Apparent Personality Prediction using Multimodal Residual Networks with 3D Convolution

Bachelor Thesis Artificial Intelligence

Stefan Iacob – S4575121

Supervisors:

prof. dr. M.A.J. van Gerven

G.E.H. Ras, MSc

July 2018

Radboud University

ABSTRACT

In this thesis we propose a 3D apparent personality prediction model as extension of the multimodal residual neural network used for first impression analysis by Güçlütürk et al. [1]. The original model was trained on audio-visual data from YouTube videos and predicts the Big Five personality traits of the people in the video. The auditory data and the visual data were randomly selected within a clip, and thus not synchronized. The novel contribution of this research is to study the effect of extending the visual information over multiple frames, and of synchronizing the two modalities on the performance of the model. The model architecture was adapted to include these changes, and several new models were trained. Each performed better than the baseline models trained on the same dataset. Moreover, we provide evidence that temporal information improves the performance. However, a different network architecture is needed to prove the effect of the synchronization.

TABLE OF CONTENTS

Abstract 2

1 Introduction 3

2 Background and Related Work 6

3 Methods 8

3.1 Baseline Model 9

3.1.1 Data Preprocessing 9

3.1.2 Model Architecture 9

3.1.3 Network Training 10

3.2 3D Deep Impression Model 11

3.2.1 Data Preprocessing 11

3.2.2 Model Architecture 13

3.2.3 Network Training 15
3.3 Technical Specifications 16
4 Results 17
4.1 Training 17
4.2 Testing 18
4.3 Evaluation 21
5 Conclusion 23
5.1 Contribution 23
5.2 Limitations 24
5.3 Future work 25
6 References 26

1 INTRODUCTION

The combination of multiple modalities in the brain is crucial for humans to make sense of the world. Parts of the brain such as the parietal cortex are dedicated to integrating sensory information of different modalities [1]. Hence, conceptual representations are formed, which contain more information than just the separate sensory streams would provide. The robustness, as well as the failure of such mechanisms can be observed for example in the McGurk effect [2]. In this experiment, subjects have to report what syllable they hear. When the visual information shows a person pronouncing a different syllable, they fail to hear the correct syllable. This suggests that visual information is also included in the decision, even though the question is solely about auditory information. Moreover, ambiguous sounds can be interpreted in multiple ways, depending on the available visual information. Thus, valuable information is added by a second modality. The existence of such neuronal processes provides a strong biological motivation for exploring the possibilities of designing multimodal neural

networks. Another important reason for developing and improving multimodal neural networks is simply the great availability of information on different modalities. For example, in video mining problems a second synchronized modality is usually available: the sound in a video.

Many researchers have been making use of multimodal machine learning techniques. Some applications, such as the models in [3] and [4] on apparent personality classification, have achieved high accuracy in predicting several apparent personality traits based on audiovisual data of faces. Certain personality features appear to be important predictors for success in different job types and are thus likely to be considered during interviews. For example, high levels of conscientiousness are typical of an individual that is likely to be dutiful and responsible - a trait that is valuable for many positions [5]. According to the Big Five model, human personality can roughly be characterized by five traits: openness, extraversion, agreeableness, conscientiousness, and neuroticism [6]. There is evidence that at least four of the Big Five traits can be predicted accurately from facial features [7].

Moreover, apparent personality and first impressions are of great importance during job interviews and have significant influence on interviewers (at the very least on a subconscious level), and on their decisions in particular in situations when candidates have comparable qualifications. This is why neural network models could provide interviewers with a secondary, more objective measure, to aid in the decision of hiring that candidate. The advantage of such models is that they are trained on an apparent personality data set labeled by a large number of people, which eliminates to a large extent any trace of bias and/or subjectivity in the assessments produced by the model. This ensures that the objectivity of such models is preserved. Furthermore, the model can be applied consistently and uniformly to all applicants, irrespective of the context of such job vacancies. Furthermore, there is some evidence that predicted levels (using a CNN model) of personality traits, such as “Rule-consciousness”, “Openness”, “Perfectionism”, and “Tension” are correlated with self-reported personality as shown in [8], which strengthen the above objectivity claim.

The audiovisual model proposed by [3] achieved an accuracy of 0.9109 in predicting the Big Five traits from audiovisual information of people facing a camera. In this study, we attempt to improve this model by revising and strengthening the biological plausibility of the neural

network architecture described in [3]. To this end, the problem we address is in regard to the timing of audio and video. To integrate audio and video, [3] used a random fragment of the audio data and a random frame of the video data, which were fed through separate streams in the neural network and concatenated in the last layer. The audio contains temporal information, whereas the visual input only represents a single snapshot in time. We argue that this design choice leaves room for improvement, as this approach does not take into account the possibility of additional bi-modal and temporal patterns, such as correlations between voice intonation, and facial movements. This limitation primarily has two causes. First, only one frame is picked from the video data, hence most of the temporal information is lost. Secondly, the audio and video modalities are not synchronized: they can span different time fragments and can thus contain different emotional expressions. This motivates the underlying assumption of this study, namely the usage of a longer video sequence, synchronized with the corresponding audio fragment. This means the audio and video data span the same time interval.

Given the above considerations, the main research question is formulated as follows:

Does using multiple video frames per sample along with a synchronization of the audio and the video streams improve the performance of the model by Güçlütürk et al. [3] *in predicting Big Five personality traits?*

The hypothesis put forward, and the novel contribution of the current study is that this adaptation will increase biological plausibility, and subsequently, the performance of the model by exploiting the additional temporal and bimodal information included in a sequence of video frames. Here, we define temporal information as the additional information present in the changes between the videoframes, and bimodal information as the information that is added by certain combinations of audio features and video features. We test this hypothesis by designing, implementing and testing a *3D-Apparent Personality Prediction* (3D-APP) model, which is an extension of the Güçlütürk et al. [3] model.

Concerning the research methodologies followed during this research, it should be noted the we combined systematic literature review (to ensure that we have an exhaustive coverage of all

relevant literature) as prescribed by [9] and [10], with the principles of design science research, as formulated by [11], and [12].

The remainder of this document is organized as follows. Section 2 gives an overview of the relevant literature for this research. In Section 3, we describe the steps of the research process, the design and training of the baseline model (Section 3.1), and of the 3D-APP model (Section 3.2). More precisely, in the new 3D-APP model the overall architecture, training process, and data of the baseline model have been preserved, in order to make possible the performance comparison of the original and new models. However, as it will be explained in Section 3.2, the 3D-APP model introduces some key differences related to the way the audio and video data samples are used. The experiments, tests, and performance evaluation of the 3D-APP model are discussed in Section 4. We conclude this report with a summary of the main contribution, a discussion of the results and limitations of the proposed solution, and some pointers to future work.

2 Background and Related Work

In this section we discuss the literature and topics that are relevant to this research.

Multimodal Machine learning is the research area focusing on the development of models that are capable of processing information from multiple modalities [13]. The simultaneous processing of different modalities helps create associations between them, which should result in a better classifier performance compared to using single modalities. Multimodal neural networks have been used for a number of problems, such as language models [14], image captioning [15], emotion detection [16] and apparent personality classification [3] [4] [17], which is also the focus of this thesis. However, some interesting questions are still unanswered, such as how to combine in a meaningful way the different modalities in the network, or how, and at what point, to join the modality-specific networks into a single multi-modal network, to achieve a meaningful data fusion [13].

