Right Inflight? A Dataset for Exploring the Automatic Prediction of Movies Suitable for a Watching Situation

Michael Riegler[1], Martha Larson[2], Concetto Spampinato[3], Pål Halvorsen[1], Mathias Lux[4]
Jonas Markussen[1], Konstantin Pogorelov[1], Carsten Griwodz[1], Håkon Stensland[1]
[1]Simula Research Laboratory & University of Oslo
[2]TU Delft & Radboud University Nijmegen
[3]University of Catania
[4]University of Klagenfurt
michael@simula.no, m.a.larson@tudelft.nl, cspampin@dieei.unict.it, mlux@itec.aau.at

ABSTRACT

In this paper, we present the dataset *Right Inflight* developed to support the exploration of the match between video content and the situation in which that content is watched. Specifically, we look at videos that are suitable to be watched on an airplane, where the main assumption is that that viewers watch movies with the intent of relaxing themselves and letting time pass quickly, despite the inconvenience and discomfort of flight. The aim of the dataset is to support the development of recommender systems, as well as computer vision and multimedia retrieval algorithms capable of automatically predicting which videos are suitable for inflight consumption. Our ultimate goal is to promote a deeper understanding of how people experience video content, and of how technology can support people in finding or selecting video content that supports them in regulating their internal states in certain situations. *Right Inflight* consists of 318 human-annotated movies, for which we provide links to trailers, a set of pre-computed low-level visual, audio and text features as well as user ratings. The annotation was performed by crowdsourcing workers, who were asked to judge the appropriateness of movies for inflight consumption.

CCS Concepts

•**Information systems → Information retrieval; Multimedia and multimodal retrieval;**

Keywords

Multimedia; Intent; Context; Data Set

1. INTRODUCTION

Increasingly, researchers are interested in developing multimedia analysis techniques that can predict the affective impact of video on viewers, and in releasing datasets that will support this work [26, 2]. Such work focuses on the way that viewers experience video. It fails, however, to take into account that a viewer does not view a video in a vacuum. Rather, viewing a video involves simultaneously experiencing *the video* and also *the context in which the video is viewed.* These two experiences interact, giving rise to a new challenge for multimedia research, which we call *Context of Experience.* Two major considerations underlie the importance of Context of Experience. First, we anticipate that techniques that are able to determine the suitability of videos for particular contexts of experience would be applicable for a wide range of users. For cases in which context has a strong impact on how a viewer experiences video, we expect that context could be an important predictor of viewer preference for videos, overshadowing personal taste or mood. Second, Context of Experience is closely related to user viewing intent, i.e., the reason why a person is watching a particular video. If we are able to predict the suitability of a video for a context, we are able to give viewers more useful tools for finding and selecting content that will help them in a given situation, for example, self-regulating in a situation of psychological tension.

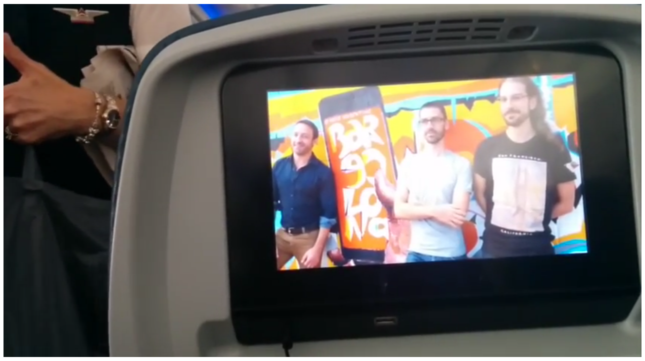


Figure 1: A set of conditions, including small screen and confined, crowded space, characterize the context of watching a movie on an airplane.

