Multimodal Emotion Recognition for Visualizing Storyline in a TV Series

Tanja Crijns (1), Metehan Doyran (2), Maurits van der Goes (1), Cecilia Herrera (3), Heysem Kaya (2),
Osman Semih Kayhan (4), Rana Klein (1), Vincent Koops (1), Cas Laugs (1,2), Daan Odijk (1),
Albert Ali Salah (2), Alexander Serebrenik (5), Yasemin Tımar (6), Anja Volk (2),
(1) RTL Netherlands
(2) Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
(3) Noldus Information Technology b.v., The Netherlands
(4) Delft University of Technology, Delft, The Netherlands
(5) Eindhoven University of Technology, Eindhoven, The Netherlands
(6) Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
Tanja.Crijns@rtl.nl, m.doyran@uu.nl, Maurits.van.der.Goes@rtl.nl, c.herrera@noldus.nl,
h.kaya@uu.nl, o.s.kayhan@tudelft.nl, rana.klein@rtl.nl, Vincent.Koops@rtl.nl,
c.a.j.laugs@students.uu.nl, Daan.Odijk@rtl.nl, a.a.salah@uu.nl, a.serebrenik@tue.nl,
yasemin.timar@gmail.com, a.volk@uu.nl

***Abstract*—Automatic analysis of video archives is a topic long-researched in multimedia. In this work, conducted with RTL Netherlands, we investigated methods for developing an integrated tool for analysis and visualization of the storyline in a TV series by combining a range of technologies in affective computing and multimedia analysis. The input to the proposed system is a set of episodes from a TV series, in proper temporal order, including subtitles. We analyse the input in audio, video, and text modalities, and identify characters in each scene. We accumulate information about the interactions of the characters and create an interactive visualisation that helps visualizing the episodes of the series, as well as accessing specific information. Our results are potentially useful for creating a tool that will help directors in creating promotional material, for multimedia summarization, and for creating visual interfaces into multimodal archival material. We also analyze the language of soap operas, how music and sound are used, and how different modalities are used to create certain affective results.**

***Index Terms*—Affect Analysis, Multimedia, Information Retrieval.**

I. Introduction

Automatic summarization of multimedia material facilitates access to large multimedia archives, helps understand the temporal dynamics in the material, and offers new perspectives to media scholars. Furthermore, some tasks that rely on indexing and retrieval can be accomplished much faster through such summaries, especially if they are connected to interactive visualizations linked to the actual content. In this report, we describe a concept that uses off-the-shelf tools for multimodal affect analysis to analyse and visualize a TV series. Our case study is a popular Dutch soap opera called Goede Tijden, Slechte Tijden (Good Times, Bad Times), which is the longest-running Dutch soap opera, running since 1 October 1990 on RTL4[1].

Our aim in this report is to create a set of tools that will help index affective content and personal interactions in multimedia material. This type of analysis can help in accessing the emotional content of many hours of video RTL NL produces every day. We know from directors that using emotionally salient material is an integral part of the content production.

Fig. 1 illustrates a hand-drawn infographic about how a movie or a TV series can be turned into an interaction graph. This can further be enhanced with emotional tags, as well as place markers, and serve as a guide to AI=based replacements to time-consuming tasks, such as automatic summarization and promo generation.

The UXMood tool, while developed for a different purpose, has a similar information processing pipeline with the approach we take here [2]. In this tool, a user's interaction with an interface is analyzed via multiple modalities, and the results are combined into a temporal visualization that illustrates the affect changes of the user. The main differences are that this study uses a single person, and the environment is a controlled, frontal setting with little variation, whereas in multimedia material, the acquisition conditions are much more challenging.

In this report, we describe our approach to the problem, and the set of tools we have worked on during the Lorentz Workshop in Leiden. Section II describes the multimedia material we worked on. In Section III, we provide details into the modules of the system, and in Section IV provide some experimental results. Section V summarizes our progress and concludes the report.

