Multimodal Emotion Recognition: RTL & UU

Dr. Daan Odijk & Dr. Hendrik Vincent Koops
Lead Data Scientist & Data Scientist
RTL Nederland

Prof. dr. Albert Salah
Professor Social and Affective Computing
University Utrecht

About RTL
RTL Nederland is the largest commercial broadcaster in the Netherlands, with a yearly turnover of around €350m. RTL has the mission to tell unmissable stories that touch the heart and mind. With a team of 10 data scientists, RTL works on challenges such as personalization, prediction and metadata generation.

Our Challenge
Every day, RTL produces many hours of video content that aims to touch the heart and mind of viewers. Understanding the emotional content of this video content is a critical part of how we tell stories. A better understanding of emotion in video content will allow us to produce better content, improve storytelling and unlock new use cases, thereby benefiting both RTL and its viewers.

In particular, we see a clear opportunity to improve video production by facilitating the making of content that better matches viewers. For example, we know from directors that using emotional salient material is an integral part of the production of promotional content, as emotion compels attention and elicits arousal. A better understanding of emotional content would on the one hand facilitate directors in finding emotional salient material, and on the other hand unlock an important aspect of the automatic generation of promotional video content.

In first pilots, we have found that current cloud solutions such as Google Cloud Video Intelligence [1] and Microsoft Video Indexer [2] can be leveraged to provide emotional metadata based on audiovisual content. However, these do not reach the level of detail nor the emotional depth that our use cases call for. Many existing models either limit themselves to analyzing a single modality or only a particular dimension of emotion (such as positive versus negative). The challenges from the MediaEval 2016-17-18 benchmarking initiative included an Emotional Impact of Movies Task. For this task a dataset of 160 professionally made and amateur movies were annotated for fear, valence and arousal. To illustrate the difficulty of this task, the fear task in MediaEval was considered unsuccessful, as it was both rare and very difficult to properly model.

We are seeking to close the gap in vocabulary between our use cases and automatically generated metadata. To achieve this, models that understand content on a deeper semantic level, including emotional expressions across modalities are needed. Combining the content and applied expertise of RTL with the academic expertise of Prof. Salah will make for exactly the right atmosphere for tackling this challenge.

[1] https://cloud.google.com/video-intelligence/
[2] https://vi.microsoft.com/
[3] http://www.multimediaeval.org/mediaeval2018/emotionalimpact/index.html

Input for workshop participants
Before the workshop, we will invite interested participants for a preparatory afternoon at RTL, giving them a first introduction of the use case, a tour of the RTL and a studio, and an informal chat with some of the intended end users of the solution.

As working with video content is computationally intensive, we will prepare a first baseline approach before the workshop, if possible, with input from the participants. We can provide the participants with a selection of video material, subtitles and the added metadata as training material. We will also look at the aforementioned MediaEval task and other publicly available datasets to ensure potential for reuse beyond our use case.

For our daily soap opera GTST we have a video dataset with shot-level emotion annotated on 8 emotion categories. This dataset consists of around 1.000 annotated shots and will be available for participants.

Where needed and relevant we can provide participant with the output of the aforementioned cloud video analysis services and other general-purpose text, audio and visual content descriptors. For these pre-computed features, we consider ImageNet-trained deep neural net features per frame, face detection bounding boxes, and if there is a face, AffectNet output for the face, and OpenSmile sound features sampled regularly from the content, and NLP features derived from subtitles.