In this paper, we focus on the case of viewers watching movies on an airplane. Here, independently of personal preferences, viewers share the common goal, which we consider to be a *viewing intent*, of passing time and keeping themselves occupied by being entertained while being confined in an uncomfortable small space of an airplane cabin. The dataset is designed to help answer the question of whether it

MMSys'16, May 10 - 13, 2016, Klagenfurt, Austria

ISBN 978-1-4503-4297-1/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2910017.2910619

is possible to predict which movies allow viewers to achieve the goal of passing time, relax or distracting themselves, given the context. Furthermore, the context of airline travel includes also limitations of the employed technology (e.g., screen size) and the environment itself (e.g., background noise, interruptions, presence of strangers). We have chosen the airplane scenario as the role of stress and viewers' intent to distract themselves is widely acknowledged [29]. 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.

Although the scope of the proposed dataset is limited to the airplane scenario, 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. The dataset was initially developed, but not public released, as part of the MediaEval Multimedia Benchmark [12, 7] and a preliminary description can be found in [18].

The dataset comprises a list of 318 movies, including links (movies or trailers are not provided because of copyright issues) to descriptions and video trailers, as well as a set of pre-extracted visual, audio and text features from movies, along with annotations created by human judges, in this case, crowdsourcing workers. Additional information about the context such as metadata and user votes/rates is also given.

The dataset is designed to support binary classification of movies as either `+goodonairplane` or the `-goodonairplane` class. For this reason, the ground truth of the task is derived from two sources: A list of movies actually used by a major airline [6], as well as user judgments on movies that are collected via a crowdsourcing platform [4].

In order to address the Context of Experience challenge instantiated by this dataset, researchers can form their own hypothesis to find out what is important for users. This can be done for example by using most appropriate low-level features extracted from airplane's movies, and, accordingly, design approaches using appropriate features, classifiers, recommender system or decision function. The value of the dataset 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. The *Right Inflight* dataset can be addressed with a variety of multimedia-related methods, like for example, recommender systems, social computing (intent), machine learning (classification), multimedia content analysis, multimodal fusion and crowdsourcing.

Further, we hope that the insights that can be gained with this dataset will be useful for content providers. If it is possible to understand how user intent contributes to user satisfaction, it would be possible to provide users with more sophisticated content recommendation and delivery services.

2. RELATED WORK

The challenge of Context of Experience stands at the intersection of research efforts currently ongoing in two different disciplines. First, in the field of multimedia, it is related to work on the impact of video content on viewers. Several datasets and benchmarks have contributed to supporting research that develops algorithms capable of automatically predicting the emotional impact (affective impact) of video content on the viewer. Within the MediaEval benchmark [12], these have been an early task on predicting viewer experienced boredom [26] and a current task on the affective impact of movies [25]. Moreover, in the field of multimedia, extensive work has been carried out on Quality of Experience, including [16, 15, 8, 14, 3]. Finally, Context of Experience is related to multimedia research in the area of viewer intent [17], since the intent of users (i.e., the reason why they want to watch movies on the airplane) is a strong influencing force on what they watch [17].

Second, in the field of recommender systems, Context of Experience is related to work on context-aware recommendation [1, 24]. Researchers have devoted significant effort into organizing challenges in the area of context-aware movie recommendation [22, 23]. There is, however, a critical difference between the challenge of Context of Experience and the challenge of context-aware movie recommendation. Context of Experience assumes that the experience of viewing a movie interacts with the context in which a movie is viewed. As a result, the movie is actually able to *change* the context. By conceptualizing context as Context of Experience we focus on the possibility that viewers might choose to view a movie driven by a particular intent, i.e., a goal. In the context of airline travel, which we assume has a strong interaction with the movie viewing experience, we assume the goal of the viewers is to be more comfortable and past time. Addressing Context of Experience means that we are not 'just' matching movies with personal tastes, but actually helping users accomplish goals. Although, personal preference without doubt plays a key role in determining which movies that people most enjoy during air travel, it is important that recommender systems are also able to exploit the general, context-related, tendency for people to find certain movies more suitable than others for watching on an airplane.

Datasets for research in computer science are an important tool to allow researchers to exchange and compare methods, techniques and algorithms. In information retrieval, large collections of document are used to evaluate for instance new ranking mechanisms or relevance functions. Due to the ever-changing nature of available data, new datasets are necessary. Recently there has been a move to develop datasets that consist of Creative Commons material. This movement helps the community to overcome the challenge of dealing with licensing restrictions, which effectively limit both the collection and the redistribution of data. Some datasets are released with the idea that they will be used for multiple purposes, for example, YFCC100M, a large-scale Flickr image dataset [28]. Whereas other datasets are released with annotations.

