, feature-based fusion, support vector machine.

1. INTRODUCTION

Automatic recognition of personality traits based on audio and visual signals is in the focus of research groups all over the world. While enabling machines to recognize personality traits may be useful in various applications, e.g., user-specific system behaviour, the recognition performance is still not satisfying. Meanwhile, computer systems are already deeply involved in recognizing, interpreting, or even synthesizing human emotions and another paralinguistic aspects. For instance, Affective Computing [11] is one of the fields which deal with human affects.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MAPTRAITS '14, November 12, 2014, Istanbul, Turkey.
Copyright © 2014 ACM 978-1-4503-0480-1/14/11...$15.00.
http://dx.doi.org/10.1145/2668024.2668028

The following personality and social aspects are under examination in the study: the Big Five personality traits (extraversion, agreeableness, conscientiousness, neuroticism, and openness) and four additional social dimensions. These social dimensions are engagement, facial attractiveness, vocal attractiveness and likability.

The Big Five personality traits are five broad factors of personality that are used to describe human personality. Thus, openness reflects the degree of curiosity, adventure, appreciation for art, and so forth. Conscientiousness shows a tendency to be organised and dependable, preference to act according to the plan rather than spontaneously. Extroversion is a level of sociability, tendency to express positive emotions and energy. Agreeableness shows a tendency to be cooperative rather than antagonistic with others. Neuroticism is a tendency to experience unpleasant emotions and situation easily. Four additional social dimensions include engagement (how engaged the person appears in the interaction), facial attractiveness (how attractive the person appears based on the face), vocal attractiveness (how attractive the person appears based on the voice) and likability (how one likes the person in the given context).

An increasing number of scientific work considers different ways of paralinguistic-based recognition. Since different corpora or performance metrics are used, these approaches are not always comparable. Therefore, challenges as MAPTRAITS provide a general framework enabling objective comparison and thus setting up state-of-the-art approaches. Further challenges include the INTERSPEECH 2012 Speaker Trait Challenge [16], the Audio/Visual Emotion recognition Challenges (AVEC) in their first [18], second [19], and third [24] edition, the first and second Emotion Recognition in The Wild Challenges (EmotiW) [6], and the Computational Paralinguistics Challenge (ComParE) [17].

The four key components being addressed by most approaches on multimodal paralinguistic recognition are the audio and the video feature extraction procedure, the modeling algorithm, and the fusion techniques. Scientific groups have used a wide variety of approaches and their combinations which will be briefly presented in the following.

For audio feature extraction, a variety of Low-Level Descriptors (LLDs) and their functionalities has been used, based on their successful applications on problems as automatic speech recognition or speaker and gender identification. Several functional-based feature sets have been created especially for their use in paralinguistic recognition challenges. Thus, feature sets of 384 and 1,582 functionals have been proposed at the INTERSPEECH 2009 Emotion challenge [14]...
and the INTERSPEECH 2010 Paralinguistic Challenge [15] respectively. The latest audio-based feature set consisting of 6,373 features has been suggested for the ComParE [17]. All of these features can be easily extracted out of speech waveforms using openSMILE software [7]. S’anchez-Lozano et al. [12] used three different audio feature sets for affect recognition: 16 MFCCs with their delta and acceleration (48), energy and spectral features (34), and 2268 functionals which have been derived from LLDs. Busso et al. [3] have achieved reasonable results in emotion recognition using the following set of audio-based features: the means, the standard deviations, the ranges, the maximum values, the minimum values and medians of the pitch and the energy. Regarding the learning stage of emotion recognition using audio cues, the most common algorithms applied are Multi Layer Perceptron (MLP) [22] and Support Vector Machine (SVM) [9], as well as their combination [25].

Classical feature sets for the visual part including Local Binary Patterns (LBP) [1], Gabor, LPQ, PHOG [5] and their combination [23] have been successfully utilised for facial expression recognition and face detection. Recently suggested LPQ-TOP and LGBP-TOP [2] have also been successfully used for emotion recognition. Concerning the learning phase of visual-based emotion or affect recognition, the most common method used is the SVM for both classification and regression tasks.