II. Dataset

The dataset we worked on consisted of 30 episodes from the 2017 season of the Dutch soap opera "Goede Tijden, Slechte Tijden" (episodes 5491–5520). Each episode runs for

[1] https://en.wikipedia.org/wiki/Goede_tijden,_slechte_tijden

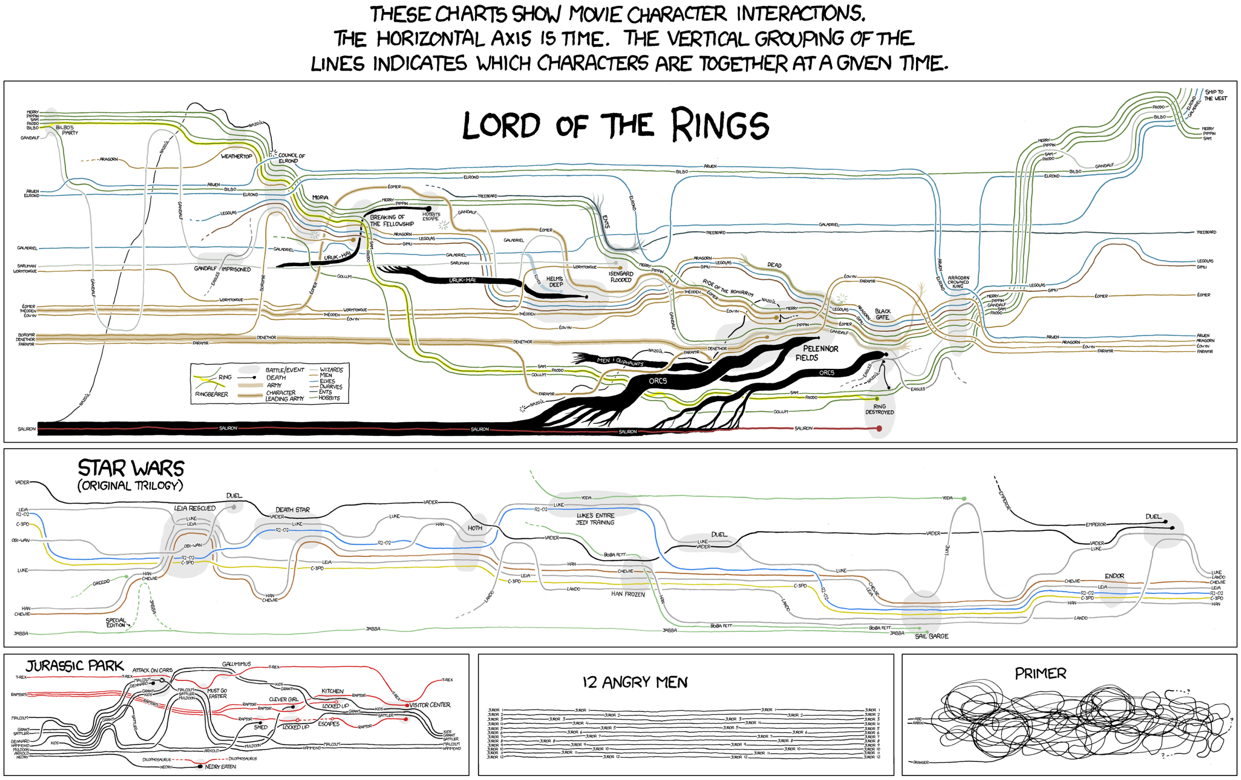


Fig. 1. Movie interaction graphs, created by Randall Munroe [1].

23 minutes and consists of three parts: overview of the relevant fragments from the previous episodes, the main part of the episode and the preview of the next episode.

The dataset has been provided by RTL Netherlands. In addition to the above dataset, a 1000-shot annotated corpus from the same series is provided. Annotated shots' length varies from 3 to 21 seconds. This corpus was annotated by RTL Data Science team using video only, subtitles only, video and sub-titles modalities. This corpus is used for audio-based emotion recognition after filtering out shots having multi-emotion annotations.

III. Methodology

The analysis is done on multiple modalities such as face, text, acoustic, speech, music. Scene classification is implemented for detecting scene boundaries in time domain which is crucial for fine-grained analysis of episodes over scenes. Additionally we train a shot type classifier which gives insights about the cinematographic decisions of the film maker.