A key example is the LIRIS-ACCEDE (Annotated Creative Commons Emotional DatabasE) dataset for affective video content analysis [2], already mentioned in the Introduction.

Across the areas of multimedia and recommender systems, it is notable that few datasets focused on the actual intent of the users and the context. To the best of our knowledge, there is only one dataset including multimedia data (images in this case) as well as the photographers' intent, namely [11]. To create this dataset, photographers on Flicker were asked for permission to include their photo in the dataset as well as to take part in a survey, which aimed at uncovering the actual reason why the people took the photo. Possible answers ranged from to publish it online, to capture a moment, to preserve a feeling. The data then was double checked in an evaluation run on Amazon Mechanical Turk. Both the photo survey as well as the results from Mechanical Turk are part of the dataset.

3. DATA COLLECTION

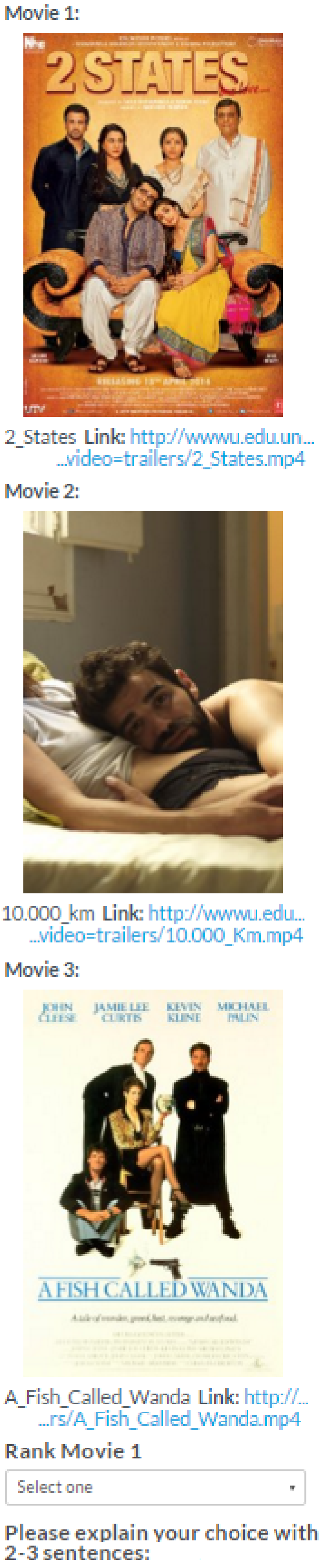
The dataset was collected in a series of steps. First, we collected the names of all the movies that were shown on flights by KLM between February 2015 and April 2015 from the KLM website [6]. We ended up with 201 movies for February, 196 for March and 200 for April. The movies were also ordered into 7 categories by KLM. The categories were *Latest, Recent, The collection, Family, World, Dutch movies* and *European movies.* Some of the movies appeared several times in different months. In the final list of movies, each movie only appeared once. The selection of movies that we included in the dataset contained 318 movies containing videos collected from KLM as positive examples and carefully selected negative examples from movie databases. For negative examples, we chose movies of the same categories and released around the same time of positive samples, but not used in the KLM system.

For the movies in this list, we crawled (i) metadata from popular movie ranking websites like IMDb and Rotten Tomatoes, etc. and (ii) links to movie trailers and posters. Afterwards, we conducted a crowdsourcing study using the Crowdflower [4] platform in two steps. First, we asked the study participants about their flying experience and their experience with movies in order to identify crowdworkers (people who do tasks on crowdsourcing platforms) who had watched movies during a flight. When we collected a large enough subset of flight experienced workers, we performed a second study.

In the second part, we asked the workers to rank the movies of our first collected list in terms of how likely they would watch the movies during a flight. This study is described in more detail in the next section.

