Introduction to a Task on Context of Experience: Recommending Videos Suiting a Watching Situation

Michael Riegler[1], Martha Larson[3], Concetto Spampinato[2], Jonas Markussen[1]
Pål Halvorsen[1], Carsten Griwodz[1]
[1]Simula Research Laboratory, Norway
[2]University of Catania, Italy
[3]Delft University of Technology, Netherlands
{michael, jonassm, paalh, griff}@simula.no, cspampin@dieei.unict.it, m.a.larson@tudelft.nl

ABSTRACT

We propose a Context of Experience task, whose aim it is to explore the suitability of video content for watching in certain situations. Specifically, we look at the situation of watching movies on an airplane. As a viewing context, airplanes are characterized by small screens and distracting viewing conditions. We assume that movies have properties that make them more or less suitable to this context. We are interested in developing systems that are able to reproduce a general judgment of viewers about whether a given movie is a good movie to watch during a flight. We provide a data set including a list of movies and human judgments concerning their suitability for airplanes. The goal of the task is to use movie metadata and audio-visual features extracted from movie trailers in order to automatically reproduce these judgments. A basic classification system demonstrates the feasibility and viability of the task.

1. INTRODUCTION

The challenge of the Context of Experience task is to automatically predict viewers' judgments on whether video content is suitable for a particular watching situation. Ultimately, the aim is to build a recommender system that would provide viewers with recommendations of content for a given context. Currently, the majority of work on video content recommendation focuses on personal preferences, and overlooks cases in which context might have a strong impact on preference relatively independently of the personal tastes of specific viewers. Particularly strong influence of context can be expected in psychologically stressful or physically uncomfortable situations.

For our task, we choose one such situation, with which a large number of people have quite frequent experience: watching movies on an airplane. In this situation, a large majority of viewers share a common goal, which we consider to be a *viewing intent*. The goal is to pass time as pleasantly and meaningfully as possible, while confined in the small space of an airplane cabin, which is characterized by a number of distractors. We take the large number of websites discussing movies to watch on airplanes (e.g., [9]) as evidence that this viewing intent is dominant among air travellers. Although the scope of this task is limited to the airplane scenario, we emphasize that the challenge of Context of Experience is a much broader area of interest. Other examples of stressful contexts where videos are becoming increasingly important include hospital waiting rooms, and dentists offices, where videos are shown during treatment.

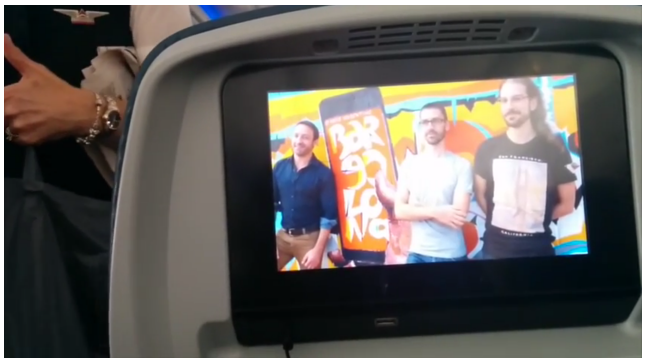

Figure 1: A set of conditions, including small screen and confined, crowded space, characterize the context of watching a movie on an airplane.

2. TASK DESCRIPTION

For the task we provide the participants a list of movies, including links to descriptions and video trailers. The assignment of the task is to classify each movie into +goodonairplane / -goodonairplane classes. Therefore, the ground truth of the task is derived from two sources: A list of movies actually used by a major airline[1], as well as user judgments on movies that are collected via a crowdsourcing tool[2]. Task participants should form their own hypothesis about what is important for users viewing movies on an airplane, and design an approach using appropriate features and a classifier, or decision function. Figure 1 gives an impression of a screen commonly used on an airplane and the very specific attributes regarding size and quality of the video. The value of the task lies in understanding the ability of content-based and metadata-based features to discriminate the kind of movies that people would like to watch on small screens under stressful or somehow not normal situations. Since the multimedia content that users watch on flights can influence their well being and overall experience this task is related to the quality of multimedia experience work like for exam-

[1]http://www.klm.com/travel/no_en/prepare_for_travel/on_board/entertainment/onboard_movies.htm
[2]https://crowdflower.com/

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

ple [6, 5, 3, 4, 1]. Apart from that, the task also includes the area of user intent since the intent of the users, why they want to watch movies on the airplane, is a strong influencing force on what they watch [7, 8]. Task participants are provided with a collection of videos, i.e., trailers as a representative for the movie because of copyright issues, and the context, e.g., video URL, metadata, user votes etc. Apart from that we also provide different pre-extracted features, including visual and audio features. The participants are asked to develop methods that will predict to which intent class the video belongs, respectively, good or bad to watch on an airplane.

To tackle the task it can be addressed by leveraging techniques from multiple multimedia-related disciplines, including such as social computing (intent), machine learning (classification), multimedia content analysis, multimodal fusion, and crowdsourcing. Further we hope that it will be useful for content provider, since the exploitation of intent in combination with users' satisfaction could lead to more sophisticated ways to develop methods of providing a better service to the users.