A. Face detection and recognition

The first module uses an off-the-shelf approach for face detection and recognition. For face detection, MTCNN is used which is a implementation of David Sandberg's (FaceNet's MTCNN) in Facenet. It is based on the paper Zhang, K et al. [3]. Detected faces are cropped and fed to face recognition and emotion detection pipeline.

For face recognition, at first, separate pipeline is executed to calculate face embeddings of all the actors (male & female) in the episode. Actors face photos which are avaliable online in various sites are used to build this dataset. After that each cropped face in each frame is matched with the most similar face embedding in actor dataset. This recognition models are based on https://github.com/ageitgey/face_recognition

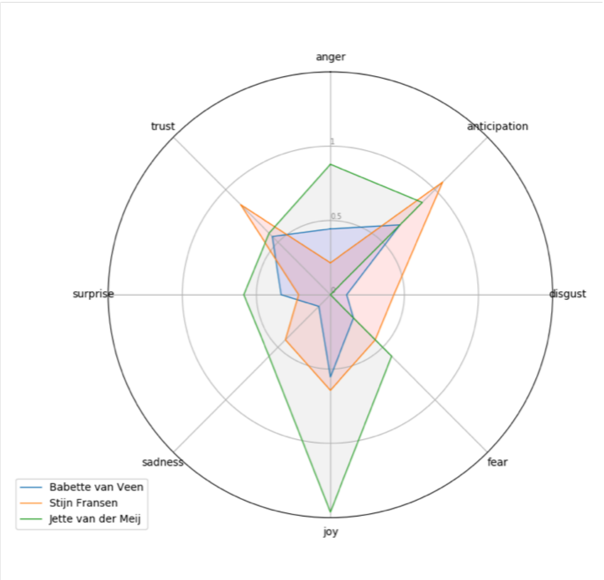


Fig. 2. Characters can be compared by their emotional displays and appearance characteristics during the episodes.

B. Face emotion detection

For each cropped face in the frame, emotion is estimated with https://github.com/oarriaga/face_classification. This tool categorizes each face based on the emotion shown in the facial expression: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral. Emotion depicted by selected characters were visualized over the shot number corresponding to each frame to show the change of emotions of a character over time. This can be used to compare emotions of characters interacting at the same time.

C. Acoustic emotion detection

From the audio signal of the annotated 1000-shot GTST dataset, we have extracted a standard set of acoustic suprasegmental features using freely available openSMILE [4] toolkit with INTERSPEECH 2013 configuration. The same set of features are extracted from publicly available speech emotion datasets in other languages, such as eNTERFACE (English), DES (Danish), EMODB (German) and RUSLANA (Russian) to assess cross-corpus and cross-language acoustic emotion recognition performance. To minimize the mismatching acoustic characteristics that may stem from various factors such as identity, language and recording conditions, we used a cascaded feature normalization scheme as in [5].

D. Text emotion detection

The dataset provided recorded polarity and subjectivity scores for each subtitle computed using Pattern [6]. We enrich the dataset by including more fine-grained emotion scores using the Dutch version of the NRC lexicon [7]. The lexicon contains 14182 Dutch words and expressions manually annotated with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust).

E. Scene classification

To classify the scenes, two different approaches are conducted: (i) unsupervised, (ii) supervised classification.

1) Unsupervised scene classification: First, features of each shot from the video are obtained by using (imagenet pretrained [8]) Resnet18 [9] network. Instead of using fine-grained features, we used intermediate features to capture global information about the scene. The high dimensional features from each shot are visualized by t-sne [10] algorithm. Afterwards, K-means [11] clustering is applied on those selected features by t-sne to classify the scenes.

2) Supervised scene classification: In this part, we used Resnet18 pretrained network with Places 365 dataset [12] to classify the scene. Random frames are chosen from the video and tested with pretrained network. Because the real place labels are not available for the dataset, the evaluation is proceeded qualitatively.