4. CROWDSOURCING OF MOVIE PREFERENCES

Since crowdsourcing of subjective information is quite challenging, we followed the principles discussed in [30] and [19]. In our crowdsourcing study, we collected opinions concerning whether people would like to watch a movie on an airplane or not. Each worker was given 3 trailers to watch plus a short video intended to help them recall the situation of being on an airplane [1]. Figure 3 shows the task description presented to the crowdsourcing workers. After they looked at the trailers, we asked some questions. First, we asked them to provide us the title of each movie in order to check whether the crowdworkers actually watched the movies or just rushed trough the questions. After that, we asked them to rank the videos from 1 to 3 according to the likelihood (1 the most likely, 3 the least likely) they would watch those videos during a flight. Crowdworkers were also asked to provide a short explanation/motivation of their ranking as well as their favorite movie genre. For each movie, we collected at least five rankings from different users. From these



Movie 1:

2_States Link: http://wwwu.edu.un...
...video=trailers/2_States.mp4

Movie 2:

10.000_km Link: http://wwwu.edu...
...video=trailers/10.000_Km.mp4

Movie 3:

A_Fish_Called_Wanda Link: http://...
...rs/A_Fish_Called_Wanda.mp4

Rank Movie 1

Select one ▾

Please explain your choice with 2-3 sentences:

Figure 2: Task presentation to the workers. Each worker was given 3 movies to rank and some additional questions to answer.

[1] https://youtu.be/TxC3OV9dBeo

Run	Features used	Precision	Recall	F1-score
i	Metadata + user ratings	0.581	0.6	0.583
ii	Only user ratings	0.371	0.609	0.461
iii	Metadata + Visual	0.584	0.6	0.586
iv	Only visual information	0.447	0.476	0.458
v	Only metadata	0.524	0.516	0.519

Table 1: Classification in terms of weighted average of precision, recall and F1-score for different types of input data used.

rankings, we calculated the average rank that was used to determine the label `+goodonairplane` or `-goodonairplane`. For movies, for which we could not make a clear decision, we collected more crowdsourcing data to break the tie.

All in all, we had 548 different workers participating in the task who provided 1644 judgments. From these 1644 judgments, we used 1590 after discarding workers who provided answers that clearly reflected that they did not take the task seriously. To detect such non-serious workers, we checked over the provided movie titles and questionnaire completion times. A very fast finishing time was defined as faster as the average of three people in our laboratory could read and finish the task if they tried to do it very fast. Which was circa 3 minutes plus the time of the trailers. We discarded around 20% of all submitted tasks using this method. The participants came from a lot of different countries (varying from USA to India). Around 53% where from Europe with Spain having the highest share with almost 5%. Circa 19% of the workers where from Asia, 14% from India and 14% from USA. Figure 2 shows the final design of the task as presented to the workers.

5. DATASET DESCRIPTION

The dataset release includes 318 movie titles and links to gather online metadata and trailers for movies. We do not provide the video files because of copyright restrictions. The trailers were downloaded from IMDb [27] and YouTube [31]. Furthermore, we provide metadata collected from IMDb, Rotten Tomatoes [21] and Metacritic [13] including user comments.

The dataset includes also low-level visual, audio and text features extracted from trailers, posters, metadata and user comments. The provided visual features are Histogram of Oriented Gradients (HOG) gray, Color Moments, local binary patterns (LBP) and Gray Level Run Length Matrix [10]. The audio descriptors are Mel-Frequency Cepstral Coefficients (MFCC) [9]. For text information, we provide a term frequency–inverse document frequency ($td - idf$) matrix, which gives indications about the importance of different words [20].

The dataset enables evaluation of systems both with respect to the airline's choice of movies and the crowd's choice of flight-suitable movies. Votes about the labels collected by crowdsourcing are considered as the authoritative labels.
The development set contains 95 labeled movies. The test data contains 223 movies (the split is chosen based on what we think would provide a robust evaluation of algorithms tests with the dataset). Negative and positive classes in both splits of the dataset are balanced. The majority class baseline is 0.5 for precision. For the evaluation, we recommend standard metrics such as weighted average of precision, recall and weighted F1 score.

