Deriving Temporal Protoypes from Saliency Map Clusters for the Analysis of Deep-Learning-based Facial Action Unit Classification

Bettina Finzel[1], René Kollmann[1], Ines Rieger[2], Jaspar Pahl[2] and Ute Schmid[1]

[1]*Cognitive Systems, University of Bamberg, An der Weberei 5, 96047 Bamberg, Germany,*
https://www.uni-bamberg.de/en/cogsys

[2]*Fraunhofer Institute for Integrated Circuits IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany,*
https://www.iis.fraunhofer.de/en.html

Abstract

Reliably determining the emotional state of a person is a difficult task for both humans as well as machines. Automatic detection and evaluation of facial expressions is particularly important if people are unable to express their emotional state themselves, for example due to cognitive impairments. Identifying the presence of Action Units in a human's face is a psychologically validated approach of quantifying which emotion is expressed. To automate the detection process of Action Units Neural Networks have been trained. However, the black-box nature of Deep Neural Networks provides no insight on the relevant features identified during the decision process. Approaches of Explainable Artificial Intelligence have to be applied to provide an explanation why the network came to a certain conclusion. In this work "Layer-Wise Relevance Propagation" (LRP) in combination with the meta analysis approach "Spectral Relevance Analysis" (SpRAy) is used to derive temporal prototypes from predictions in video sequences. Temporal prototypes provide an aggregated view on the prediction of the network by grouping together similar frames by considering relevance. Additionally, a specific visualization method for temporal prototypes is presented that highlights the most relevant areas for a prediction of an Action Unit. A quantitative evaluation of our approach shows that temporal prototypes aggregate temporal information well. The proposed method can be used to generate concise visual explanations for a sequence of interpretable saliency maps. Based on the above, this work shall provide the foundation for a new temporal analysis method as well as an explanation approach that is supposed to help researchers and experts to gain a deeper understanding of how the underlying network decides which Action Units are active in a particular emotional state.

Keywords

Temporal Prototypes, Layer-wise Relevance Propagation, Facial Action Unit Detection, Spectral Clustering, Affective Computing, Correlation Analysis

LWDA'21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany

bettina.finzel@uni-bamberg.de (B. Finzel); rene-josef.kollmann@stud.uni-bamberg.de (R. Kollmann); ines.rieger@iis.fraunhofer.de (I. Rieger); jaspar.pahl@iis.fraunhofer.de (J. Pahl); ute.schmid@uni-bamberg.de (U. Schmid)

0000-0002-9415-6254 (B. Finzel); 0000-0002-8694-762X (I. Rieger); 0000-0002-1301-0326 (U. Schmid)

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

Correct distinction and interpretation of facial expressions is an important part of medical treatment. For example, when patients experience pain, but cannot express how they feel, which can be the case after surgery or due to neuro-cognitive impairments like dementia, healthcare professionals have to rely on their skills to interpret facial expressions in order to choose the right level of medication against pain. Systems that automatically classify facial expressions as indicators for pain and other emotions can support the medical decision making process.

This decision process needs to be as transparent and efficient as possible. Therefore, classification outcomes have to be presented to the human decision maker in a comprehensible way with as much as information aggregation as possible. There exist approaches that classify facial expressions, in particular so-called Action Units (AUs) [1] from video material with the help of convolutional neural networks [2]. Neural networks usually remain intransparent black-boxes without applying further methods to explain their decisions. Methods to make black-box models transparent are summarized under the term "Explainable Artificial Intelligence" (XAI) [3]. XAI presents the challenge to create explainable models embedded in an explanatory framework while still achieving state-of-the-art performance. For an overview of methods and terminologies refer to Schwalbe et al. [4]. In order to aggregate information, methods have been developed for example for visual explanations by clustering saliency maps that denote the relevance of individual pixels or pixel groups with respect to a class decision [5]. Beyond clustering, prototypes provide an explanation that is reduced to the most representative characteristics of a class.

