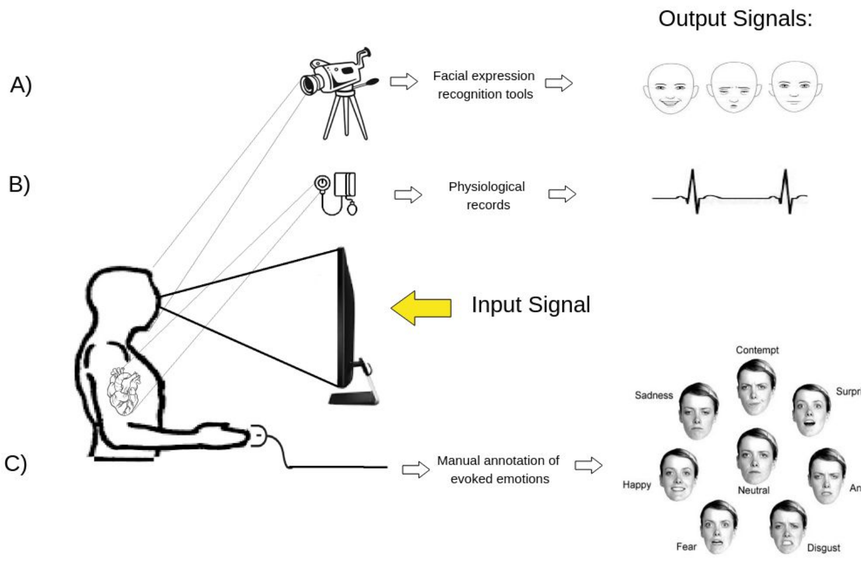
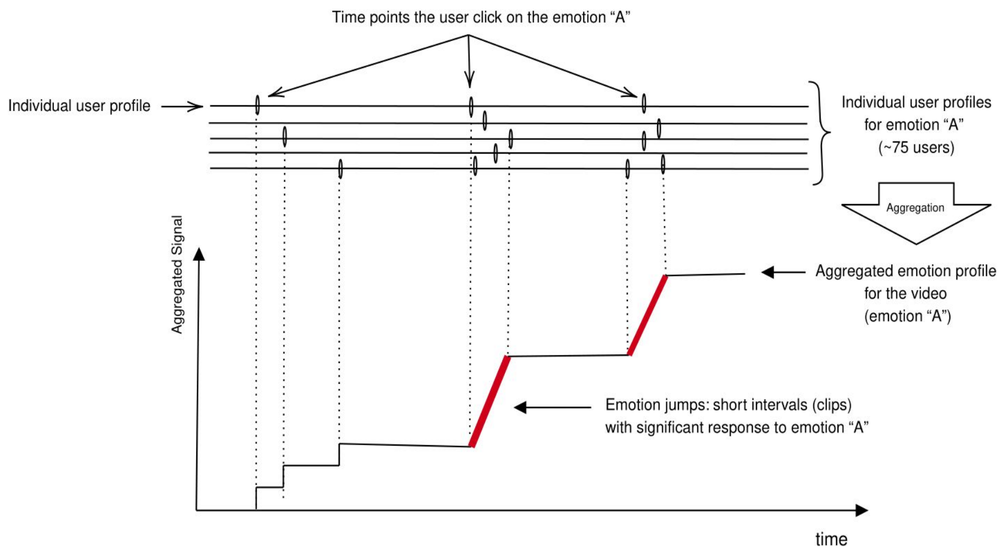
Predicting emotional responses triggered by video adverts

Alexey Antonov[1], Steeve Wood[2], William Headley[2] and Giovanni Montana[1],
[1]WMG University of Warwick, [2]System1

Being able to predict the emotions that a video clip is likely to evoke from its viewers is a challenging and still largely unsolved problem. An automated video captioning system capable of indicating frame-by-frame which emotions are expected to be experienced by viewers offers several benefits, from enabling improved video indexing and summarisation to delivering mood-based personalised content and assisting with advertising content generation [1]. Here we present initial efforts made towards developing such a system through state-of-the-art deep learning methodologies for video analytics. We introduce analyses of a large scale dataset of video ads manually annotated with emotions evoked by the video content at different time points. The dataset comprises 30 thousands video ads annotated on average by 75 viewers with 8 standard emotions resulting in 2.3 millions of emotion profiles (a pair of one video annotated by one viewer). We translate this data into a video classification problem by defining for each of the 8 emotions a representative set of short (5 seconds) video clips that have evoked significant response. We refer to them as ”emotion jumps”. We have extended a state-of-the-art deep learning architecture for video understanding with audio modality to capture specific needs of evoked emotion recognition. We demonstrate that audio modality is important and improves performance for classification of ”emotion jumps”. Finally we evaluate the ability of our trained neural network to localize ”emotion jumps” in the full length video ads. The classification performance for some emotions (Happiness, Sadness) was promising for practical applications.