As mentioned earlier, the focus of the current research lies in the adaptation of the model proposed in [3]. The work in [3] was largely motivated by *Chalearn Looking at People* challenge

[18], in which it won the third place. The changes proposed in this research are partly present in the model proposed in [4], which ended-up in the second place in the same challenge. Subramaniam et al. [4] select multiple data frames from a video and perform a 3D convolution on them, hence implementing the same idea of temporally ordered frames. However, our approach essentially differs from [4] in the way audio data is preprocessed. Instead of using raw audio input, in [4] several features are extracted, which are then used to train the network. These features represent global characteristics of the whole audio fragment, so the temporal patterns in the audio channel are not explicit. Thus, no correlation between the audio and video patterns is possible. We argue that, when feeding the network raw audio data, no restrictions are posed on the relevant feature extraction, and hence such correlations are possible.

Regarding the design of the network architecture, we make use of the Residual Neural Network framework defined in [19]. The building blocks of a residual network are called *Residual Blocks*. A schematic representation of a residual block can be seen in Figure 1. Residual networks enable a much deeper network architecture while avoiding the degradation problem described in [19].

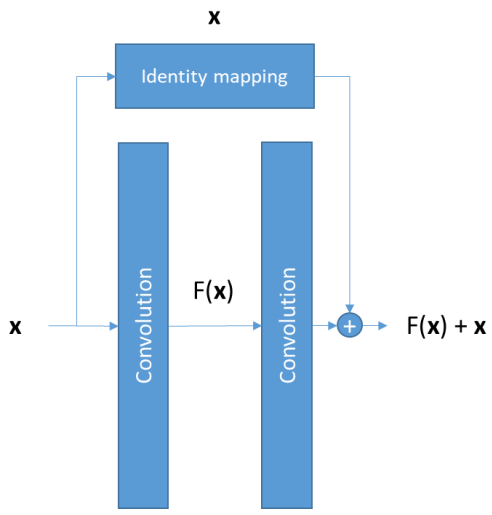


Figure 1. Residual block. The output of each residual block is added to an identity mapping of the input.

3 METHODS

In this section, we describe the steps of the general approach we followed for designing the architecture of the 3D-APP model, carrying out the data processing, and conducting the experiments. The neural network underlying the 3D-APP model is an adaptation of the residual neural network designed by [3], which is also used as baseline model for the purpose of performance benchmarking. The overall approach is shown in Figure 2 (in the form of a simple workflow diagram) together with the sections where the respective steps have been discussed.

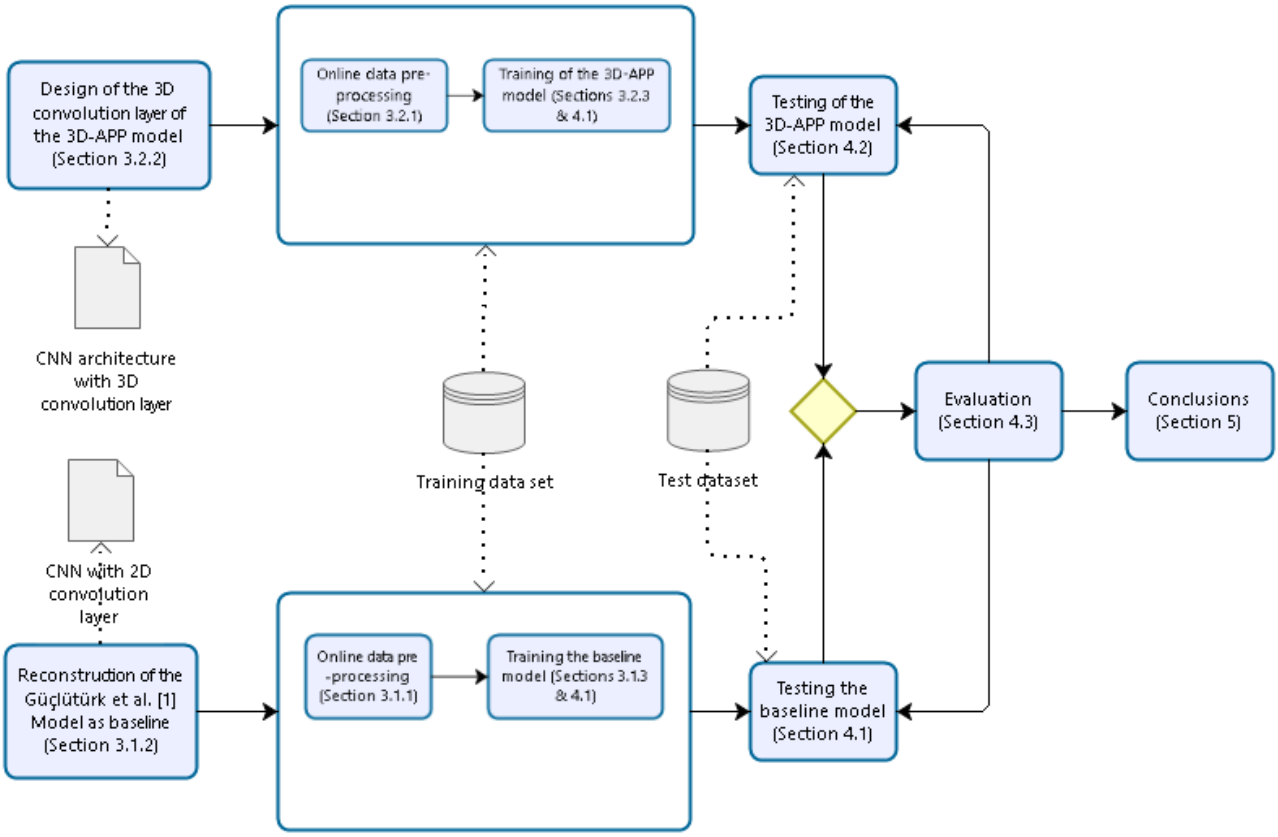


Figure 2. Research approach

3.1 BASELINE MODEL

In this subsection the data input, model architecture, and training of the baseline model is described.

3.1.1 Data Preprocessing

The dataset used for the baseline model consists of 10000 mp4 files, with an average duration of 15 seconds. The videos show people speaking English, while facing the camera. The training, validation and test ratio was 3:1:1. The data was labeled in terms of the Big Five personality traits, with each trait value ranging from 0 to 1. For each of the clips, the data set was split into visual input and audio input. These pairs were used as single training samples, fed through the network in mini-batches of 32 samples. The audio was resampled to a sampling frequency of 16000 Hz. For each audio clip, a random temporal crop of 50176 samples was made, corresponding to a time frame of 3.136 seconds. To obtain the visual input, a random frame was selected from the clip. A random 224 by 224 pixels spatial crop was taken from this frame. Güçlütürk et al. [3] argue that selecting random crops is the best approach, as this makes no assumptions on the importance of the different locations of visual patterns in the frame. For example, video background could also contain some information about the personality of the individual featured by that video sequence.