6. APPLICATION OF THE DATASET

To show the usefulness of the dataset, we conducted some initial experiments. The findings of these experiments are presented here. To confirm the viability of the dataset for supporting identification of movies suitable to be watched on an airplane, and show the possibilities that it opens we carried out some basic classification experiments. For these experiments, we used the WEKA machine learning library[2]. As a 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 from which the best leaves are used as rules for the classifier [5]. Table 1 shows the results of our four initial experiments.

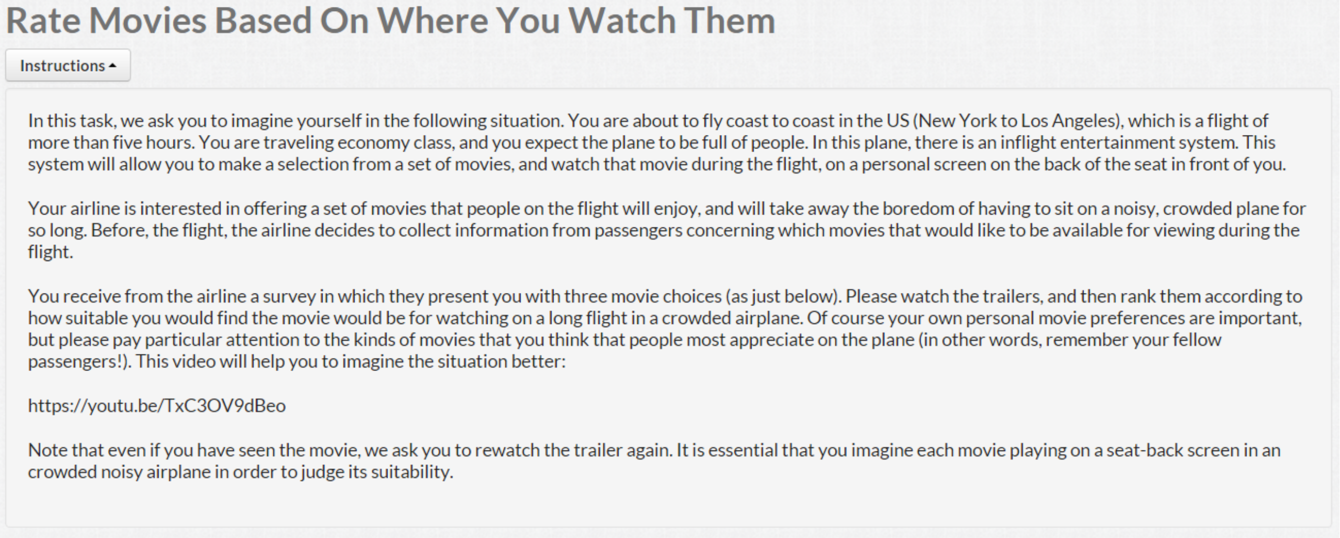
The first experiment (i) uses metadata (language, year published, genre, country, runtime and age rating) in combination with user ratings as input for the classifier. This run is our best performer. It clearly outperforms the naive baseline, which is 0.5 (precision, recall and F1-score).

The second run (ii), uses user ratings only (collected from IMDb, Rotten Tomatoes and Metacritic). 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. Which is a very important message because it means, that using user ratings from standard platforms only does not lead to the best recommendations. This is a strong indicator that the watching situation is an important impact factor. Taken together, the first two runs confirm that the task is non-trivial, and that it is also viable.

The only user ratings experiment (ii) achieves virtually the same performance as Metadata + Visual run (iii). The results are not, however, exactly identical. We take this as motivation to perform further experiments in the future (different features, audio features, etc.).

The only visual information run (iv) uses global visual features for the classification. This run scores below the naive baseline. However, the approach to visual classification here was relatively simple. We only used one global image feature, namely Joint Composite Descriptor (JCD). JCD is a combination of Fuzzy Color and Texture Histogram (FCTH) and Color and Edge Directivity Descriptor (CEDD) [32]. It combines color, textural and edge information in one descriptor. This makes it a good choice for initial tests since the most promising parts are included. 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.