F. Shot type classification

In film and TV grammar shot types are defined with body parts visible within the frame. One can assume that face is the most important part that can be used to figure out how much of actor's body is visible. For practical reasons, we have defined constant thresholds for face to frame ratio to define the different types of shots (”1-Extreme Long Shot”, ”2-Long Shot”, ”3-Medium Long Shot”, ”4-Medium Shot”, ”5-Medium Close-Up”, ”6-Close-Up”, ”7-Extreme Close-Up”,) taken in episode. Apparently, if there is no face we assume it is an extreme long shot.

G. Visualization

WordCloud python package [13] is utilised for scene, episode, and season level visualisations of the word frequencies from the subtitles. Subtitles are preprocessed by applying stemming and stop word removal models trained on Dutch corpora.

IV. Experimental Results

We have focused on Episode 5492. Fig. 11 shows the actors detected frame by frame, using a combination of face recognition and speech recognition. It shows that the soap opera has very few sequences without a face visible on the screen.

A. Face emotion and apparent personality detection

Figure 4 shows categories of emotions (angry, fear, happy, neutral, sad, and surprise) detected from selected characters over time. The characters selected were Stijn Franssen, Erik de Vogel and Babette van Veen, respectively. The dots in each category show the periods when faces were detected. The heights of the dots show the likelihood (probability) each emotion was detected. The higher the probability, the more likely the particular emotion was shown by the character. Comparing different categories per time period, once can see the progression of emotion shown by a character. By comparing the emotions depicted by two characters in the same time period, one can see the interaction of characters in the given time.

B. Speech and sound emotion detection

Applying a within-corpus evaluation using the annotated portion we obtained binary arousal and valence average recall performances of 62.7% and 56.2%, respectively. The seven-class (composed of Ekman's six basic emotions and neutrality) classification performance is about 25%. These results indicate that the problem is highly challenging even in within corpus setting. We further attribute the low acoustic emotion recognition performance to the 'gold standard': a single annotator annotation (per file) without listening to audio. This could be improved by employing multiple (preferably three) annotators who should better observe all modalities to cast their decision.

Having noted the low within corpus performance, we next carried out the cross-corpus experimentation to reveal i) the predictive power of other corpora and ii) whether readily available corpora in other languages could be useful in improving performance in an 'in the wild' Dutch speech emotion corpus. The results are summarized in Figure 5. Here we observe that

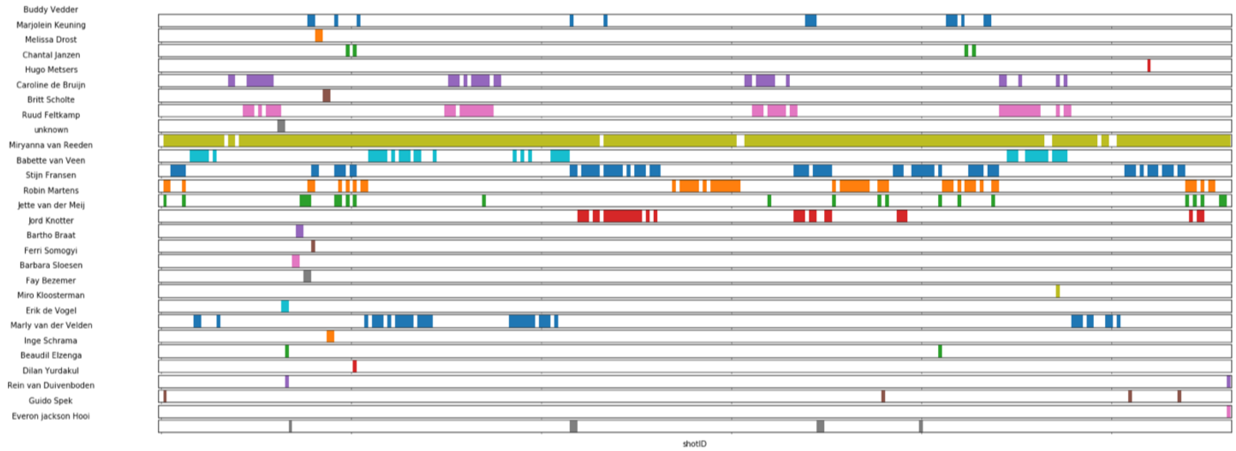


Fig. 3. Detected actors for one episode. The long yellow line is the 'unknown' class, where a face is detected during a shot, but the identity is not reliably determined.