Comparison of our study with other studies:
A) A synchronized recording of viewer facial expression is used to convert it into emotional response to the watched video content (commonly done automatically by available facial emotion expression recognition tools) [2]. B) A synchronized recording of various physiological signals is used as a response variable (a surrogate information about emotions the viewer is feeling). C) In our study we use manually annotated videos with evoked emotions by ***face tracing*** platform.



Adcumen data: signal quantification schema:
Each video was annotated by n users (75 on average). Each user profile for each emotion represent a list of time points (clicks) where the user have indicated the emotion. Each click of a user increases the value of aggregated signal by 1/n. Short intervals with significant number of clicks are referred to as emotion jumps (shown in red).

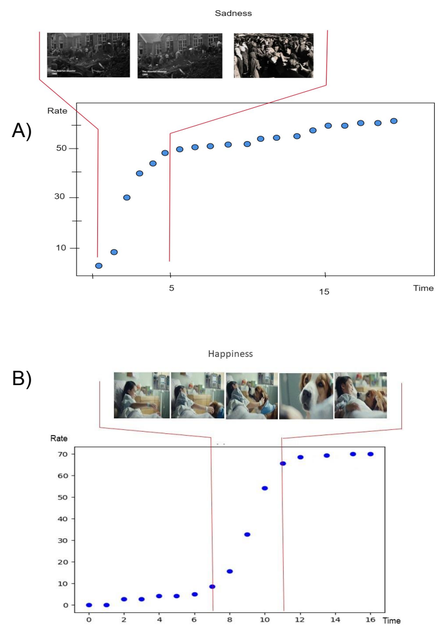


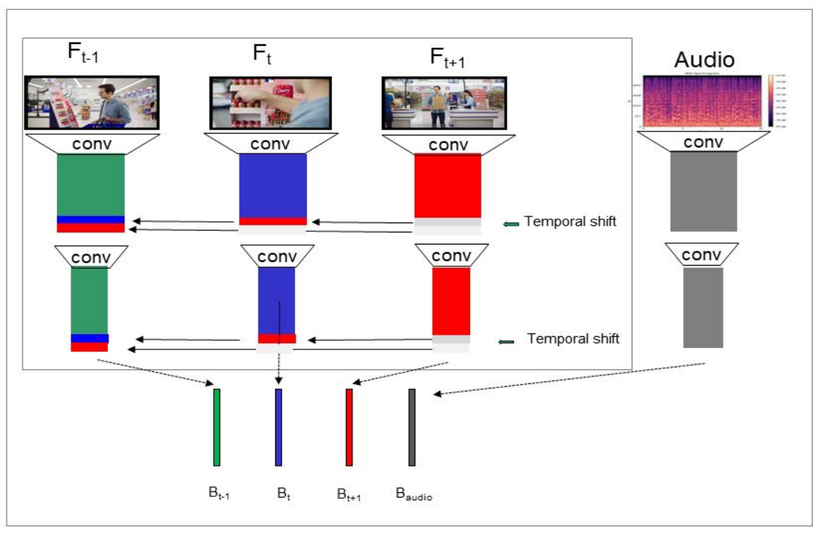
Illustration of emotion jumps:
A) between 0 and 5 seconds in about 50\% of users indicate Sadness
B) between 7 and 12 seconds in about 55% of users indicate Happiness

Translation to video classification problem:

We translated ***adcumen*** *data* into 8-class (by the number of annotated emotions) video classification problem. Each class is represented by 5 second clips that evoked significant response according to adcumen data. So the input for classification represents the set of 5 seconds clips labeled with one emotion. The problem has 8 classes by the number of annotated emotions. Table 1 reports the number of clips selected for each class (emotion).

To solve the this problem we use state-of-the-art deep learning architecture from the action recognition domain [3] extended with audio modality (see figure Video Classification Module) .

VCM uses TSM resnet50 backbone to fuse (shift) features between temporally neighboring frames. Audio input (Mel-Spectrogram) is processed by the same resnet50 backbone but without shifting of features.