In this paper, we present an approach that is based on explainable Action Unit prediction with a convolutional neural network and that is designed to explain co-occurrence of Action Units in video sequences of human facial expressions in different emotional states. Our aim is to provide an effective explanation method based on prototypes that contain the features most relevant to a class decision. Our approach shall further provide an efficient explanation method by reducing the number of saliency maps to be checked by a human decision maker per video sequence. To the best of our knowledge, our approach is the first to consider the temporal co-occurrence of facial expressions by means of explaining Action Unit predictions over time. In general, our approach could be used to aggregate visual explanations for classification tasks on image sequences and to detect temporal prototypes that are representative for certain classes, such as different emotional states. The development of our proposed approach was mainly motivated by two research questions:

- Are correlated predictions similar enough to be aggregated into prototypes?
- Are there reoccurring patterns in temporal prototypes that are specific to classes?

This paper is organized as follows: First, we summarize related work in section 2. Afterwards, we present how to derive temporal prototypes from saliency map clusters in section 3. This section introduces the application use case of facial expression recognition, describes how to create saliency maps and how to cluster them based on spectral relevance analysis. We further explain how temporal prototypes are derived from Action Unit clusters and visualized for the human decision maker. In section 4 we present the results of our evaluation. We conclude and present future research directions in section 5.

2. Related Work

Affective Computing and especially Action Unit detection has evolved tremendously in the last years, fueled by new, extensive datasets appearing as well as new algorithmic approaches. For detailed information refer to the following overview papers: Martinez et al. [6] give a thorough explanation about the components of an AU-detector, discusses relevant databases, and identifies challenges for the transfer to real-world scenarios. Zhi et al. [7], another survey, puts more focus on modern Deep Learning approaches and has a very detailed overview of the AU detection challenges over the last few years. Action Units are often a feature in downstream tasks such as categorical emotions analysis [8] or pain detection [9]. The increased research efforts have also spawned multiple commercial applications, some of which have been compared by Dupré et al. [10].

Recent challenges such as the ABAW Challenge [11] at the Face and Gesture Conference 2020 have provided a benchmark opportunity for AU algorithms and are continuously challenging those algorithms with even more realistic application scenarios. Furthermore, they give a good overview on the current state-of-the-art [12, 13, 2].

Visualization of the Neural Network's decision is one of the most prominent XAI methods for making emotion and Action Unit detectors more transparent. Heimerl et al. [14] apply XAI methods specifically for emotions in their introduced framework NOVA. This annotation tool incorporates recent XAI techniques such as a confidence value or visual explanations. In NOVA, the user can interactively change the predicted labels by also relying on the additional information provided, where after the model is retrained with the users feedback. Heimerl et al. [14] show that XAI techniques help non-expert users to understand the inner workings of a machine learning system better. This also motivates our work.

Chu et al. [15] propose a hybrid model incorporating spatial, temporal, and correlation information of Action Units. They were the first to visualize the learned concepts for Action Unit models. Zhao et al. [16] propose a region learning approach for Action Unit detection and use saliency maps to compare different approaches qualitatively. Sánchez et al. [17] and Ntinou et al. [18] use heatmap regression that uses predefined heatmaps of the expected areas of the Action Units as ground truth for training an Action Unit model. Examples of wrongly and correctly predicted heatmaps along with the ground truth heatmaps are shown in the paper for the reader's qualitative evaluation. As opposed to our approach, the visualizations used in all described approaches are computed on a per frame basis for qualitative evaluation or as training feedback and information and are not aggregated over time in order to constitute temporal prototypes.

Our approach uses LRP as visualization method for Action Unit detection. Rieger et al. [19] were the first to apply LRP for Action Unit detection. They propose a verification pipeline based on bounding boxes to quantitatively evaluate if the Neural Network's prediction is based on relevant facial regions. To the best of our knowledge this is still the only approach using LRP for Action Unit Network evaluation.

At this point in time, we are the first to utilize temporal information in our visualizations, let alone in combination with LRP and SpRAy.