[2] http://www.cs.waikato.ac.nz/ml/weka/



Rate Movies Based On Where You Watch Them

In this task, we ask you to imagine yourself in the following situation. You are about to fly coast to coast in the US (New York to Los Angeles), which is a flight of more than five hours. You are traveling economy class, and you expect the plane to be full of people. In this plane, there is an inflight entertainment system. This system will allow you to make a selection from a set of movies, and watch that movie during the flight, on a personal screen on the back of the seat in front of you.

Your airline is interested in offering a set of movies that people on the flight will enjoy, and will take away the boredom of having to sit on a noisy, crowded plane for so long. Before, the flight, the airline decides to collect information from passengers concerning which movies that would like to be available for viewing during the flight.

You receive from the airline a survey in which they present you with three movie choices (as just below). Please watch the trailers, and then rank them according to how suitable you would find the movie would be for watching on a long flight in a crowded airplane. Of course your own personal movie preferences are important, but please pay particular attention to the kinds of movies that you think that people most appreciate on the plane (in other words, remember your fellow passengers!). This video will help you to imagine the situation better:

https://youtu.be/TxC3OV9dBeo

Note that even if you have seen the movie, we ask you to rewatch the trailer again. It is essential that you imagine each movie playing on a seat-back screen in an crowded noisy airplane in order to judge its suitability.

Figure 3: Crowdsourcing task description. It also includes a link to a video that should help the workers to get in the feeling of a flight situation.

Finally, the last experiment (v), using only metadata, 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.

7. LIMITATIONS OF THE DATASET

The collected data and the idea behind it is very novel and opens some promising directions in the field of multimedia. Nevertheless, it also comes with some limitations.

The crowd-sourcing study is carefully prepared with enough means to check for the subjects' reliability. However, the data for each movie are collected from five subjects only which can be seen as on the lower end considering the subjectivity and difficulty of the task. Moreover, this makes it hard for a statistical analysis which should be performed on any data collected from subjects.

Furthermore, the methodology of splitting the dataset into suitable and not suitable based on the ranks is questionable. To tackle this problem, all crowdworkers votes and rankings are included in the dataset. That should allow possible users a more detailed insight. Even though the task is well described for the observer, and the initial video to *place the subject into the situation* is well prepared, it is very hard to be sure that subjects fully understood the task and can picture themselves in the situation.

The data is also collected based on the trailers only, while the ranks from the databases are for the whole movies which can lead to some biases. A further limitation is that the data is only collected from one airline (KLM) so far. Although, investigating different airlines revealed that the used movies over the used time were almost identical.

This lead us to the conclusion that airlines most probably follow recommendations based on the popular rating sites. Taking all these limitation into consideration, we still believe that the obtained ground-truth data can give a first signal and open a new direction but any conclusions should therefore be drawn with taking them into account.

8. CONCLUSION AND OUTLOOK

We have presented *Right Inflight*, a dataset that allows researchers to explore the next challenge of predicting whether video content is suitable for a particular watching context. We choose to focus on airline travel, since the relative familiarity of the situation, and the relatively extremeness of the distractors, allow us to more easily tap into general opinions of people about the content suited for the context. The resulting dataset poses a challenge for multimodal classification that is extremely difficult. However, contrary to what one might expect, given the subjective nature of individuals' preferences for movies, inferring which movies are considered suitable for watching on an airplane is not impossible.

Our ambition is that the novel use case addressed by the dataset may inspire multimedia researchers to delve deeper into research questions that involve user viewing intent and the context of multimedia experience. As mentioned in the introduction, we believe that Context of Experience is important in helping people to decide which kinds of content is suitable for stressful situations including waiting rooms, airports, and during medical treatments, such as dental procedures. We hope that our dataset can help to raise awareness about the topic, but also provide an interesting and meaningful use case to researchers already working in related fields.

