Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 155-158 recibido 29-03-2019 revisado 07-05-2019 aceptado 17-05-2019

Project CAVIAR
CApturing VIewers' Affective Response

Inferencia de la respuesta afectiva de los espectadores de un vídeo

Fernando Fernández-Martínez[1], Zoraida Callejas[2], Ricardo Kleinlein[1], Cristina Luna Jiménez[1], Juan Manuel Montero[1], José Manuel Pardo[1]
[1]Universidad Politécnica de Madrid, Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Madrid, Spain
[2]Universidad de Granada, Department of Languages and Computer Systems, Granada, Spain
fernando.fernandezm@upm.es; Tel.: +34-915495700 (ext. 4228)

Abstract: In this project we propose the automatic analysis of the relation between the audiovisual characteristics of a multimedia production and the impact caused in its audience. With this aim, potential synergies are explored between different areas of knowledge including, among others: audiovisual communication, computer vision, multimodal systems, biometric sensors, social network analysis, opinion mining, and affective computing. Our efforts are oriented towards combining these technologies to introduce novel computational models that could predict the reactions of spectators to multimedia elements across different media and moments. On the one hand, we study the cognitive and emotional response of the spectators while they are watching the media instances, using neuroscience techniques and biometric sensors. On the other hand, we also study the reaction shown by the audience on social networks by relying on the automatic collection and analysis of different metadata related to the media elements, such as popularity, sharing patterns, ratings and commentaries.
Keywords: Opinion mining, aesthetics, multimedia information retrieval, metadata, video recommendation, indexation, biometrics, emotion, subjective response prediction

Resumen: Este proyecto propone el análisis de la posible dependencia entre el contenido audiovisual de una producción multimedia y el impacto causado por ésta en sus espectadores. Para ello, nos apoyamos en diferentes áreas de conocimiento tales como comunicación audiovisual, visión por computador, sistemas multimodales, sensores biométricos, análisis de redes sociales, análisis de opinión o computación afectiva, entre otras, con el objetivo de diseñar nuevos modelos computacionales que permitan predecir las reacciones de los espectadores de un vídeo de forma transversal a los medios y momentos en que éstas se producen. Trabajamos principalmente con dos tipos de respuesta: la respuesta cognitiva y emocional inmediata de los espectadores durante el visionado, que medimos utilizando técnicas de neurociencia y sensores biométricos, y la reacción expresada en redes sociales, cuyo impacto es cuantificado mediante el análisis automático de diferentes metadatos recabados para dichos vídeos, tales como popularidad, patrones de compartición, valoraciones y comentarios realizados en las redes.
Palabras clave: Análisis de opinión, estética, recuperación de la información multimedia, metadatos, recomendación e indexación de vídeo, biometría, emoción, predicción de respuesta subjetiva

ISSN 1135-5948. DOI 10.26342/2019-63-19 © 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

In a world where new technologies are progressively more related with multimedia information, it is essential to have tools that simplify the treatment of this type of data. In this sense, a problem that has recently attracted the interest of the scientific community is the development of models for objectivizing what is subjective, concretely, to measure the quality and impact of audiovisual productions (Luo, Wang, and Tang, 2011).

Project CAVIAR proposes the automatic analysis of the relation between the audiovisual characteristics of a multimedia production and the impact caused in its audience. With this aim, we have gathered together different knowledge areas including: audiovisual communication, computer vision, multimodal systems, biometric sensors, social network analysis, and affective computing. Our contribution will be based in combining these technologies to study thoroughly and predict the reactions of spectators to multimedia elements across different media and moments. We will work mainly with two response types: the immediate reaction while viewing the audiovisual, and the shown reaction on social networks.

On the one hand, we will study the cognitive and emotional response of the spectators while they are watching the audiovisual, using neuroscience techniques and biometric measures based on sensors. This will allow transitioning from the typical scenario only based on opinion surveys to a more precise model that will also allow measuring reactions to specific passages (not only the general reaction).

On the other hand, the natural reaction will also be studied focusing on the audience's behaviour on social networks by means of automatic analysis of the metadata corresponding to the multimedia elements watched. We will establish techniques to evaluate the impact on the users through the study of the popularity, sharing patterns, ratings and commentaries received, as a way of quantifying the effect of the audiovisual and its implications.

Regarding the automatic inference of the spectator's perception, the audiovisual content may provoke cognitive effects (e.g. attract attention) and affective responses (e.g. arouse a particular emotional state) in the audience, which are the typical objectives of every multimedia production. However, the factors that determine if they have a positive or negative effect (or no effect at all) on the audience's response are still unknown to a great extent, and their corresponding analysis is far from automatic. To achieve this, we will identify the relevant elements of the audiovisual production in two channels: video and sound, and will study their correspondence with the perception of the spectators in the lines described earlier, generating an automatic prediction system.