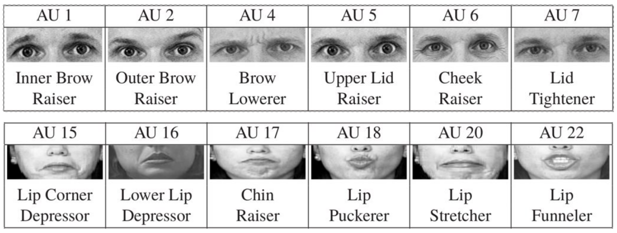

Figure 1: A subset of upper (top row) and lower (bottom row) facial Action Units with corresponding example images (from Tian et al. [21], according to Ekman [1])

3. Deriving Temporal Prototypes from Saliency Map Clusters

This section introduces our approach to create temporal prototypes to visually explain Deep-Learning based classification of facial expressions. The proposed solution utilizes a combination of different methods. First, we create saliency maps as a basis to visual explanations for each individual video frame. Afterwards, we apply spectral clustering in order to group similar saliency maps. Next, we create temporal prototypes by combining the prototypes that represent individual clusters according to their predicted temporal occurrence. Finally, we visualize the computed temporal prototypes based on an existing heatmapping approach that was adapted by us in order to increase interpretability [20].

3.1. Application to Facial Expression Recognition

As introduced in the previous section, we are applying our proposed approach to classify and explain human facial expressions based on Action Units specified by the Facial Action Coding System (FACS) [1]. The FACS includes specifications for 46 Action Units. A subset of those with its corresponding descriptions is shown in Figure 1. Action Units often co-occur in specific emotions displayed in the face. Due to space restrictions, the reader may refer to Farnsworth [22] or the work of Du et al. [23] that has been conducted based on the FACS.

The Action Unit model for the LRP analysis and coding of Action Units was provided by Pahl et al. [24]. Their convolutional neural network approach allows to combine multiple datasets in a single training set even when the combination would lead to missing values. This is achieved by training a multi-label neural network which is enabled to handle missing labels by a batch-wise adaptation of the vector fed to the loss function. We chose this Action Unit approach since it provides competitive performance in comparison to state-of-the art algorithms.

We then subsequently evaluate the existence and meaningfulness of temporal prototypes on the Extended Cohn-Kanade Dataset (CK+) (Lucey et al. [25]) as well as the Actor Study Dataset (AS) (Seuss et al. [26]). The CK+ dataset provides FACS coded Action Unit annotations as well as validated emotion labels for some of the included video sequences. This makes CK+ a well-suited initial dataset to evaluate temporal prototypes and their relation to the shown emotion. The dataset includes 593 sequences from 123 subjects that show posed facial expressions of seven different emotions: angry, disgust, fear, happy, sadness, surprise and contempt. To verify the results from the CK+ dataset, our approach is also applied to a subset of the AS dataset, namely high quality video sequences of 3 actors that pose emotions from different camera angles.

3.2. Creating Saliency Maps with Layer-wise Relevance Propagation

Likewise to [19] we utilize Layer-wise relevance propagation (LRP) to create visual explanations, namely saliency maps, as introduced by Bach et al. [27]. LRP is a back-propagation and decomposition method that assigns a relevance value to each input parameter. For neural networks that classify images LRP assigns relevance to each pixel. In other words, for an image *x* LRP computes the contribution of each pixel to the classification result *f(x)*.

Since a neural network typically consists of multiple layers, the connection of prediction and pixel values as shown in equation 1 cannot be derived directly. A propagation of relevance values is performed from the output layer back to the input layer, in order to find out a single pixel's contribution to the prediction. Equation 1 describes the principle of relevance propagation (with $R_i^{(l)}$ being the relevance score of neuron i at layer l [28]):

$$\sum_d R_d^{(input)} = ... = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = ... = f(x) \tag{1}$$

Importantly, the contribution of a pixel d to the classification is not necessarily positive such that R_d can be positive or negative. A positive relevance value indicates that the pixel contributed evidence for the existence of the prediction, whereas a negative relevance value indicates proof against the prediction. The relevance values of each pixel are saved in a so-called *relevance map*, which represents the relevance values of a whole image as a matrix.

Our implementation complies with the best practices presented in Montavon et al. [29] and Kohlbrenner et al. [30] with respect to computation of relevance. We use the decomposition rules as provided by the iNNvestigate toolbox (Alber et al. [31]), the *LRPSequentialPresetAFlat* composition rule in particular.

3.3. Clustering Relevance Maps with Spectral Relevance Analysis

Relevance maps can be further clustered based on *similarity* to derive representative prototypes. For that purpose we use the Spectral Relevance Analysis framework (SpRAy, Lapuschkin et al. [5]) that clusters similar relevance maps provided by LRP. Similar relevance maps most likely focus on the same features, such that if those have been identified, a human observer is able to interpret them. This way, SpRAy provides a notion of different strategies a network uses to identify a class, but requires a human observer to interpret them.