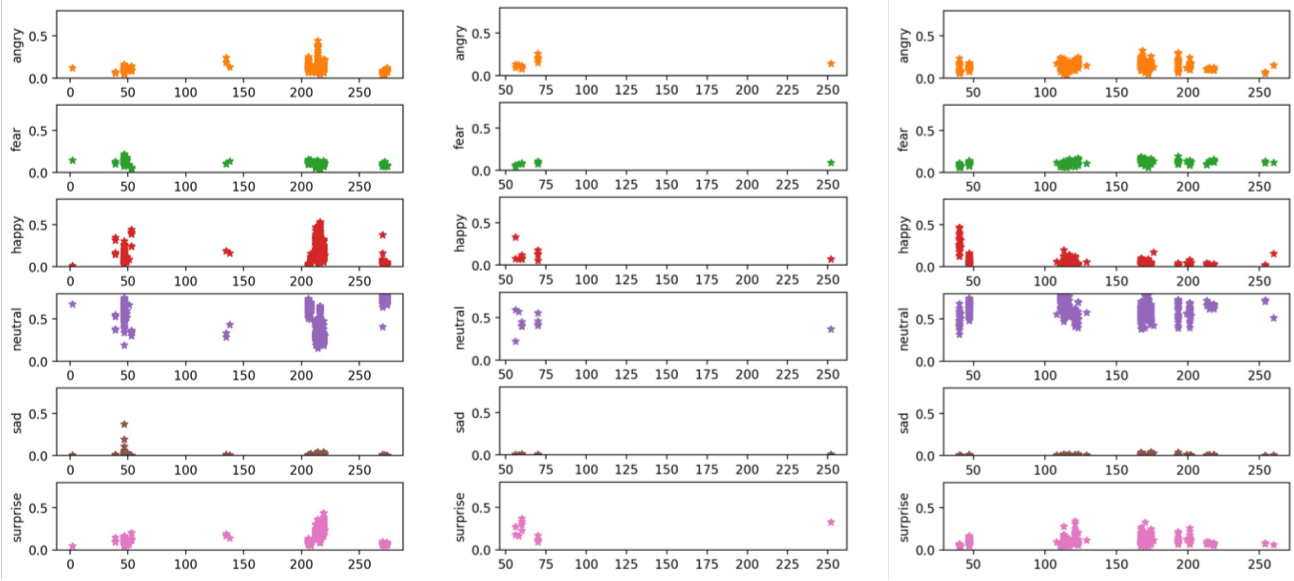
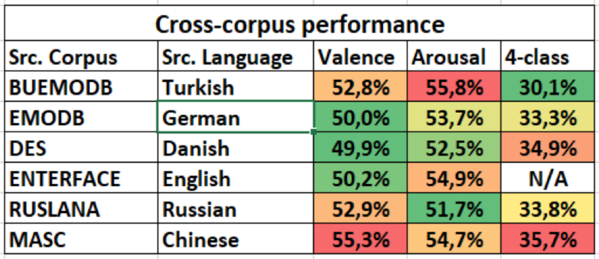


Fig. 4. Distribution of facial emotions for three characters in an episode. The x-axis denotes time (shot id) in the series. Gaps in vertical directions are periods with no detected faces for that person.

the performance is slightly over the chance level and hence under the gold standard of the test set, the models trained on other corpora do not provide sufficiently good results.



Cross-corpus performance

Src. Corpus	Src. Language	Valence	Arousal	4-class
BUEMODB	Turkish	52,8%	55,8%	30,1%
EMODB	German	50,0%	53,7%	33,3%
DES	Danish	49,9%	52,5%	34,9%
ENTERFACE	English	50,2%	54,9%	N/A
RUSLANA	Russian	52,9%	51,7%	33,8%
MASC	Chinese	55,3%	54,7%	35,7%

Fig. 5. Emotion recognition results from speech input. In the cross-corpus, cross-language setting, models are trained on corpora in various languages and tested on the annotated 1000-shot GTST corpus.