Several fusion techniques may be utilised for merging the multimodal information. Thus, audio-visual affect personality traits recognition are based on three types of fusion methods: feature-level fusion (also called early fusion), decision-level fusion and model-level fusion [20, 26]. Feature-level fusion can be implemented by combining the features of different modalities into a single feature set. Feature-level fusion has the disadvantage of combining different types of modalities and their different metric levels. Furthermore, a high dimensionality of the resulting feature set can also pose a serious problem. In decision-level fusion, each input stream of the multimodal data is treated independently and the single-modal hypotheses are combined at decision level using some criterion (weighted sum, etc.). Decision-level fusion overcomes the problems of feature-level fusion but it is based on the assumption that the data for each modality is independent of other modalities. However, that is not necessarily true in reality. The Hidden Markov Model, in which different components are responsible for different data sources, represents an example of model-level fusion.

We present a multimodal personality traits recognition system using SVMs (trained by the Sequential Minimal Optimization (SMO) algorithm [8]) and feature-level fusion. For this, the data must be segmented into smaller chunks. Hence, a number of segmentation methods has been explored and their corresponding results have been analysed for the above presented nine personality aspects. To derive a final hypothesis out of the segment hypotheses, two different aggregation methods have been analyzed. Moreover, the performance on the different modalities as well as on the combined feature set has been evaluated.

The rest of the paper is organized as follows: Section 2 describes MAPTRAITS’14 corpus; Section 3 reviews the audio and video feature extraction methods as well as segmentation techniques; the proposed system is described in Section 4; Section 5 presents the experiments and results obtained with the test set; finally, Section 6 presents the conclusions and future work.

2. CORPUS DESCRIPTION

For all experiments within the MATPRAITS challenge [4], a subset of the SEMAINE [10] database consisting of 44 videos, which are in four situational contexts of 11 subjects. These 44 clips have been assessed by 6 raters along the 9 dimensions both with visual-only and audio-visual modalities. The dimensions were scored on a Likert scale with ten possible values mapped into the range from 1 to 10, from strongly disagree to strongly agree. For each setting (audio-visual and visual-only), the ground truth labels were generated by taking the average of the ratings per clip and dimension. The recordings have been curtailed to 15 seconds and divided into a training (6 subjects, 24 videos) and a testing (5 subjects, 20 videos) set.

3. FEATURES

The recordings in the challenge are quite long, therefore segmentation must be applied. The organisers of the challenge have been realized multimodal audio-visual features for a variety of segments in order to perform experiments. A brief description of the suggested features and corresponding segmentation methods are in the following Section.

3.1 Audio Features

The audio part of the recordings has been characterised by a 6,376 dimensional feature vector. The following set of audio features has been selected to represent the challenge recordings: energy- and spectral-based LLDs, voicing LLDs, delta of energy/spectral features, delta of voicing features, and voiced/unvoiced duration features. The LLDs cover the state-of-the-art features of emotion recognition such as loudness, MFCCs (1-16), jitter, shimmer, energy in different bands, and probability of voicing. Functional features comprise the most descriptive functional dependencies of LLDs: standard deviation, quartiles, linear regression slope and quadratic regression coefficient a with corresponding approximation errors, mean, maximum, minimum, etc.

A number of segmentation schemes has been proposed in order to conduct the experiments. Full segmentation uses only one average feature vector per recording. The second segmentation, called lld, includes a set of LLDs-based features which have been extracted each 0.01 sec. The next set of segmentation methods consider overlapping segments of 2 and 4 seconds shifted towards at a rate of 0.5 and 1 second (2s0.5s, 2s1s, 4s0.5s and 4s1s, correspondingly).

3.2 Video Features

For the visual part, Quantised Local Zernike Moments (QLZM) [13] have been utilised as video-based features. Face localisation by detecting 49 landmark points per frame has been performed with the publicly available Xiong and De la Torre face detector. The face is divided into subregions by applying a 5x5 outer grid and a 4x4 inner grid which yielded a 656-length feature vector. Another video-based feature set has been created by determining the two eye areas and the mouth area, and extracting QLZM features on these parts. For part-based representation, the eye and mouth areas have not been divided into subregions. This resulted in feature vectors having a length of 48.
Table 1: Results of personality traits recognition, test partition, audio-only modality. The bold values outperform the baseline [4] results.