SpRAy can be summarized in five basic steps. First, relevance maps are generated with LRP. Afterwards, relevance maps are pre-processed, if needed (e.g., reshaping them into comparable dimensions). Next, the similarity between every relevance map is computed and recorded in a similarity matrix. Based on this matrix, spectral clustering is applied to identify groups. Here, the number of clusters is determined by Eigen-values near or equal to zero in accordance to the method presented in [5]. Finally, the identified clusters can be visualized.

3.4. Creation of Temporal Prototypes based on Action Unit Clusters

After clustering the relevance maps for each sequence of frames to calculate prototypes (the average heatmap from a cluster), we consider the co-occurrences of Action Unit prototypes over

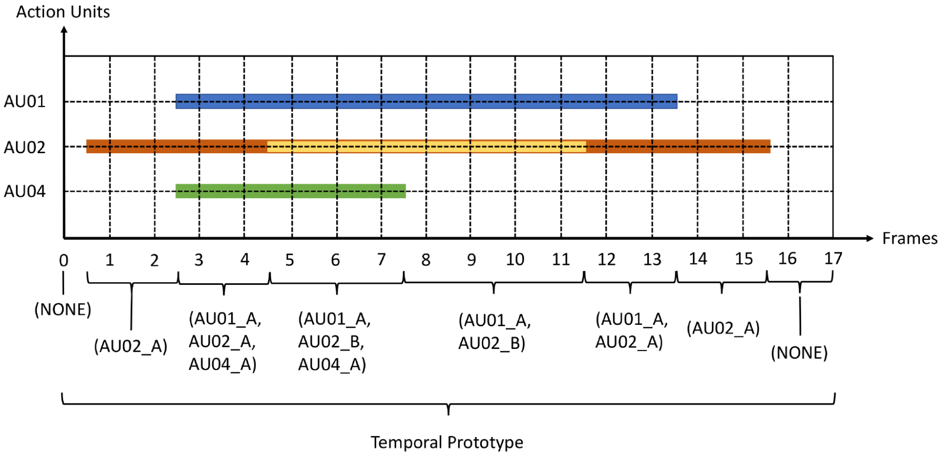

Figure 2: Illustration of how temporal prototypes are derived from a sequences of frames

time based on the prediction of the neural network. The assumption is that as a consequence of the similarity of the input also relevant features to the network remain similar for a certain amount of time. Under this assumption it is possible to define different *stages* within a video sequence by grouping together similar frames until there is a significant change. Those changes might be a transition from one Action Unit of being inactive into being activated or a sudden increase in the intensity of the Action Unit. Note that predictions are only attributed to the same stage if they are indeed similar. Whereas similarity is not just decided by the prediction decision itself (i.e. Action Unit is active or not), but by the similarity of the explanation or reason for the decision, which we compute based on the relevance data. The aggregation of a set of similar predictions into one entity is referred to as building a *prototype* in our approach, such that a prototype serves as a representative for certain duration of the video sequence. The temporal ordering of those prototypes (i.e. the succession of all prototypes in a video sequence) is then called a *temporal prototype*.

As mentioned in the introduction, we aim to answer two research questions: First, are correlated predictions actually similar enough such that they can be aggregated? And second, when comparing temporal prototypes of different video sequences, are there reoccurring patterns?

If temporal prototypes do show patterns, a meaning can be assigned to them. This allows them to be an instrument for knowledge discovery as they provide a condensed description of what the network saw.

Given the artificial example of Figure 2, a sequence of 18 frames (including frame 0) is shown. The network predicts the presence of AU01 (blue), AU02 (orange and yellow) and AU04 (green) at different frames of the sequence. Recognizing the same predictions over multiple frames allows to aggregate them into a prototype by clustering. The first occurrence of a cluster for an Action Unit is always labeled by "A", the second by "B" and so on which allows to identify if a cluster reoccurs during a sequence.