C. Text emotion detection

In episode 5492, 13% of the subtitles contains one or more words associated with emotions in the NRC lexicon. The most common emotions are joy, trust and anticipation, while the least common ones are disgust, fear and sadness.

D. Music analysis

Background music is present in the lion's share of the scenes, though often it is hardly noticeable, as the sound level is very soft in comparison to the speech sound level. The difficulty for human listeners to even decide whether music is playing or not has been reported in earlier studies on automatic music detection in television production [14], also showing that soap operas contain much more music than other television genres [14]. In general, it has been argued that the soundtrack is of greatest importance in soap operas [15], as it can be basically watched without looking at the screen,

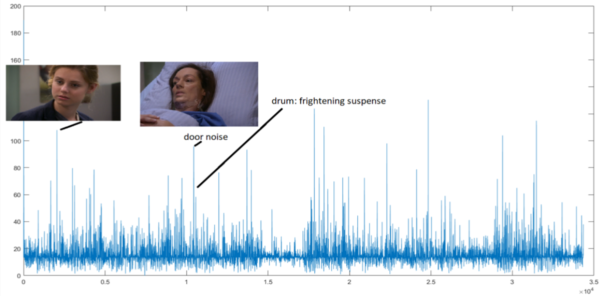


Fig. 6. The changes in loudness during an episode. Loud noises or sounds that induce suspense can be detected.

arguing for a “primacy of sound” [15] in TV productions such as soap operas.

Subtitles in “Goede tijden, slechte tijden” repeatedly refer to music used to build up suspense (e.g., “onheilspellende muziek” or “spannende muziek”) or to underline romantic moments (e.g., “romantische muziek”). Specific music instruments, e.g., piano or strings, are rarely mentioned.

As opposed to such soap operas as “Muhteşem Yüzyıl” [16], “Goede tijden, slechte tijden” does not make use of specific themes or leitmotifs associated with a certain character or situation. Similarly to de Bruin [17] we observe that non-diegetic music often connects several related scenes, stressing the relation between the scenes rather than indicating a break between scenes. This observation is in line with Butler’s [15] characterization of one important function of music in soap operas, namely the support of the narrative by ”counteracting segmentation” and hence connecting several different scenes by a continuous stream of musical events. A second important function of music in soap operas relating to emotions, such as setting the mood, expressing hidden emotions of characters and marking intense emotions, supports another main characteristic of soap operas, namely never arriving at a full closure or a full resolution of enigmas [15]. As argued in [15] “it is critical to soap opera form that emotions are never quite fully discharged, traces always linger”, hence music in “Goede tijden, slechte tijden” expresses these lingering emotions such as shown in subtitles (e.g., “onheilspellende muziek”).

E. Scene classification

Unsupervised scene classification of 2 scenes is obtained by using 15 different shots. Features from each scene are yielded by using first 3 blocks of Resnet18. t-sne is used to visualize the features in 2-dimensional plane (Figure 7). Afterwards, K-means algorithm is applied on selected features of t-sne to construct scene clusters (Figure 8). We obtained 74% of classification accuracy with unsupervised method. In addition, 55% of accuracy is achieved for 3 scenes from 27 different shots.

Additional, randomly chosen frames from different scenes only tested by using Resnet18 network pretrained with Places 365 dataset. The aim is to see if the network yields reasonable outputs. Only qualitative results are obtained as ground truth labels for scenes are not available. When the outputs of the network is checked visually, subjectively, the predictions are

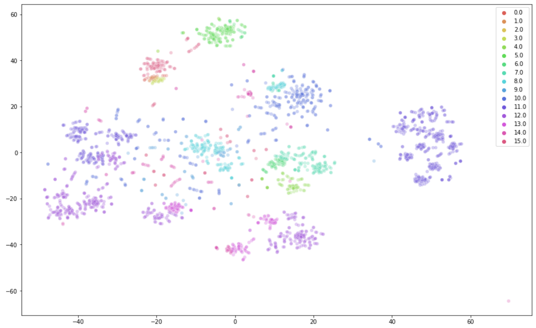


Fig. 7. t-sne visualization of 15 shots. Each color represents different shots.