3.1.2 Model Architecture

The baseline architecture proposed in [3] consisted of audiovisual models and language models. For the purpose of this research, only the former is relevant. The audiovisual model consists of a visual stream and an audio stream. Both are 17 layers deep residual neural networks. They are joined together in a final fully connected layer. Each of the two streams consists of one convolutional layer and eight residual blocks, where each residual block consists of two convolutional layers (see Figure 3).

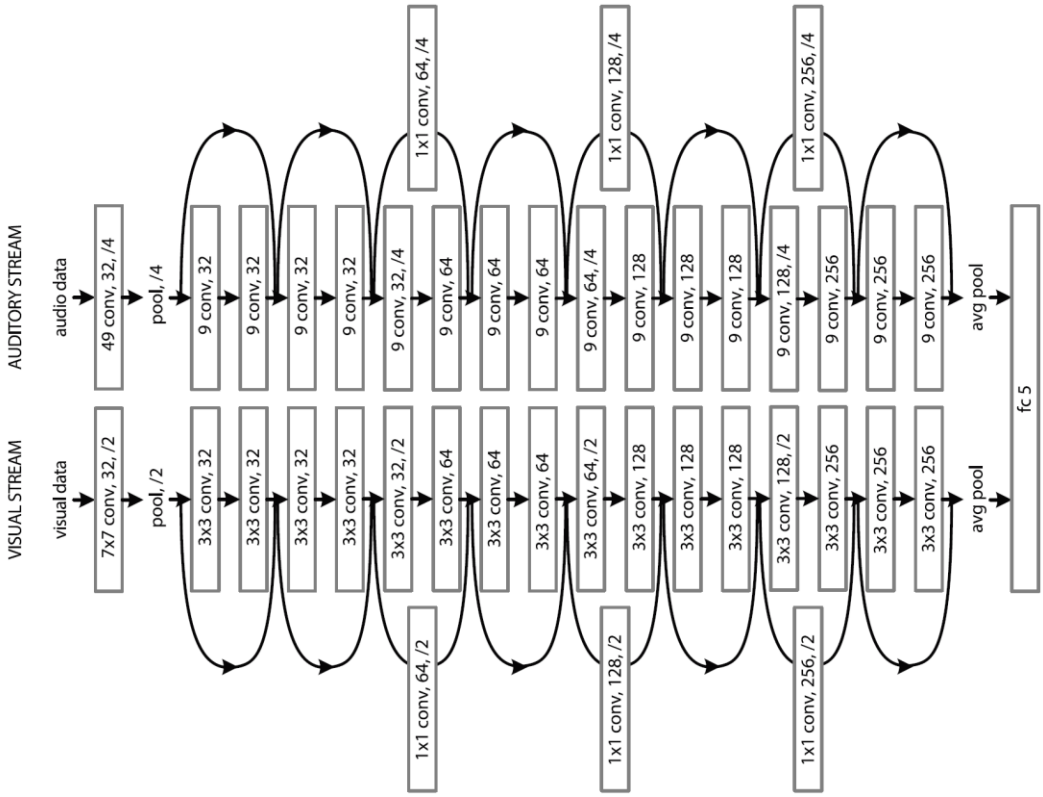


Figure 3. Architecture of the baseline model (source: [3])

It should be noted that two types of residual blocks are used in the baseline model. "Residual block A" consists of convolutional layers with a kernel size of 3 by 3 in the visual stream, and a kernel size of 9 in the audio stream, both having a stride of 1. This means there is no down-sampling of the data. On the other hand, the convolutional layers in "Residual block B" have a kernel size of 1 by 1 with a stride of 2 in the visual layer, and a kernel size of 1 with a stride of 4 in the audio layer. Residual block B down-samples the data.

3.1.3 Network Training

The training of this network was done using a stochastic gradient optimization method called Adam [20] using parameters $\alpha = 0.0002$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and a minibatch size of 32 samples. The training consisted of 900 epochs, where each epoch had 187 training steps consisting of samples selected randomly from the 6000-sample training set. Every 300 epochs, α was reduced with a factor 10. During the testing, the model achieved an average accuracy of 0.9109, ending up in the third place in the Chalearn challenge [18].

3.2 3D Apparent Personality Prediction Model

In Figure 4, a schematic overview of the new 3D-APP model pipeline can be seen.

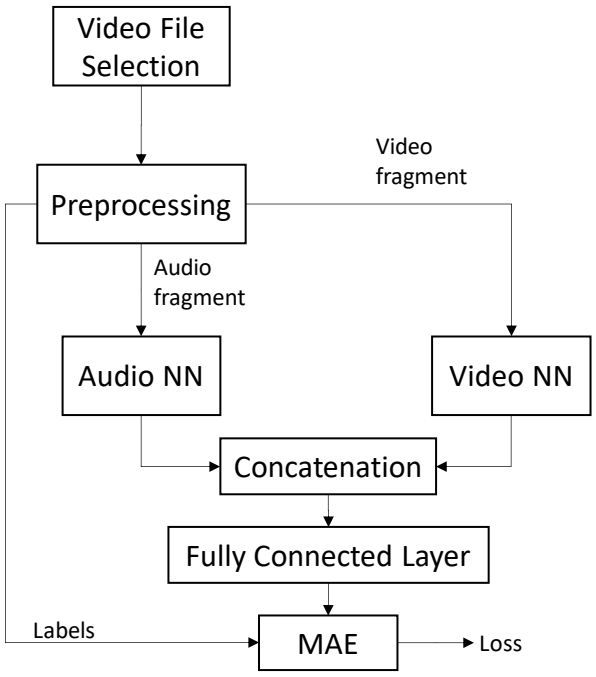


Figure 4. Detailed specification of the online preprocessing and training

When designing the new architecture of the network to allow for synchronized visual and audio streams, two aspects are important:

- How should the video frames and audio samples be selected and pre-processed?
- How are the two information streams synchronized?

In the next two subsections, the design decisions we took with regard to these two problems are described and motivated.

3.2.1 Data Preprocessing

We assume that there exists an underlying association between the auditory and visual information. Thus, in order to create stronger associations and connections between the two separate streams, temporal information needs to be included in the network. This effectively means synchronizing the visual input with the audio input. This can be achieved in several ways. One is to pair every input frame with corresponding audio information, i.e. in the form of raw audio segments in the duration of that single frame. However, with an average frame rate of 25

frames per second, the duration of a single frame is 0.04 seconds. Given that the average duration of a word uttering is much longer than 0.04 seconds, the amount of added information would be very low. Using such a short corresponding audio fragment is therefore unlikely to increase performance of the model. Furthermore, using single frame-audio pairs leaves out the temporal information over multiple frames. A more feasible approach to this problem is to pick a sequence of frames within in a specified time interval, and feed them through the network simultaneously, along with an audio fragment corresponding to that interval. This means that input data of the visual stream is represented as 4D matrices of the shape *time* x *width* x *height* x *color*, rather than single frame 3D matrices. This approach to temporally order video-frames is similar to that proposed in [4], [21], and [22].