The example of Figure 2 shows two different colors for AU02 because even though the prediction stayed the same the relevance maps for AU02 of the frames one to four and twelve to

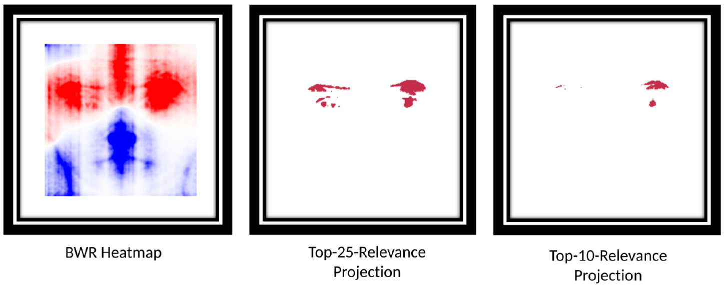

Figure 3: Heatmap produced based on a BWR color scheme (left) compared to reduced heatmap only showing most relevant features that aggregate 25 % (middle) and 10 % (right) of the relevance mass present in the corresponding relevance map (shown prediction is for AU02, the "outer brow raiser")

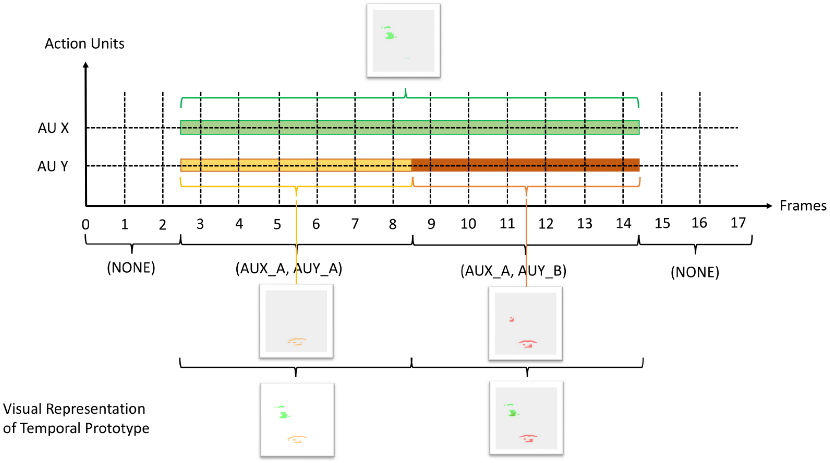

Figure 4: Illustration of how a visual representation of a temporal prototype is derived from a given clustering of Action Unit predictions

fifteen are different to the relevance maps of frame five to eleven. In contrast, relevance maps for AU01 and AU04 are similar throughout the sequence, such that there is only one cluster for each. The first occurrence of a cluster for an Action Unit is always labeled by "A", the second by "B" and so on which allows to identify if a cluster reoccurs during a sequence. No prediction of any Action Unit in a frame is marked by "NONE". The toy example of Figure 2 yields the following temporal prototype:

(NONE) → (AU02_A) → (AU01_A, AU02_A, AU04_A) → (AU01_A, AU02_B, AU04_A) → (AU01_A, AU02_B) → (AU01_A, AU02_A) → (AU02_A) → (NONE)

Thus, the temporal prototype in Figure 2 provides an aggregation of 18 individual frames into eight stages (note that only seven of them are unique since cluster AU02_A occurs twice).

CK+ dataset labels:	Fear	Disgust	Happy	Sadness	Surprise	Angry	Contempt	None	Entire Dataset
Num. of sequences with emotion label	25	59	69	28	83	45	18	266	593
Avg. num. of frames per sequence	21,84	14,71	19,27	19,53	16,0	22,7	12,9	18,26	18,1
Num. of temporal prototypes (only AU co-occurrence considered)	22	58	56	28	36	43	16	214	455
Num. of temporal prototypes (AU co-occurrence and spectral clustering)	25	59	65	28	56	45	17	243	525
Avg. num. of stages (changes in co-occurrence or appearance) per temporal prototype	7,88	6,85	6,57	6,89	**3,4**	8,07	4,28	6,09	6,06
Avg. reduction in description length	64 %	52 %	65 %	63 %	**78 %**	62 %	65 %	66 %	66 %

Table 1

A summary of results, denoting the number of video sequences in the CK+ dataset per emotion label, the average number of frames per sequence, the number of temporal prototypes per emotion label based on Action Unit co-occurrence as well as the number of temporal prototypes per emotion label based on Action Unit co-occurrence combined with Action Unit-wise spectral clustering. Further, the summary contains the average number of stages per temporal prototype resulting from changes in Action Unit co-occurrence or appearance. The last row denotes the average reduction in description length per emotion label that results from using temporal prototypes based on spectral clustering to represent a video sequence instead of considering all frames of a video sequence. To highlight the level of aggregation by temporal prototypes, we denote high aggregation (dark green) and low aggregation (light green).