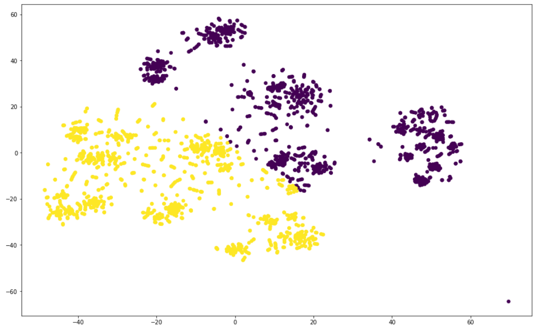


Fig. 8. K-means clustering. Yellow class and blue class demonstrates the prediction of scene-1 and scene-2 respectively.

acceptable. In Figure 9 bottom-right image, hospital room is correct classified even the actor is not in the scene. Conference center and art school frames are incorrectly classified.

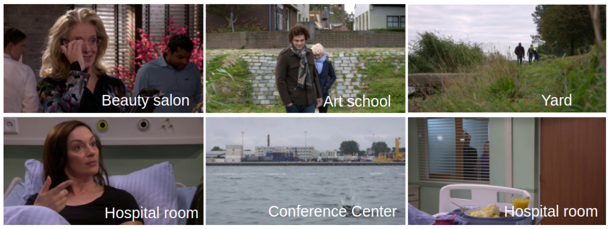


Fig. 9. Shot samples with supervised place classification outputs. Hospital rooms and yard images are correctly classified. Art school and conference center frames are misclassified. Beauty salon image does not include sufficient information whether it is correctly classified.

F. Shot type classification

Our preliminary experiments show that the shot type classifier is as successful as the pre-trained face detector we use. Since in TV series most of the shots consist of one or more people where their faces are frontal and close to the center area of the frame, face detector performs accurately compared to other in-the-wild datasets where faces are occluded or far from frontal angle.

V. Conclusion

Our multi-modal analysis and various visualisations show that sentiment and emotion play a prominent role in TV series.

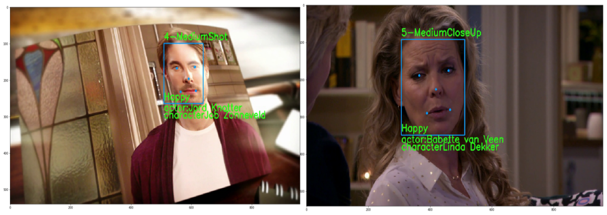


Fig. 10. Detection of actor faces and shot types, based on face size.

These techniques can further be combined into an interactive tool for analysing TV series or movies from different aspects. Shot type classification can be further improved into a cinematographic analyser which can provide more info such as camera angle, zooming speed, angle change speed. These features can provide rich indicators for the current mood of the scene since these cinematographic techniques are widely adopted by directors.

We have identified the following future work for our project:

- Each modality performance will be further improved prior to fusion.
- Videos will be re-annotated observing all modalities for better ground truth.
- Character-based video shots to be modeled for apparent personality traits.
- Background music processing will be investigated more:
  - An unusually promising, under-researched source for affect analysis
  - Novel problems in affective computing, like predicting suspense
  - Acoustic scene recognition to be fused with visual scene recognition
  - Contact composers and editors for getting insight into soap opera music.

Acknowledgments

Authors thank the organizers of the ICT with Industry workshop and gratefully acknowledge the hospitality and financial support of the Lorentz Center and NWO.

References

[1] R. Munroe, “Movie narrative charts (comic 657). courtesy of xkcd.com,” in *8th Iteration (2012): Science Maps for Kids, Places Spaces: Mapping Science*, K. B{´ orner and T. Theriault, Eds., 2009.