Video Classification Module (VCM)

Emotion	Threshold	Clips (Videos)
Anger	8.1	3739 (1645)
Contempt	10.6	3651 (1385)
Disgust	10.6	3662 (828)
Fear	8.1	3692 (787)
Happiness	32.5	3608 (1488)
Neutral	16.7	3658 (1394)
Sadness	22.2	3590 (859)
Surprise	22.7	3693 (1330)

Table 1. Parameters used to translate acumen data into video classification problem.

Modality	Frames	PreTrained	Accuracy (test)
RGB	8	imagenet	38.9
RGB +audio	8	imagenet	43.2
RGB	16	imagenet	38.8
RGB +audio	16	imagenet	42.2
RGB	8	INET21K	38.9
RGB +audio	8	INET21K	43.5
RGB	16	INET21K	40.9
RGB +audio	**16**	**INET21K**	**43.6**

Table 2. Emotional jumps: classification results. Balanced classification accuracy (8-class) are reported for test set.

Emotion	Accuracy
Anger	28.4
Contempt	42.3
Disgust	26.7
Fear	50.7
Happiness	55.8
Neutral	47.8
Sadness	60.2
Surprise	51.1

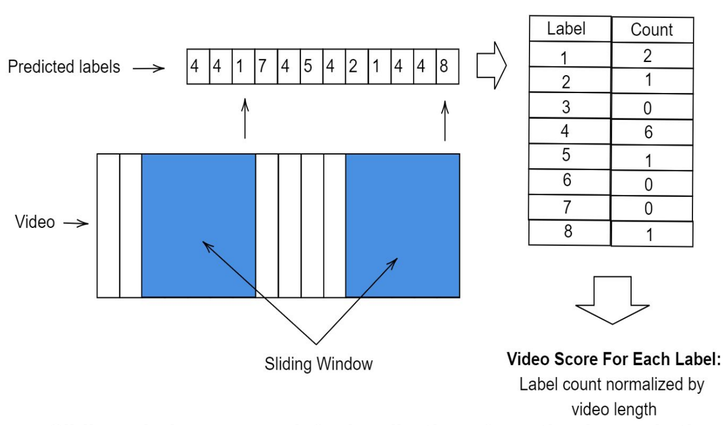
Table 3. Test set accuracy for each emotion (For the best model from table 2)

Localization of emotion jumps in the full length video ads:

We estimated the ability of our video classification module to localize (presence/absence) emotion jumps in the full length video ads. For this purpose, we labeled each video in the test sets (3055 videos) with 1 if it has at least one "true" emotion jump (for the considered emotion) and -1 otherwise.

We implemented a sliding window approach (see figure). Each incoming video is sliced into five seconds clips starting from 0,1,.. seconds so for a 30 second video we would get 25 clips. For each clip using our trained VCM model we predict the clip label. Thus as output we get an integer array of size 25 (for a 30 second video). We refer to this value for a given emotion as video emotion score.

The final binary label (has at least one emotion jump for a given emotion) is assigned to the full video based on the video emotion score by setting a threshold. To evaluate performance for localization of emotion jumps in the full length video ads we use area under the curve (AUC) metrics (table 4).

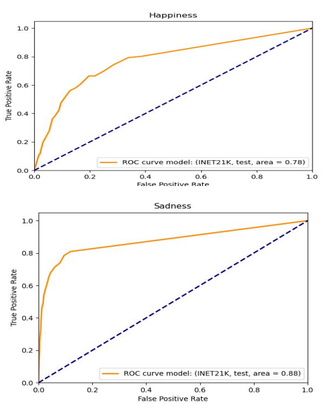


Sliding window approach for localization of emotion jumps in the full length video ads

The ability to localize emotion jumps varied significantly among emotions (table 4). For Sadness the validation and test AUC score was 0.91 and 0.88 respectively. In general, we could see an obvious trend - a good prediction performance for the well defined emotions like Sadness, Happiness and Fear while poor performance for a less clearly defined like Contempt or Neutral.

Emotion	AUC
Anger	0.67
Contempt	0.67
Disgust	0.69
Fear	0.83
Happiness	0.78
Neutral	0.57
Sadness	0.88
Surprise	0.73

Table 4. Area Under the Curve (AUC) for localization of emotion jumps in the full length video ads.