9. ACKNOWLEDGEMENTS

This work is partly funded by the FRINATEK project ”EONS” (#231687) and the BIA project *PCIe* (#235530) funded by the Norwegian Research Council and by the EC FP7 project CrowdRec (#610594).

10. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

[2] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content

analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

[3] 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.

[4] Crowdflower. Crowdflower crowdsourcing platform. http://crowdflower.com. [last visited, March. 10, 2016].

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

[6] KLM-Royal-Dutch-Airlines. KLM on board entertainment system content. http://www.klm.com/. [last visited, March. 10, 2016].

[7] M. Larson, B. Ionescu, M. Sjöberg, X. Anguera, J. Poignant, M. Riegler, M. Eskevich, C. Hauff, R. Sutcliffe, G. J.F. Jones, Y.-H. Yang, M. Soleymani, and S. Papadopoulos. Proceedings of the MediaEval 2015 multimedia benchmark workshop. *The MediaEval 2015 Workshop*, 2015.

[8] 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.

[9] B. Logan et al. Mel frequency cepstral coefficients for music modeling. In *Proc. of ISMIR*, 2000.

[10] M. Lux. Lire: Open source image retrieval in java. In *Proc. of MM*. ACM, 2013.

[11] M. Lux, D. Xhura, and A. Kopper. User intentions in digital photo production: A test data set. In *MultiMedia Modeling*, pages 172–182. Springer, 2014.

[12] MediaEval Benchmarking Initiative for Multimedia Evaluation. MediaEval benchmark homepage. http://www.multimediaeval.org/. [last visited, March. 10, 2016].

[13] Metacritic. Metacritics critics and ratings. http://www.metacritic.com/. [last visited, March. 10, 2016].

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] M. Riegler, M. Larson, C. Spampinato, J. Markussen, P. Halvorsen, and C. Griwodz. Introduction to a task on context of experience: Recommending videos suiting a watching situation. In *Proceedings of the MediaEval 2015 Workshop*. CEUR-WS.org, 2015.

[19] M. Riegler, V. Reddy G, M. Larson, P. Halvorsen, and C. Griwodz. Crowdsourcing as self fulfilling prophecy: Influence of discarding workers in subjective assessment tasks. In *Proc. of CBMI*, 2016.

[20] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 2004.

[21] Rotten-Tomatoes. Rotten tomatoes critics and ratings. http://www.rottentomatoes.com/. [last visited, March. 10, 2016].

[22] A. Said, S. Berkovsky, and E. W. De Luca. Putting things in context: Challenge on context-aware movie recommendation. In *Proceedings of the Workshop on Context-Aware Movie Recommendation*, CAMRa '10, pages 2–6, New York, NY, USA, 2010. ACM.

[23] A. Said, S. Berkovsky, and E. W. De Luca. Group recommendation in context. In *Proceedings of the 2Nd Challenge on Context-Aware Movie Recommendation*, CAMRa '11, pages 2–4, New York, NY, USA, 2011. ACM.

[24] Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, 47(1):3:1–3:45, May 2014.

[25] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The MediaEval 2015 Affective Impact of Movies Task. In *Proceedings of the MediaEval 2015 Workshop, CEUR-WS.org*, 2015.

[26] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic. Corpus development for affective video indexing. *IEEE Transactions on Multimedia*, 16(4):1075–1089, 2014.

[27] The-Internet-Movie-Database-IMDB. Imdbs critics and ratings. http://www.imdb.com/. [last visited, March. 10, 2016].

[28] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[29] 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, 2015].

[30] B. Winther, M. Riegler, L. Calvet, C. Griwodz, and P. Halvorsen. Why design matters: Crowdsourcing of complex tasks. In *Proc. of CrowdMM*. ACM, 2015.

[31] Youtube. Youtube movie sharing platform. http://www.youtube.com/. [last visited, March. 10, 2016].

[32] K. Zagoris, S. A. Chatzichristofis, N. Papamarkos, and Y. S. Boutalis. Automatic image annotation and retrieval using the joint composite descriptor. In *Informatics (PCI), 2010 14th Panhellenic Conference on*, pages 143–147. IEEE, 2010.