Challenges in face expression recognition from video

Sergey Redyuk[1] and Erik A. Billing[1]

[1]Interaction Lab, University of Skövde, Sweden
{*sergey.redyuk, erik.billing*}*@his.se*

Identification of emotion from face expressions is a relatively well understood problem where state-of-the-art solutions perform almost as well as humans. However, in many practical applications, disrupting factors still make identification of face expression a very challenging problem. Within the project DREAM[1] - Development of Robot Enhanced Therapy for Children with Autism Spectrum Disorder (ASD), we are identifying face expressions from children with ASD, during therapy. Identified face expressions are used both in the online system, to guide the behavior of the robot, and off-line, to automatically annotate video for measurements of clinical outcomes.

This setup puts several new challenges on the face expression technology. First of all, in contrast to most open databases of face expressions comprising adult faces, we are recognizing emotions from children between the age of 4 to 7 years. Secondly, children with ASD may show emotions differently, compared to typically developed children. Thirdly, the children move freely during the intervention and, despite the use of several cameras tracking the face of the child from different angles, we rarely have a full frontal view of the face. Fourthly, and finally, the amount of native data is very limited.

Although we have access to extensive video recorded material from therapy sessions with ASD children, potentially constituting a very valuable dataset for both training and testing of face expression implementations, this data proved to be difficult to use. A session of 10 minutes of video may comprise only a few instances of expressions e.g. smiling. As such, although we have many hours of video in total, the data is very sparse and the number of clear face expressions is still rather small for it to be used as training data in most machine learning (ML) techniques.

We therefore focused on the use of synthetic datasets for transfer learning, trying to overcome the challenges mentioned above. Three techniques were evaluated: (1) convolutional neural networks for image classification by analyzing separate video frames, (2) recurrent neural networks for sequence classification to capture facial dynamics, and (3) ML algorithms classifying pre-extracted facial landmarks.

The performance of all three models are unsatisfactory. Although the proposed models were of high accuracy, approximately 98%, while classifying a test set, they performed poorly on the real-world data. This was due to the usage of a synthetic dataset which had mostly a frontal view of faces. The models which have not seen similar examples before failed to classify them correctly. The accuracy decreased drastically when the child rotated her head or covered a part of her face. Even if the frame clearly captured a facial expression, ML algorithms were not able to provide a stable positive classification rate. Thus, elaboration on training datasets and designing robust ML models are required. Another option is to incorporate voice and gestures of the child into the model to classify emotional state as a complex concept.

[1]DREAM is funded by the 7th Framework Programme of the EU (grant #611391). www.dream2020.eu