3.5. Generating a Visual Representation of Temporal Prototypes

Relevance maps can be normalized and mapped to a predefined color spectrum which allows a visualization of the results in the form of heatmaps. To achieve a consistent mapping, relevance values are normalized such that they are always within the interval of [-1, 1]. As a result, both, the highest and lowest relevance value, are always mapped to a specific color. In foresight of what is required to be able to provide plots for temporal prototypes, a heatmapping approach is suggested which is designed to emphasize the most relevant features for the decision. First, we remove all negative relevance from a relevance map. Then, we determine the sum of all positive relevance (relevance mass). Next, we compute the most important input features that contribute to x percent of the relevance mass (i.e. combining those features with the highest relevance score add up to x percent of the entire relevance mass). Only those top features are plotted in the heatmap, which we call a *Top-X-Relevance Projection*, where x determines the portion of the positive relevance mass. For clarification, a $x = 25$ means that only those features are plotted which have the highest relevance scores that add up to 25% of the entire positive relevance (25% quantile).

Figure 3 shows how the reduced visualization emphasizes the most relevant features compared to a traditional heatmap (here, the *BWR* color scheme was used). The heatmaps show a prediction for AU02 “outer brow raiser". All three visualizations indicate that the positive relevance is indeed located close to the target (i.e. the brow), however, only the Top-25- and Top-10-Relevance Projection clearly show the network indeed put special focus on the brow. In our visualizations we use the Top-25-Relevance Projection.

Clustering ensures that all relevance maps within one cluster are relatively similar. Consequently, also their heatmap representations are similar. Consider Figure 4 that illustrates how heatmaps, colored differently for each Action Unit, are put together to derive a sequence of prototypes (the temporal prototype).

4. Evaluation

In the first step, we checked whether different sequences of the same emotion share identical temporal prototypes. For that, table 1 compares the number of sequences grouped by emotion (first row) with the number of unique prototypes (third row). We see that for emotion "Surprise" several sequences share the same temporal prototype and therefore reach the highest aggregation rate, namely 78 %. The most common prototype in "Suprise" is (AU01_A, AU02_A, AU25_A). We found that all common prototypes only include variations of AU01, AU02 and AU25 clusters, which aligns with the findings of Du et al. [23]. An illustration of temporal prototypes for "Surprise" is provided in Figure 5.

For other emotions the level of aggregation is lower. Note that two prototypes are only considered identical if they share the exact same sequence of transitions into the exact same stages. This is quite restrictive as even a small difference will create two distinct temporal prototypes. This is reflected in the results of table 1 since for all other emotions the number of unique prototypes is roughly equal (or even exactly even) to the number of sequences (third and fourth row).

Nevertheless, this indicates that the temporal prototypes preserve the individual differences of subjects with respect to displaying emotions (e.g., appearance and intensity of Action Units). We re-run the experiments without clustering among individual Action Units, considering only the occurrence of Action Units as predicted by the neural network. The results are shown in the third row of Table 1 and indicate that, indeed, in this case, temporal prototypes may occur multiple times across sequences. However, without clustering, individual differences in the faces of subjects may be disregarded, although the network may have deemed them relevant.

Since temporal prototypes are constructed such that only similar predictions are grouped together, it can be seen as a generalization of frame-based predictions into a concise, transition-based representation. Table 1 compares the average frames (second row) of a sequence to the average number of stages of the corresponding temporal prototype for the CK+ dataset (fifth row). Even though the sequences of the CK+ dataset are relatively small (approx. 18 frames per sequence), describing the sequence as a temporal prototype reduces its description length on average by approximately 66% with on average approximately six stages for the entire dataset.