When selecting the visual samples, the size of these clips must be taken into account. Using full frames would take a lot of memory and slow down the batch making process. As mentioned earlier in Data Preprocessing, Güçlütürk et al. [3] try to make as little assumptions as possible on importance of visual features by selecting random crops. It is possible that background information is relevant for personality prediction, but this could only be the case if the video creators consciously or unconsciously chose their own video background and location. Although this is true for this dataset, such an assumption would prevent the generalization to other videos, in which this is not the case. Furthermore, according to [7], internal and external facial features are accurate predictors of four of the five personality traits. Hence, facial information is highly relevant to personality prediction. Lastly, we attempted to create better associations, on the one hand between visual information and auditory information, and on the other hand between different visual frames in time. For the former aspect, the visual information must somehow be relevant to the auditory information. For the latter, there must exist some continuity between the visual frames. Facially centered crops would provide both of these aspects: lip movements and sound are correlated, and sequential facially centered frames contain relevant information about movement.

For the reasons mentioned above, sequences of 208 by 208 pixels facially centered frames were chosen, along with audio segments corresponding to the time window of the frames (see Figure 5). The pre-processing of the audio was performed in a similar way as in the baseline model.

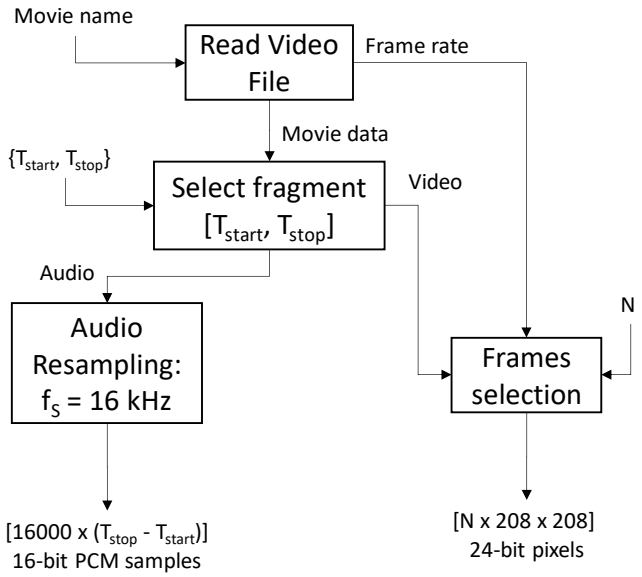


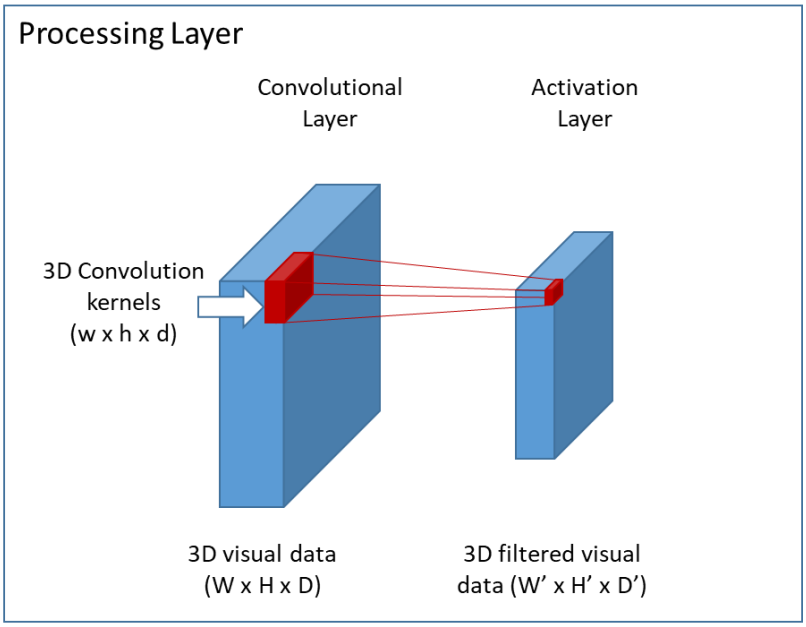
Figure 5. The data preprocessing step

Due to the relatively large time interval necessary for the audio sample, the volume of data corresponding to the video frames in this time interval is too large to be processed efficiently by the network, resulting in a prohibitive training time. On the other hand, the change in visual information between two consecutive frames is very low. Therefore, it is much more efficient to select a subset of the frames, equally spaced over the duration of the selected time interval. The time duration of the intervals, the number of frames, and the audio sampling frequency are fixed.

3.2.2 Model Architecture

The changes that are made in the architecture are limited to the visual stream. Therefore, in this subsection the auditory stream is not further discussed.

To implement this approach of simultaneously feeding multiple frames through the network, we replace the original 2D convolutional layers in the visual stream with 3D convolutional layers, hence also including the time dimension, which is inspired by models such as [4], [21], and [22]. A schematic representation of 3D convolution can be seen in Figure 6.



$$W' = (W - w + 2P_W)/stride_W + 1$$
$$H' = (H - h + 2P_H)/stride_H + 1$$
$$D' = (D - d + 2P_D)/stride_D + 1$$

Figure 6. Processing Layer with 3D convolution. The formulas show the new dimensions of the visual sample after convolution.

A detailed overview of the 3D-APP model architecture can be seen in Figure 7. Similar to the model architecture described in section 3.1.2, the 3D-APP model makes use of the residual blocks A and B. In order keep the 3D-APP model similar to the baseline model, the kernel sizes for the width and height of the frames are kept the same. However, to account for the relatively limited number of frames, the depth of the kernel must also be small. For the convolutional layers in residual block A, we choose a kernel size of 3 by 3 by 1, with a stride of 1. Convolutional layers in residual block B use a kernel size of 3 by 3 by 3 with a stride of 2. Hence, the depth of the sample is reduced.