<table>
<thead>
<tr>
<th>Label</th>
<th>Mode-based aggregation</th>
<th>Average-based aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feat.</td>
<td>MSE</td>
</tr>
<tr>
<td>EN</td>
<td>2s0.5s</td>
<td>1.05</td>
</tr>
<tr>
<td>VO</td>
<td>4s0.5s</td>
<td>0.6</td>
</tr>
<tr>
<td>EX</td>
<td>4s0.5s</td>
<td>0.65</td>
</tr>
<tr>
<td>AG</td>
<td>4s0.5s</td>
<td>1.75</td>
</tr>
<tr>
<td>CO</td>
<td>4s0.5s</td>
<td>0.4</td>
</tr>
<tr>
<td>NE</td>
<td>4s0.5s</td>
<td>3.4</td>
</tr>
<tr>
<td>OP</td>
<td>2s0.5s</td>
<td>0.7</td>
</tr>
<tr>
<td>LI</td>
<td>2s0.5s</td>
<td>1.8</td>
</tr>
</tbody>
</table>

4. PERSONALITY TRAITS RECOGNITION WITH MULTIMODAL FEATURES

A Support Vector Machine (SVM) trained using the Sequential Minimal Optimization (SMO) algorithm [21] has been deployed for solving the regression tasks.

For performance evaluation, a number of metrics has been proposed, namely, Mean Square Error (MSE), the sample Pearson’s correlation coefficient (COR), coefficient of determination (R²), and unweighted average recall (UAR). Let $y_k$ and $\hat{y}_k$ be the ground-truth and predicted values, respectively, MSE and COR can be defined as:

$$MSE = \frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{y}_k)^2,$$

(1)

$$COR = \frac{\sum_{k=1}^{N} (y_k - \mu_{yk}) (\hat{y}_k - \mu_{yk})}{\sqrt{\sum_{k=1}^{N} (y_k - \mu_{yk})^2} \sqrt{\sum_{k=1}^{N} (\hat{y}_k - \mu_{yk})^2}},$$

(2)

where $N$ is the number of predictions, $\mu_{yk}$ and $\mu_{\hat{yk}}$ are the sample mean of the ground-truth and the predicted values, respectively. $R^2$ and UAR are defined as

$$R^2 = 1 - \frac{\sum_{k=1}^{N} (y_k - \hat{y}_k)^2}{\sum_{k=1}^{N} (y_k - \mu_{yk})^2},$$

(3)

$$UAR = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{\sum_{k=1}^{N_i} \delta(y^{i,k}, \hat{y}^{i,k})}{N_i},$$

(4)

where $N_c$ is the number of classes, $y^{i,k}$ and $\hat{y}^{i,k}$ are the ground-truth and the predicted values of $i$-th class, correspondingly. $\delta(x, y)$ is equal to 1 if $x = y$, 0 otherwise.

Since there is only one ground-truth label per recording for the paralinguistic measurements, the same label has been duplicated so that the number of features and the labels are equal. Analogically, there are 50 visual-based features per second; as the audio-based functional features have been obtained at different rates, average values of corresponding visual features have been calculated in order to equal the rates of the audio- and visual-based features for performing feature-level fusion.

For conducting the experiments on the train set, a speaker-based leave-one-out (leave-one-speaker-out) cross-validation strategy has been chosen. In order to figure out the optimal segmentation type, all segmentation methods have been tested with the train partition. Furthermore, a segmentation method which has shown the lowest MSE value with the training set has been applied for training portion (having a model trained with the whole training data set).

To fit the visual features to the audio features, the visual features have been averaged correspondingly. Thus, in case of the full segmentation method, only one average feature vector has been calculated. For the 2s1s method, every 100 visual features have been averaged regarding the constant overlapping. Analogically, a similar procedure has been applied for the rest of the segmentation techniques. The given label has been multiplied for all segments of the corresponding recording. In case of visual-only and audio-visual modalities, 48- and 656-dimensional visual feature sets have been evaluated as well.

Firstly, audio and video features have been treated separately, then the feature-based fusion technique without further feature selection or reduction methods has been utilised.

In order to rescale the attributes of different units and scales, the statistical normalization procedure $Z$-transformation has been applied as a preprocessing step.

For the paralinguistic hypotheses, the SVM has been evaluated on the training set using leave-one-speaker-out cross-validation. Hence, one single SVM has been created addressing the personality traits of every speaker in the database.