[2] R. Y. da Silva Franco, R. Santos do Amor Divino Lima, M. Paixão, C. G. Resque dos Santos, B. Serique Meiguins *et al.*, “Uxmood—a sentiment analysis and information visualization tool to support the evaluation of usability and user experience,” *Information*, vol. 10, no. 12, p. 366, 2019.

[3] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[4] F. Eyben, F. Weninger, F. Groß, and B. Schuller, “Recent developments in openSMILE, the Munich open-source multimedia feature extractor,” in *Proceedings of the 21st ACM International Conference on Multimedia (ACM MM)*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[5] H. Kaya and A. A. Karpov, “Efficient and effective strategies for cross-corpus acoustic emotion recognition,” *Neurocomputing*, vol. 275, pp. 1028 – 1034, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231217315680

[6] T. De Smedt and W. Daelemans, “Pattern for python,” *J. Mach. Learn. Res.*, vol. 13, pp. 2063–2067, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2343710

[7] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word-emotion association lexicon,” *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[11] A. David and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” in *18th annual ACM-SIAM symposium on Discrete algorithms (SODA), New Orleans, Louisiana*, 2007, pp. 1027–1035.

[12] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[13] A. Mueller, J.-C. Fillion-Robin, R. Boidol, F. Tian, P. Nechifor, yoonsubKim, Peter, R. Rampin, M. Corvellec, J. Medina, Y. Dai, B. Petrushev, K. M. Langner, Hong, Alessio, I. Ozsvald, vkolmakov, T. Jones, E. Bailey, V. Rho, IgorAPM, D. Roy, C. May, foobuzz, Piyush, L. K. Seong, J. V. Goey, J. S. Smith, Gus, and F. Mai, “amueller/word_cloud: Wordcloud 1.5.0,” Jul. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1322068

[14] K. Seyerlehner, T. Pohle, M. Schedl, and G. Widmer, “Automatic music detection in television productions,” in *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, 2007.

[15] J. G. Butler, “Notes on the soap opera apparatus: Televisual style and ”as the world turns”,” *Cinema Journal*, vol. 25, no. 3, pp. 53–70, 1986.

[16] K. Bowen Çolakoğlu, R. Reigle, and Ş. Beçiroğlu, “Magnificent music: Identity in turkish soap opera soundtracks,” *Porte Akademik Müzik ve Dans Araştırmaları Dergisi*, vol. 13, pp. 53–59, 2016.

[17] M. E. J. M. de Bruin, “Het gehoor als navigator,” 2011, Bachelor’s Thesis, University of Utrecht, The Netherlands.

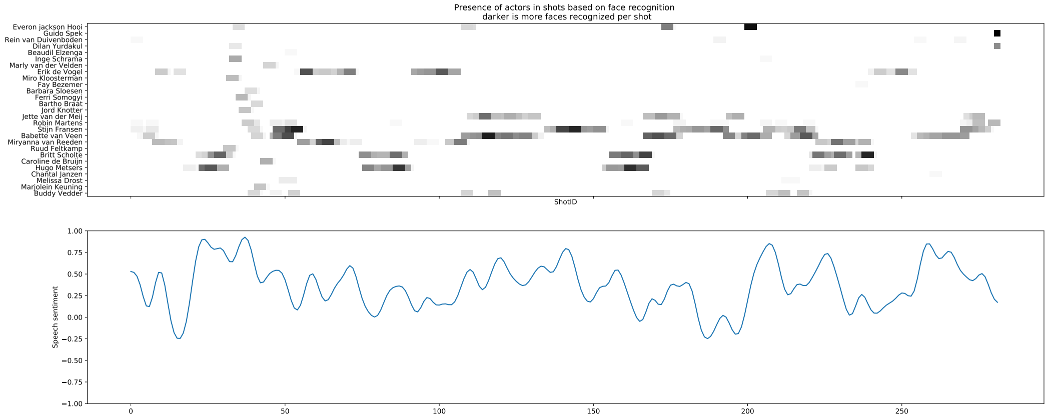


Fig. 11. Top: Presence of actors in shots based on face recognition, where darker colors mean more faces are recognized per shot. Bottom: Speech sentiment, where higher values denote positive sentiments.