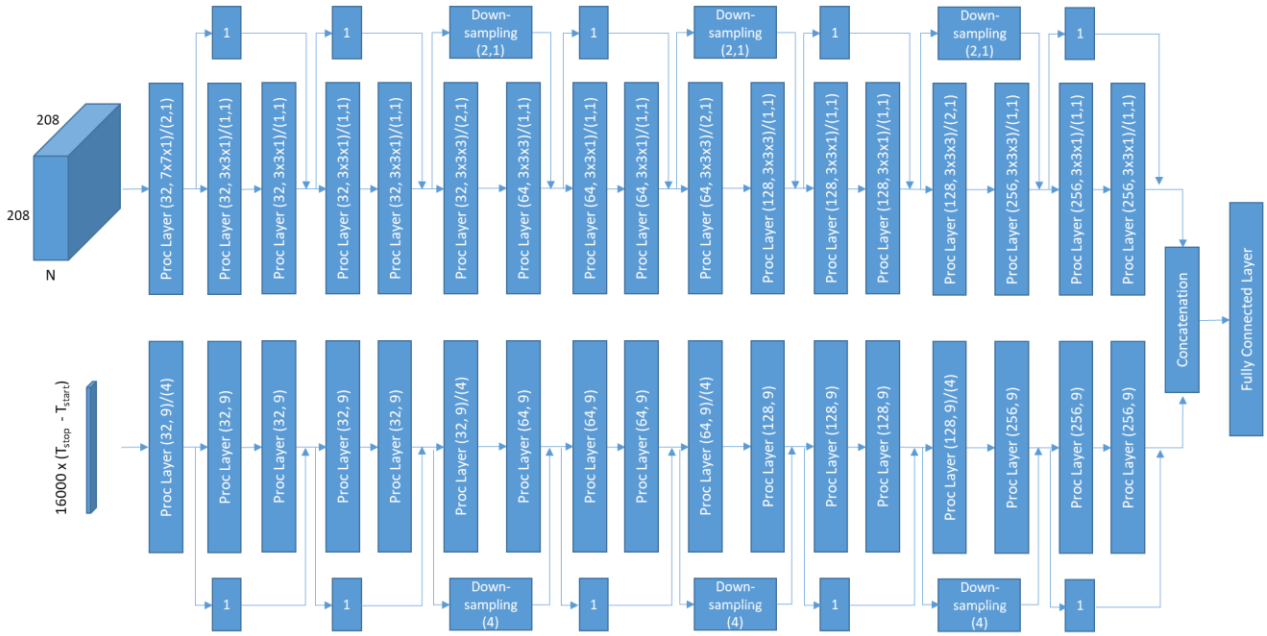


Figure 7. 3D-APP model architecture

Besides changing the shape of the input data and the mechanics of the convolutional layers in the visual stream, no changes are made to the rest of the network.

3.2.3 Network Training

The most important challenge regarding the definition of the network training was to establish which temporal parameters to use for obtaining the best performance. These parameters consist of the number of selected frames and the size of the time window, as well as the resulting frame density. Several models were trained with varying number of frames and time window sizes:

- model 2.5_5, with 5 frames over 2.5 seconds, resulting in 2 frames per second,
- model 5_5, with 5 frames over 5 seconds, resulting in 1 frame per second,
- model 5_10, with 10 frames over 5 seconds, resulting in 2 frames per second, and
- Model 8_8, with 8 frames over 8 seconds, resulting in 1 frame per second.

The reason for choosing such variations in the number of frames and fragment duration was to make possible a comparison between various amounts of temporal information in the video

stream, and thus to provide the means for evaluating the importance of this additional data dimension.

In order to create models that are comparable to the baseline model, all other training parameters were kept constant, with one exception. In the baseline model, each epoch consisted of 187 training steps using batches of 32 samples, thus covering the entire training set. In the training procedure of the 3D-APP model models, all training data was traversed as well, but occasionally a video was too short for the selected time frame. In this case, the video was skipped, resulting in a slightly shorter epoch. We worked under the assumption that this will not affect the results, since the skipped videos are a tiny fraction of the training set.

Furthermore, to account for the use of facial crops instead of full frames in the adapted model, two new baseline models were trained on the same facial crops. However, the same temporal parameters were used as in the model by [1]: one frame selected at random from the whole video, and an audio fragment of 3.136 seconds. Thus, the resulting baseline models provide a better frame of reference for the 3D-APP model.

Lastly, as mentioned in section 3.2.1, each of these models are *synchronized*. In order to make it possible to test whether the synchronization of audio and visual information indeed leads to an improvement of the model's performance (as originally hypothesized), an additional *desynchronized* model was trained, using 5 frames over 5 seconds, which can be compared with the synchronized 5_5 model. It should be noted that, besides selecting the audio samples and video frames from independently selected time windows, no other changes were made compared to model 5_5.

3.3 Technical Specifications

In this subsection, the technical details needed for the implementation of the models are summarized.

All neural network implementations were made using Chainer, CUDA, and cuDNN. For facially cropping the videos, the Python library dlib[1] was used. For frame and audio selection, we first

[1] https://github.com/davisking/dlib

attempted to use FFmpeg, which is a multimedia framework able to decode and encode media [23]. Equally spaced frames and the corresponding audio are directly selected from mp4 files, and then saved as arrays. Unfortunately, reading the data from mp4 files is too slow for a reasonable training time. It results in an average batch making time of 8 seconds. To solve this problem, we choose to save the data as HDF5 files. HDF5 is a data model, library and file format that supports unlimited datatypes [24]. This change lowered the batch making time to 1.5 seconds on average.

4 RESULTS

In this section, the different models that were trained will be discussed, compared and evaluated in terms of convergence, test loss, and performance.

4.1 TRAINING

As mentioned in the previous section, each model was trained for 900 epochs. During the training phase, the average loss over all samples of the model was recorded after each epoch. This loss was computed by taking the mean absolute error between the predicted samples and the labels. In Figure 8, the training loss of each model is plotted as a function of the epoch number. The training loss is defined as the mean absolute error of the predictions during the training phase.

Figure 8. Training loss graphs

4.2 TESTING

A quick overview of the test accuracies of the different types of model architecture can be seen Table 1. The accuracy is defined as 1 – the mean absolute error of the test predictions. Each model was trained for 900 epochs and saved every 10 epochs, resulting in 90 available models per architecture, from which the best was selected.

Table 1. Test loss of the best performing models. The abbreviations Ex, Ne, Ag, Co, and Op stand for respectively, extraversion, neuroticism, agreeableness, conscientiousness, and openness.

Model	**Average**	**Ex**	**Ne**	**Ag**	**Co**	**Op**
Baseline 1 (epoch 819)	0.900825	0.899148	0.89727	0.906019	0.900043	0.901645
Baseline 2 (epoch 529)	0.902124	0.899695	0.901094	0.909605	0.897208	0.903018
2.5 seconds, 5 frames (epoch 639)	0.905445	0.906947	0.900495	0.910346	0.903762	0.905677
5 seconds, 5 frames (epoch 319)	0.906305	0.907046	0.90363	0.910904	0.902743	0.9072
5 seconds, 10 frames (epoch 399)	0.907497	0.908668	0.905557	0.911338	0.903853	0.908067
8 seconds, 8 frames (epoch 499)	0.907736	0.911171	0.90351	0.911663	0.9048	0.907536
5 seconds, 5 frames, desynchronized (epoch 479)	0.906894	0.909574	0.90398	0.910854	0.903853	0.906208

The models were tested using a separate dataset, by computing the average loss over all test samples. Besides the mean absolute error of all output traits, separate losses for each individual trait were recorded as well. Hence, we can determine to what extent the 3D-APP model is able to classify each trait separately.

To observe the progressions in test loss throughout the training process, the mean absolute error was computed for each epoch. To obtain a fair comparison between the epochs, the same time fragments were used at each test. In the desynchronized model, the audio fragments were shifted two seconds forward compared to the video fragments.

In Figure 9, the test loss of each model architecture is plotted as a function of the epoch number. In Figure 10, a comparison is shown between model 5_5 and the desynchronized model 5_5.