In order to produce one final hypothesis for each recording, two different aggregation techniques have been examined. During the mode-based aggregation technique, the label which appears most often in a set of segment-based hypotheses has been set as a final hypothesis. When the average-based aggregation method is applied, the average value over all segment-based hypotheses has been calculated. This average value has then been rounded to the nearest integer number which has been set as final hypothesis.

The result of the described experiments are in the following Section.

5. EXPERIMENTAL RESULTS

The results of personality traits recognition using the audio-only, visual-only and audio-visual feature set, with the test set, are in Table 1, Table 3 and Table 3, respectively. There, the label columns correspond to the first two letters of the corresponding paralinguistic measurement, which were described in Section 1. Feat. columns denote the optimal segmentation technique (i.e., the segmentation method result-
Table 2: Results of personality traits recognition, test partition, visual-only modality. The bold values outperform the baseline [4] results.

<table>
<thead>
<tr>
<th>Label</th>
<th>Dim.</th>
<th>Feat.</th>
<th>MSE</th>
<th>COR</th>
<th>R2</th>
<th>UAR</th>
<th>Dim.</th>
<th>Feat.</th>
<th>MSE</th>
<th>COR</th>
<th>R2</th>
<th>UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>48</td>
<td>lld</td>
<td>0.94</td>
<td>0.12</td>
<td>-0.03</td>
<td>0.2</td>
<td>48</td>
<td>lld</td>
<td>0.94</td>
<td>0.12</td>
<td>-0.03</td>
<td>0.2</td>
</tr>
<tr>
<td>FA</td>
<td>656</td>
<td>full</td>
<td>2.5</td>
<td>-0.05</td>
<td>-0.7</td>
<td>0.25</td>
<td>48</td>
<td>2s1s</td>
<td>3.17</td>
<td>0.2</td>
<td>-1.15</td>
<td>0.4</td>
</tr>
<tr>
<td>EX</td>
<td>656</td>
<td>full</td>
<td>0.89</td>
<td>0.23</td>
<td>-0.53</td>
<td>0.27</td>
<td>656</td>
<td>full</td>
<td>0.89</td>
<td>0.23</td>
<td>-0.53</td>
<td>0.27</td>
</tr>
<tr>
<td>AG</td>
<td>48</td>
<td>lld</td>
<td>1.83</td>
<td>-0.04</td>
<td>-0.65</td>
<td>0.23</td>
<td>48</td>
<td>lld</td>
<td>1.83</td>
<td>-0.04</td>
<td>-0.65</td>
<td>0.23</td>
</tr>
<tr>
<td>CO</td>
<td>48</td>
<td>2s1s</td>
<td>0.89</td>
<td>0.07</td>
<td>-0.28</td>
<td>0.29</td>
<td>48</td>
<td>2s1s</td>
<td>1.17</td>
<td>-0.13</td>
<td>-0.68</td>
<td>0.29</td>
</tr>
<tr>
<td>NE</td>
<td>48</td>
<td>lld</td>
<td>5.5</td>
<td>-0.48</td>
<td>-0.99</td>
<td>0.09</td>
<td>48</td>
<td>lld</td>
<td>5.17</td>
<td>-0.34</td>
<td>-0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>OP</td>
<td>48</td>
<td>lld</td>
<td>0.83</td>
<td>0.12</td>
<td>-0.87</td>
<td>0.27</td>
<td>48</td>
<td>lld</td>
<td>0.83</td>
<td>0.12</td>
<td>-0.87</td>
<td>0.27</td>
</tr>
<tr>
<td>LI</td>
<td>48</td>
<td>2s1s</td>
<td>2.33</td>
<td>-0.22</td>
<td>-0.89</td>
<td>0.13</td>
<td>48</td>
<td>2s1s</td>
<td>2.11</td>
<td>-0.02</td>
<td>-0.71</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 3: Results of personality traits recognition, test partition, audio-visual modality. The bold values outperform the baseline [4] results.