3. DATA SET

The data set we provide is released, including titles and links, that allow participants to gather online metadata and trailers for movies. We do not, as already mentioned, provide the video files because of copyright restrictions. Movies are collected based on movie lists from a major international airline, in our case, KLM Royal Dutch Airlines. The final list of movies is a merged set of movies collected between February and April 2015. The video data set contains both positive and negative samples, whereas the negative examples are carefully sampled from IMDB in order to create a fair and representative negative class. The data set is split into a training set and a test set. In order to collect user judgments, we use an existing system that has been built for the purpose of collecting user feedback of this sort. We evaluate systems both with respect to the airline's choice of movies, and the crowd's choice of airline-suitable movies. Votes about the labels collected by crowdsourcing are considered as the authoritative labels. For this reason, crowdworkers are asked to rank a small set of movies with respect to how strongly they would like to watch this video on an airplane. This ranking is then combined to create the class for each movie in the training and test data.

Technical details. Overall, the data set contains 318 movies. Links to trailers are collected from IMDB and YouTube. Participants are also allowed to collect their own data such as full length movies, more metadata and user comments, etc. The goal of systems that are developed to address this task should be to automatically identify appropriate content, i.e., whether a movie should be recommended for being watched on an airplane or not. To achieve this goal, the methods should not require manual or crowdsourced input. The data set contains extracted visual, audio and text features. Furthermore, we provide metadata collected from IMDB including user comments. The visual features that are provided are: Histogram of Oriented Gradients (HOG), Color Moments, Local Binary Patterns (LBP) and Gray Level Run Length Matrix. The audio descriptors are Mel-Frequency Cepstral Coefficients (MFCC). The development set contains 95 labelled movies. The test data contains 223 movies without labels.

Features used	Precision	Recall	F1-score
Metadata + user ratings	0.581	0.6	0.583
Only user ratings	0.371	0.609	0.461
Only visual information	0.447	0.476	0.458
Only metadata	0.524	0.516	0.519

Table 1: Classification in terms of weighted average of precision, recall and F1-score.

Evaluation. For the evaluation we use the standard metrics Precision, Recall and weighted F1 score. Negative and positive classes in both data sets are balanced. Participants are asked to submit a predicted class for each movie in the test data set. The metrics then are calculated and provided to the participants. For a transparent and fair procedure, the labels used for the evaluation will be released together with the results.

Initial results. To confirm the viability of the task, and show the possibilities opened by this data set, we carried out some basic classification experiments. For the classification we used the Weka library. As classifier we choose the rule based PART classifier. This classifier uses separate and conquer to generate a decision list. From this, it builds a decision tree where the best leaves are used as rules for the classifier [2]. Table 1 show the results of our four initial runs. For the evaluation we used the weighted average of precision, recall and F1-score. The first run uses metadata (language, year published, genre, country, runtime and age rating) in combination with user ratings as input for the classifier. This run is our best performer. It clearly outperforms the naive baseline, which is 0.5 (precision, recall and F1-score). The second run uses user ratings alone. This run performs well with recall, but poorly with precision. This implies that receiving certain user ratings is a necessary, but not a sufficient condition for being a movie that is good to watch on an airplane. Taken together, the first two runs confirm that the task is non-trivial, and that it is also viable. The third run uses visual features. This run scores below the naive baseline. However, the approach to visual classification here was relatively simple. Additional exploratory experiments, not reported here, revealed that visual features do have the ability to approve results when used in combination with other features. Such combinations are interesting for future work.

Finally, the last run confirms that metadata without user ratings is able to yield performance above the naive baseline. An information gain based analysis of all features ranked genre, publication year, country, language and runtime as the top five features.

4. SUMMARY

The task is challenging due to the complex relationship between the multimedia content, and viewers' perceptions and reception. We hope that the novel use case will inspire researchers to investigation of user intent and context of experience. Understanding user intent and what users need in order to have the best experience is an important emerging topic in the area of multimedia research.

5. ACKNOWLEDGMENT

This work is funded by the Norwegian FRINATEK project "EONS" (#231687) & the EC project CrowdRec (#610594).

6. REFERENCES

[1] A. Borowiak and U. Reiter. Long duration audiovisual content: Impact of content type and impairment appearance on user quality expectations over time. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 200–205. IEEE, 2013.

[2] E. Frank and I. H. Witten. Generating accurate rule sets without global optimization. 1998.

[3] P. Lebreton, A. Raake, M. Barkowsky, and P. Le Callet. Evaluating complex scales through subjective ranking. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 303–308. IEEE, 2014.

[4] B. Rainer and C. Timmerer. A quality of experience model for adaptive media playout. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 177–182. IEEE, 2014.

[5] J. A. Redi, Y. Zhu, H. de Ridder, and I. Heynderickx. How passive image viewers became active multimedia users. In *Visual Signal Quality Assessment*, pages 31–72. Springer, 2015.

[6] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank. Factors influencing quality of experience. In *Quality of Experience*, pages 55–72. Springer, 2014.

[7] M. Riegler, L. Calvet, A. Calvet, P. Halvorsen, and C. Griwodz. Exploitation of producer intent in relation to bandwidth and qoe for online video streaming services. In *Proceedings of the 25th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 7–12. ACM, 2015.

[8] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. 2010.

[9] Tripinsurance. Best movies guide for airplanes. http://www.tripinsurance.com/tips/guide-to-the-best-moviestv-shows-to-watch-on-a-plane. [last visited, Dezember. 10, 2014].