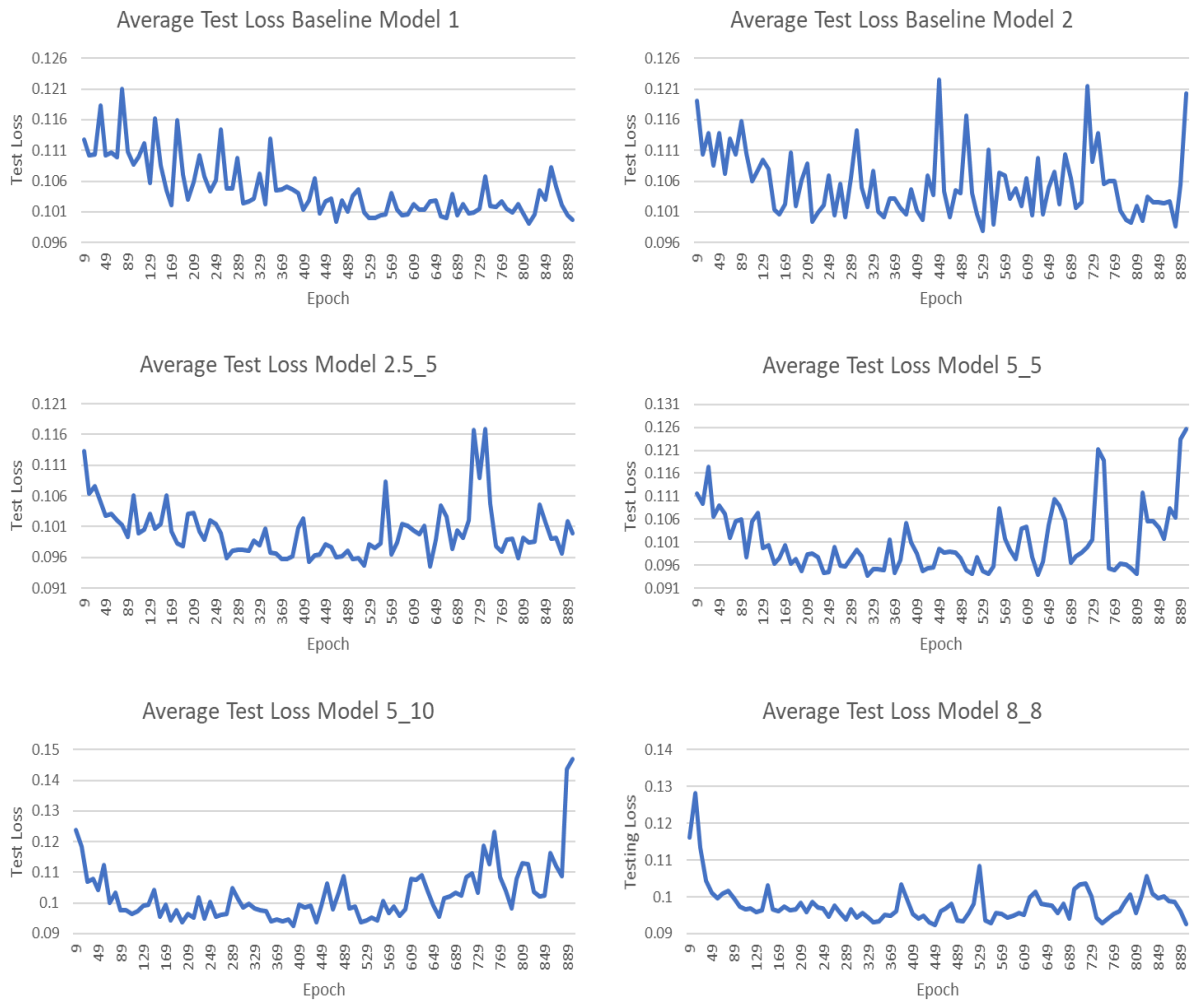


Figure 9. Mean Absolute Error of the trained models.

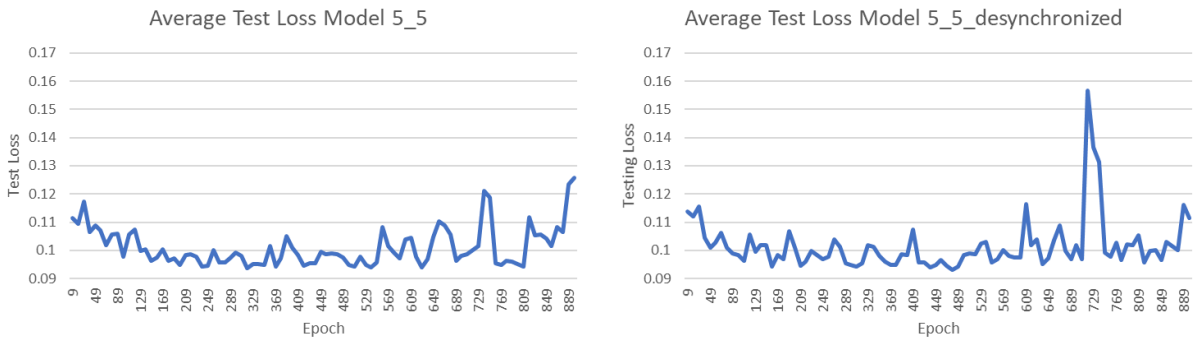


Figure 10. Comparison between a synchronized and desynchronized model.

4.3 EVALUATION

First, a comparison between the speed of convergence in the training phase can be made by looking at the graphs in Figure 8. We can see that the baseline models decrease the slowest in the training loss, followed by model 2.5_5, models 5_5 and 5_5 desynchronized, model 5_10 and lastly, model 8_8. This suggests that using more frames and using a longer time fragment improves the model's ability to fit the training data. However, this does not tell us anything about the model's ability to generalize to the test set.

Secondly, when looking at the graphs in Figure 9 and average accuracies in Table 1, we observe that model 8_8 achieves the highest performance, followed by model 5_10, model 5_5, model 2.5_5 and the two baseline models. The baseline models maintain a decreasing trend in test loss until the last epochs, whereas models trained with multiple frames have the tendency to reach their lowest test loss in an earlier stage of the training, sometimes followed by a rising trend, which could indicate overfitting.

In Figure 11, a comparison is shown between two models trained with fragments with the same duration, but different frames densities. It appears that the model with more frames (the blue graph) converges faster, reaches a slightly better performance, but also presents a rising trend after 400 epochs. A possible explanation is that the difference between successive frames is smaller than for the other fragments, and thus more frames are presented with similar visual content.