<table>
<thead>
<tr>
<th>Label</th>
<th>Dim.</th>
<th>Feat.</th>
<th>MSE</th>
<th>COR</th>
<th>R2</th>
<th>UAR</th>
<th>Dim.</th>
<th>Feat.</th>
<th>MSE</th>
<th>COR</th>
<th>R2</th>
<th>UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>48</td>
<td>4s1s</td>
<td>1.06</td>
<td>0.23</td>
<td>-0.38</td>
<td>0.24</td>
<td>48</td>
<td>2s1s</td>
<td>1.06</td>
<td>0.23</td>
<td>-0.38</td>
<td>0.24</td>
</tr>
<tr>
<td>VO</td>
<td>48</td>
<td>2s1s</td>
<td>0.5</td>
<td>0.0</td>
<td>-0.6</td>
<td>0.27</td>
<td>656</td>
<td>full</td>
<td>0.67</td>
<td>0.2</td>
<td>-1.14</td>
<td>0.32</td>
</tr>
<tr>
<td>FA</td>
<td>48</td>
<td>2s1s</td>
<td>1.72</td>
<td>0.31</td>
<td>-0.51</td>
<td>0.21</td>
<td>656</td>
<td>4s1s</td>
<td>1.83</td>
<td>0.0</td>
<td>-0.61</td>
<td>0.25</td>
</tr>
<tr>
<td>EX</td>
<td>656</td>
<td>2s1s</td>
<td>0.61</td>
<td>0.37</td>
<td>0.12</td>
<td>0.24</td>
<td>656</td>
<td>2s1s</td>
<td>0.61</td>
<td>0.37</td>
<td>0.12</td>
<td>0.24</td>
</tr>
<tr>
<td>AG</td>
<td>656</td>
<td>2s0.5s</td>
<td>1.22</td>
<td>0.44</td>
<td>-0.08</td>
<td>0.34</td>
<td>656</td>
<td>4s1s</td>
<td>1.28</td>
<td>0.3</td>
<td>-0.13</td>
<td>0.28</td>
</tr>
<tr>
<td>CO</td>
<td>48</td>
<td>2s1s</td>
<td>0.56</td>
<td>-0.15</td>
<td>-1.22</td>
<td>0.28</td>
<td>48</td>
<td>2s1s</td>
<td>0.33</td>
<td>0.08</td>
<td>-0.33</td>
<td>0.31</td>
</tr>
<tr>
<td>NE</td>
<td>48</td>
<td>4s0.5s</td>
<td>3.89</td>
<td>-0.29</td>
<td>-1.06</td>
<td>0.11</td>
<td>48</td>
<td>4s1s</td>
<td>3.33</td>
<td>-0.13</td>
<td>-0.76</td>
<td>0.14</td>
</tr>
<tr>
<td>OP</td>
<td>48</td>
<td>2s1s</td>
<td>0.83</td>
<td>-0.25</td>
<td>-0.81</td>
<td>0.29</td>
<td>48</td>
<td>2s1s</td>
<td>0.83</td>
<td>-0.25</td>
<td>-0.81</td>
<td>0.29</td>
</tr>
<tr>
<td>LI</td>
<td>656</td>
<td>full</td>
<td>1.44</td>
<td>0.45</td>
<td>-0.16</td>
<td>0.38</td>
<td>656</td>
<td>full</td>
<td>1.44</td>
<td>0.45</td>
<td>-0.16</td>
<td>0.38</td>
</tr>
</tbody>
</table>

A wide variety of segmentation methods has shown the best performance during training phase. It can be concluded that the optimal choice of segmentation method completely depends on the recognised personal trait and corresponding metric. However, the difference between the results achieved by applying different segmentation techniques might be significant. Furthermore, despite the fact that 656-dimensional visual features contain more information, the 48-dimensional feature set has been the optimal choice in most of the cases. Therefore, not all of the features from the extended set are relevant and the most important information is extracted from the eye and mouth regions.

6. CONCLUSION

We have presented an audiovisual system for recognition of personality traits. The modalities have been treated separately, as well as the merged features have been investigated. A number of proposed segmentation techniques have also been examined along with two different aggregation methods for assigning a final hypothesis.

While the SVM already provides reasonable results for audio-visual personality recognition, we are still examining its general appropriateness. The usage of other possibly more accurate identifiers may improve the performance of this system. In addition, by applying feature selection techniques (principal component analysis, genetic algorithm-based feature selection, etc.) for these problems, further performance improvement could be achieved.
7. REFERENCES