Our comparison of the CK+ and the Actor Study dataset showed that sequences in the AS dataset are significantly longer. On average, a sequence of the AS dataset includes approximately seven times more frames than a sequence of the CK+ dataset (129 versus $\sim$18 frames respectively). However, the average length of the temporal prototype only increases by a factor of $\sim$3 from 6 in the CK+ dataset to 18 in the AS dataset. As a result, the temporal prototype provides an even more concise description with an average length reduction of 86%.

So far, only the identity of prototypes has been tested. Since the shape of a temporal prototype changes by a single "misjudgment" in only one frame of a sequence, a more relaxed approach could interpret the temporal prototypes as sets which allows to define an inclusion relation, e.g.:

$$(\text{AU01 A, AU02 A}) \rightarrow (\text{AU02 A}) \subset (\text{AU01 A, AU02 A}) \rightarrow (\text{AU02 A, AU25 A})$$

Thus, common basic temporal prototypes could be found that are included in many, more specific, prototypes.

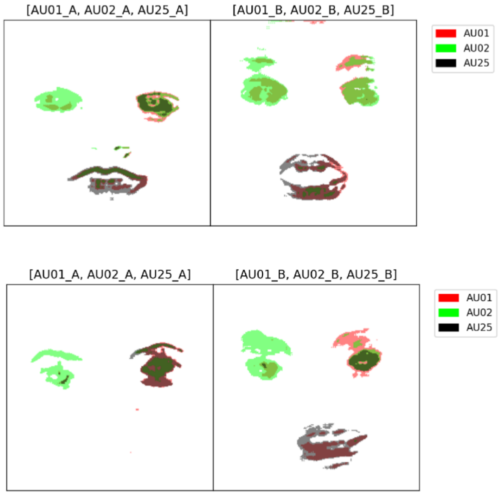

Figure 5: Two examples of two-staged temporal prototypes for “surprise” (CK+)

5. Conclusion

With the motivation to explain a deep neural network’s decision, two approaches from the field of XAI, namely Layer-wise Relevance Propagation and Spectral Relevance Analysis, have been combined to create temporal prototypes for Action Unit classification. Temporal prototypes serve as aggregated explanations based on what the neural network deemed relevant for a class.

To give an answer to the initial research question of whether there are temporal prototypes, the evaluation of the CK+ dataset showed that the higher level patterns present in sequences (i.e. temporal prototypes in sequences of similar emotions) are rather specific to video sequences, however, there are clear patterns within sequences that are usable to aggregate information. Meaning that correlated predictions are actually similar enough such that they can be aggregated, which is validated by both the CK+ and a subset of the AS dataset.

Independent of possible higher level interpretations, application to the CK+ and AS dataset shows that the temporal prototype aggregates information within a sequence well. A novel visualization approach that only plots a certain percentage of the positive relevance mass has been applied to create a visualization tool for a simultaneous, multi-class prediction. These results are a first step towards providing a local explanation for an entire video sequence. Future research may analyze temporal prototypes more thoroughly to identify common sub-sequences that are included in many of the identified prototypes. Further, the exact level of interpretability and usefulness of the visual representation of a temporal prototype either to the network developer or even to a medical expert has to be determined in the future.

Acknowledgments

The work presented in this paper is funded by grant FKZ 01IS18056 A/B, BMBF ML-3 (TraMeExCo) and by grant DFG (German Research Foundation) 405630557 (PainFaceReader).

References

[1] P. Ekman, The argument and evidence about universals in facial expressions, Handbook of social psychophysiology (1989) 143–164.

[2] I. Rieger, J. Pahl, D. Seuss, Unique class group based multi-label balancing optimizer for action unit detection, 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (2020).

[3] D. Gunning, Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2017).

[4] G. Schwalbe, B. Finzel, XAI method properties: A (meta-)study, CoRR abs/2105.07190 (2021). URL: https://arxiv.org/abs/2105.07190. arXiv:2105.07190.

[5] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature communications 10 (2019) 1–8.

[6] B. Martinez, M. F. Valstar, B. Jiang, M. Pantic, Automatic analysis of facial actions: A survey, IEEE transactions on affective computing 10 (2017) 325–347.