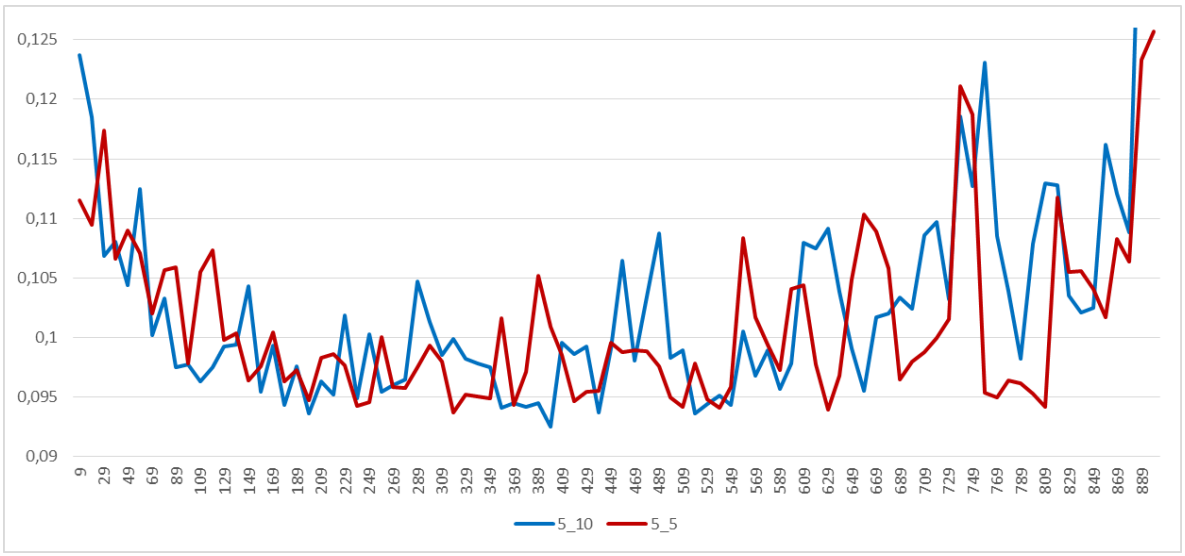


Figure 11. Test Loss of model 5_10 and model 5_5 as a function of epochs.

The presence of cross modal temporal patterns should result from the better performance of a model trained with synchronized audio and video, with respect to a model trained with desynchronized modalities.

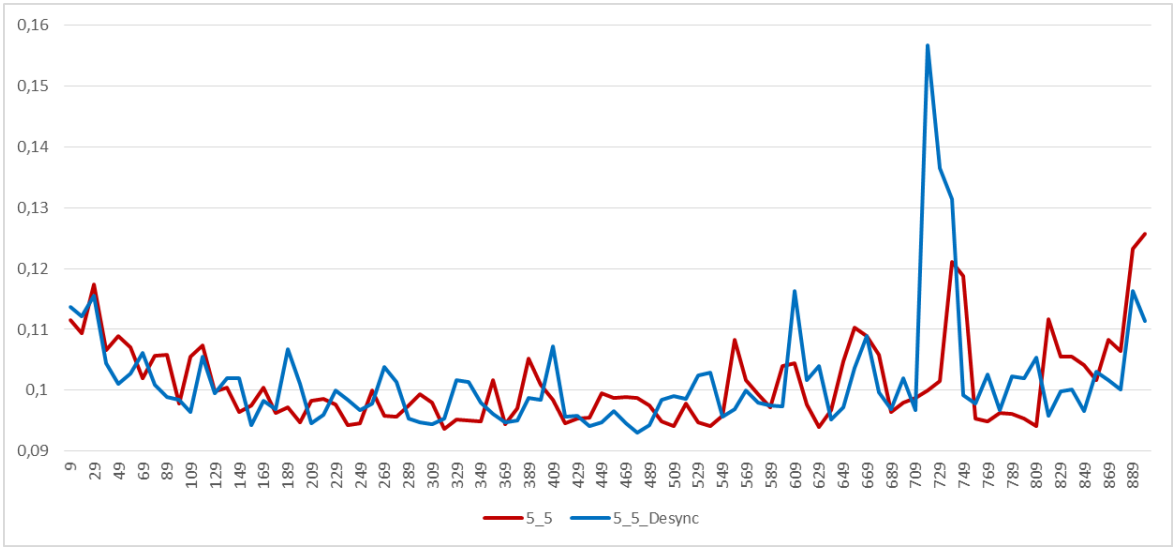


Figure 12. Test loss for the models 5_5 with synchronous and desynchronized audio, respectively.

However, from Figure 12 and from Table 1, it results that the difference between these two models is insignificant. This may be due to the fact that the fusion of the two modalities is only performed in the last layer, so cross-modal temporal patterns could not be learned. A network

architecture with early fusion needs to be implemented in order to achieve more relevant results related to the learning of cross-modal patterns.

When looking at the performance in for each individual trait, it appears that some of the five traits are more easily predictable than others. For comparison the best performing model is chosen (8 frames over 8 seconds). The best overall performance is achieved, in all models, by Agreeableness and Extroversion, while Neuroticism and Conscientiousness are the most difficult to estimate. Figure 13 shows the losses for the average between Extroversion and Agreeableness in comparison with the average between Neuroticism and Extroversion.

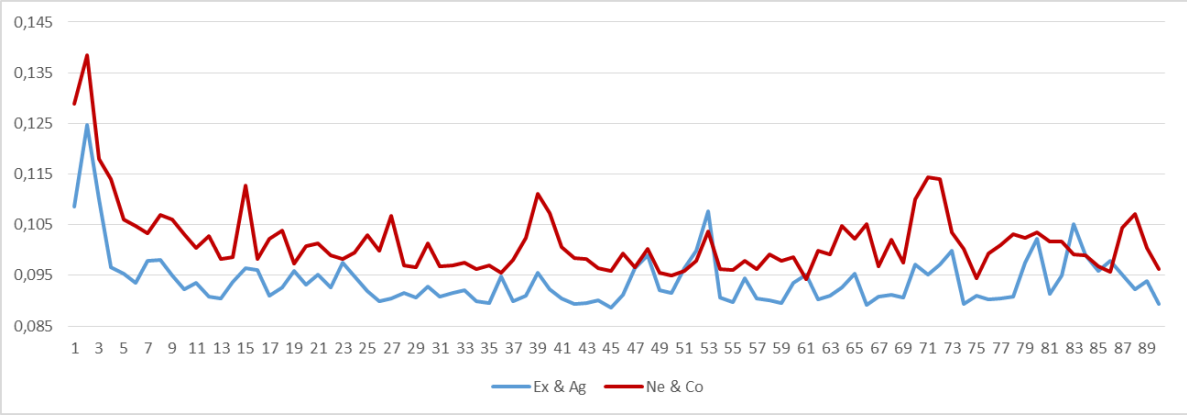


Figure 13. Comparison between the losses for different traits by the model 8_8, plotted against the epoch number.

5 CONCLUSION

In this section, we reflect back to the research question and hypothesis proposed in section 0. We conclude this thesis with an overview of the limitations of the research and some pointers to future work.

5.1 CONTRIBUTION

The hypothesis put forward in section 0 can be split up into three parts:

1. The 3D-APP design performs better than the baseline model. This is because:
2. The 3D-APP models exploit the additional temporal information, and
3. Bimodal associations are learned from a sequence of video frames synchronized with audio.

We can conclude that each of the 3D-APP models indeed achieve a higher accuracy than the baseline architecture trained on the same dataset. Thus, the first part of the hypothesis is confirmed. Moreover, to obtain a more complete comparison between the models, the testing loss throughout the training process and the speed of convergence are also taken into account. The 3D-APP models converged in fewer epochs compared to the baseline model.