ROC curves for localization of emotion jumps in the full length video ads (see table 4) for Happiness and Sadness

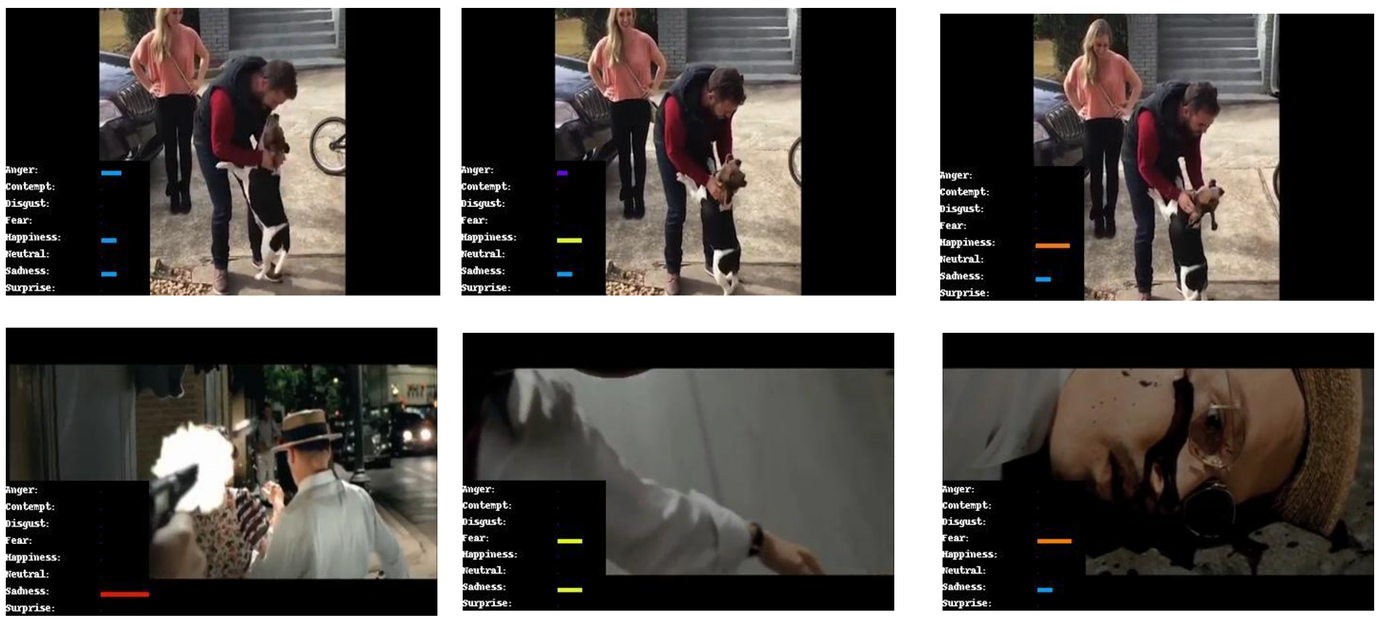
importance of emotion jumps:

Star rating (or its equivalent) of the ads is a commonly accepted measure of ad quality in the marketing industry. The star rating varies in the range between 1 to 5 where 5+ means the best possible quality and 1 vice versus. 5-Star ads leave the audience feeling happy with a good perspective of maximum commercial gain in the future. We demonstrate significant link between presence/absence of emotion jumps in the ad and the star rating it would get. The results are reported in table 5.

Star Rating	Number of videos	Happiness (percentage of videos)
1+	15186	0.9%
2+	9763	3.1%
3+	4369	14.1%
4+	1126	28.3%
5+	307	48.8%

Table 5. Distribution of ads by star rating and percentage of videos with at least one Happiness emotion jump in each star rating category.

External Validation on Youtube videos:
We use Youtube search engine to select top videos based on a searching string. We selected videos based on the search strings ”Animals Reunited With Owners” and ”Heartbreaking Moments in Movies PART”. The former videos are supposed to be enriched with emotion jumps related to Happiness while the latter ones are supposed to be movie clips provoking Sadness or Anger. For example, about 77 percent of 5 seconds intervals in the first category were predicted to evoke Happiness, while 62 percent of 5 seconds intervals in the first category were predicted to evoke Sadness/Fear/Anger.



References:
[1] M. Madhumita, “Emotion recognition: can ai detect human feelings from a face?” The Financial Times. [Online]. Available: https://www.ft.com/content/c0b03d1d-f72f-48a8-b342-b4a92610945
[2] Y. Baveye, C. Chamaret, E. Dellandr´ea, and L. Chen, “Affective video content analysis:A multidisciplinary insight,” IEEE Transactions on Affective Computing, vol. 9, no. 4, pp. 396–409, 2018.
[3] J. Lin, C. Gan, and S. Han, “TSM: Temporal shift module for efficient video understanding,” in Proceedings of the IEEE International Conference on Computer Vision, vol. 2019-October, 2019.