[7] R. Zhi, M. Liu, D. Zhang, A comprehensive survey on automatic facial action unit analysis, The Visual Computer 36 (2020) 1067–1093.

[8] D. Deng, Z. Chen, B. E. Shi, Multitask emotion recognition with incomplete labels, 2020. arXiv:2002.03557.

[9] Z. Chen, R. Ansari, D. Wilkie, Automated pain detection from facial expressions using facs: A review, arXiv:1811.07988 (2018).

[10] D. Dupré, E. G. Krumhuber, D. Küster, G. J. McKeown, A performance comparison of eight commercially available automatic classifiers for facial affect recognition, Plos one 15 (2020) e0231968.

[11] D. Kollias, A. Schulc, E. Hajiyev, S. Zafeiriou, Analysing affective behavior in the first abaw 2020 competition, arXiv preprint arXiv:2001.11409 (2020).

[12] J. Saito, R. Kawamura, A. Uchida, S. Youoku, Y. Toyoda, T. Yamamoto, X. Mi, K. Murase, Action units recognition by pairwise deep architecture, CoRR abs/2010.00288 (2020). URL: https://arxiv.org/abs/2010.00288. arXiv:2010.00288.

[13] F. Kuhnke, L. Rumberg, J. Ostermann, Two-stream aural-visual affect analysis in the wild, 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (2020).

[14] A. Heimerl, K. Weitz, T. Baur, E. Andre, Unraveling ml models of emotion with nova: Multi-level explainable ai for non-experts, IEEE Transactions on Affective Computing (2020).

[15] W.-S. Chu, F. De la Torre, J. F. Cohn, Learning spatial and temporal cues for multi-label facial action unit detection, in: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), IEEE, 2017, pp. 25–32.

[16] K. Zhao, W.-S. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3391–3399.

[17] E. Sánchez-Lozano, G. Tzimiropoulos, M. F. Valstar, Joint action unit localisation and intensity estimation through heatmap regression, in: British Machine Vision Conference

2018, BMVC 2018, Newcastle, UK, September 3-6, 2018, BMVA Press, 2018, p. 233.

[18] I. Ntinou, E. Sanchez, A. Bulat, M. Valstar, Y. Tzimiropoulos, A transfer learning approach to heatmap regression for action unit intensity estimation, IEEE Transactions on Affective Computing (2021).

[19] I. Rieger, R. Kollmann, B. Finzel, D. Seuss, U. Schmid, Verifying deep learning-based decisions for facial expression recognition, in: 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2020, Bruges, Belgium, October 2-4, 2020, 2020, pp. 139–144.

[20] R. Kollmann, Explaining Facial Expressions with Temporal Prototypes, Master's thesis, University of Bamberg, 2020.

[21] Y.-L. Tian, T. Kanade, J. F. Cohn, Facial expression analysis, in: Handbook of face recognition, Springer, 2005, pp. 247–275.

[22] B. Farnsworth, Facial action coding system (facs)—a visual guidebook, https://imotions.com/blog/facial-action-coding-system/. Accessed on July 2 (2016) 2018.

[23] S. Du, Y. Tao, A. M. Martinez, Compound facial expressions of emotion, Proceedings of the National Academy of Sciences 111 (2014) E1454–E1462.

[24] J. Pahl, I. Rieger, D. Seuss, Multi-label learning with missing values using combined facial action unit datasets, The Art of Learning with Missing Values Workshop at International Conference on Machine Learning (ICML) (2020).

[25] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 ieee computer society conference on computer vision and pattern recognition-workshops, IEEE, 2010, pp. 94–101.

[26] D. Seuss, A. Dieckmann, T. Hassan, J.-U. Garbas, J. H. Ellgring, M. Mortillaro, K. Scherer, Emotion expression from different angles: A video database for facial expressions of actors shot by a camera array, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 35–41.

[27] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[28] S. Lapuschkin, Opening the machine learning black box with layer-wise relevance propagation, Dissertation (2019).

[29] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-Wise Relevance Propagation: An Overview, Springer International Publishing, Cham, 2019, pp. 193–209.

[30] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, S. Lapuschkin, Towards best practice in explaining neural network decisions with lrp, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[31] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, P.-J. Kindermans, innvestigate neural networks!, J. Mach. Learn. Res. 20 (2019) 1–8.