Regarding the second part of the hypothesis, the contribution of the temporal information is confirmed by the fact that the model 2.5_5 performs slightly worse than model 5_5. This suggests that a longer time fragment results in better performance when the number of frames is kept constant. A possible explanation is that the variation between the different frames from a video is higher, resulting in more information per data sample. However, these models need to be trained multiple times to provide conclusive evidence.

Furthermore, the benefit of exploiting cross-modal patterns should have been proven by a worse performance of the desynchronized model 5_5 compared to the synchronized model. This did not result from our experiments. A possible explanation is that the visual and auditory channels are only merged in the last layer, so the association between these is minimal. An early fusion approach should be used to get a stronger conclusion about the value of cross-modal patterns. This is explained in more detail in Section 5.3.

Lastly, we can observe clear differences in the prediction performance of individual traits. Figure 13 shows that the models consistently have a higher prediction error for traits neuroticism and conscientiousness. A higher error for conscientiousness is in accordance with the research from [7], as they were unable to predict this trait from internal and external facial features.

5.2 LIMITATIONS

In this subsection, we discuss several aspects of this research that can be improved in order to strengthen the evidence for the improved performance of the 3D-APP model.

Due to time constraints, each 3D-APP model architecture was trained once, and the baseline architecture was trained twice. Although the results of the testing have shown a clear

performance improvement of the 3D-APP models, in to order to be able to conclude with a high degree of certainty that the 3D-app model consistently performs better than the baseline model, the models should be trained multiple times.

Secondly, more research is needed on the internal associations between audio and video. We cannot yet conclude that the increase in performance and the faster convergence can be attributed to the earlier mentioned bimodal temporal patterns, as the synchronized model does not show improvement compared to a desynchronized model in the current state of the network architecture.

Furthermore, the parameters of the 3D convolution were not varied, and may be suboptimal. Further experimentation is necessary to establish the settings of the optimal network architecture. Such parameters include the number of kernels, stride and kernel size.

Lastly, further experimentation with the parameters of the optimization algorithm could improve the training result.

5.3 FUTURE WORK

The applications of audiovisual neural networks go further than personality detection. In the case of models such as the 3D-APP model, which synchronizes audio and video data, interesting future applications might include the detection of discrepancies between facial cues and vocal cues. A particular field that might benefit from this, is lie detection. There is evidence that cross modal discrepancies of personality perception are associated with dishonesty [25]. The architecture of this network is ideal for such an application, as it uses both modalities and can be trained on short videos containing a statement, labelled either true or false.

Secondly, it would be interesting to investigate whether some personality traits are more accurately predicted from just one of the audio, or the video streams. Such research could lead to more fine-tuned trait-specific models that exhibit a better prediction accuracy.

Lastly, in its current state, the 3D-APP model performs a concatenation followed by a fully connected layer at the end of the network. Thus, data fusion occurs late in the processing, and as a consequence, the associations formed between synchronized audio and video is limited, as

described in section 5.1. Shifting the data fusion to an earlier point in the processing of the two streams could prove beneficial in the context of synchronized modalities. This might also provide the opportunity for studying the activation of neurons after the point of fusion, which could lead to a better understanding of the underlying patterns in the data. For example, multimodal neurons may arise, that do not respond to either video or audio patterns, but rather to a combination of the two.

6 REFERENCES

[1] R. A. Andersen, "Multimodal integration for the representation of space in the posterior parietal cortex," *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* vol. 352, no. 1360, pp. 1421-1428, 1997.

[2] H. McGurk and J. Macdonald, "Hearling lips and seeing voices," *Nature,* vol. 264, pp. 746-748, 23 December 1976.

[3] Y. Güçlütürk, U. Güçlü, X. Baró, H. J. Escalante, I. Guyon, S. Escalera, M. A. J. Van Gerven and R. Van Lier, "Multimodal First Impression Analysis with Deep Residual Networks," *IEEE Transactions on Affective Computing.,* 2017.

[4] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian and A. Mittal, "Bi-modal First Impressions Recognition using Temporally Ordered Deep Audio and Stochastic Visual Features," in *European Conference on Computer Vision*, Amsterdam, 2016.

[5] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: A meta analysis," *Personnel Psychology,* vol. 44, no. 1, pp. 1-26, 1991.

[6] L. R. Goldberg, "The structure of phenotypic personality traits," *American Psychologist,* vol. 48, no. 1, pp. 26-34, 1993.

[7] R. S. S. Kramer and R. Ward, "Internal facial features are signals of personality and health," *The Quarterly Joirnal of Experimental Psychology,* vol. 63, no. 11, pp. 2273-2287, 2010.

[8] T. Zhang, R.-Z. Qin, Q.-L. Dong, W. Gao, H.-R. Xu and Z.-Y. Hu, "Physiognomy: Personality Traits Prediction by Learning," *International Journal of Automation and Computing,* vol. 14, no. 4, pp. 386-395, 2017.

[9] D. Budgen, and P. Brereton, "Performing systematic literature reviews in software engineering," in *Proceeding of the ICSE '06 Proceedings of the 28th international conference on Software engineering*, 2006.

[10] B. Kitchenham, P. Brereton, D. Budgen and M. Turner, "Systematic literature reviews in software engineering – A systematic literature review," *Information and Software Technology,* vol. 51, no. 1, pp. 7-15, 2009.

[11] A. Hevner, "A Three Cycle View of Design Science Research," *Scandinavian Journal of Information Systems,* vol. 19, no. 2, pp. 1-6, 2007.

[12] K. Peffers, T. Tuunanen , M. Rothenberger and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems,* vol. 24, no. 3, pp. 45-77, 2007.

[13] T. Baltrusaitis, C. Ahuja and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2018.

[14] R. Kiros, R. Salakhutdinov and R. Zemel, "Multimodal Neural Language Models," in *International Conference on Machine Learning*, Beijing, 2014.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang and A. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," in *ICLR*, 2015.

[16] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller and G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing,* vol. 31, no. 2, pp. 153-163, 2013.

[17] F. Gürpinar, H. Kaya and A. A. Salah, "Multimodal Fusion of Audio, Scene, and Face Features for First Impression Estimation," in *International Conference on Pattern Recognition*, Cancún, 2016.

[18] V. P. Lopez, B. Chen, A. Clapes, M. Oliu, C. Corneanu, X. Baro, H. J. Escalante, I. Guyon and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions - dataset and results," *ChaLearn Looking at People Workshop on Apparent Personality Analysis. ECCV Workshop proceedings,* 2016.

[19] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR,* vol. abs/1512.03385, 2015.

[20] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, San Diego, 2015.

[21] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 1, pp. 221-231, Jan 2012.

[22] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, Fellow, IEEE and J. Dawson, "3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition," *IEEE,* no. 99, 2017.

[23] "FFmpeg," [Online]. Available: https://www.ffmpeg.org/about.html. [Accessed 29 05 2018].

[24] "HDF5 Support," The HDF Group, 29 11 2017. [Online]. Available: https://support.hdfgroup.org/HDF5/. [Accessed 29 05 2018].

[25] C. U. Heinrich and P. Borkenau, "Deception and Deception Detection: The Role of Cross-Modal Inconsistency," *Jorunal of Personality,* vol. 66, no. 5, pp. 687-712, 2002.