2 Project objectives

The main purposes of this project are:

- To present a novel approach based on different metadata (such as related opinions and comments) extracted from social networks (such as YouTube) for the automatic annotation of videos in terms of their impact or their expected or potential perceived value.
- To explore neuroscience techniques as an alternative and totally different automatic annotation solution by measuring biometric indicators of the cognitive and affective response of the audience to the video content stimuli.
- To expand upon existing research to investigate how audiovisual content can influence perception and emotions and use related findings for the construction of effective computational models.
- To develop technology for the automatic assessment of the effectiveness of a video and to provide novel automatic tools for multimedia document retrieval, focusing on the automatic analysis of the audiovisual content of a video as a mean to derive suitable principles for predicting the audience's perception.

3 Methodology and viability

3.1 Impact on social networks

CAVIAR researchers have already tested the feasibility of this proposal by means of a pilot study (Fernández-Martínez, Hernández-García, and de María, 2015) which has been conducted by choosing car commercials and related metadata downloaded from YouTube

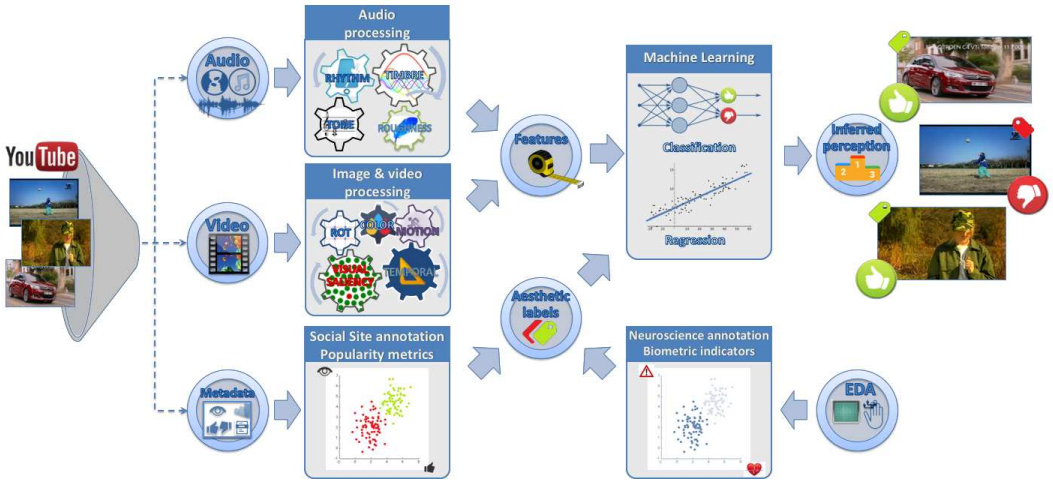

Figure 1: Overview of the proposed approach

as the evaluation domain for some of the suggested approaches (figure 1 shows the overall pipeline of the proposal).

To that end, a novel approach to automatically annotate a corpus composed of 138 videos is presented. Videos are grouped into 2 classes corresponding to different satisfaction levels, by means of a regular k-means algorithm applied to the video metadata related to users feedback. Evaluation results show that simple linear logistic regression models based on the 10 best visual descriptors and on the 10 best audio descriptors individually perform reasonably well, achieving a classification accuracy of roughly 70% and 75%, respectively. Combination of audio and visual descriptors (Hernández-García, Fernández-Martínez, and Díaz-de María, 2016) yields better performance, roughly 86% for the top-20 selected from the entire descriptor set, but tipping the balance in favor of the audio ones (i.e. 17 vs 3). Audio content bigger influence in this domain is also evidenced by a side analysis of the video comments (Fernández-Martínez et al., 2014).

3.2 Biometric response

Electrodermal activity, EDA, is a psychophysiological indicator which can be considered a somatic marker of the emotional and attentional reaction of subjects towards stimuli. EDA measurements are not biased by the cognitive process of giving an opinion or a score to characterize the subjective perception, and group-level EDA recordings integrate the reaction of the whole audience, thus reducing the signal noise.

Thanks to our collaboration agreement with Sociograph company (`www.sociograph.es`), a new corpus of audiovisual contents has already been annotated from recorded physiological signals (EDA) measured on 270 participants while they watched a concatenation of videos. These videos correspond to 136 short films selected from the Jameson Notodofilmfest Short Film Festival 2015.

In (Hernández-García, Fernández-Martínez, and Díaz-de María, 2017), we predict the levels of emotion and attention, derived from EDA records, by means of a small set of low-level visual descriptors computed from the video stimuli. Linear regression experiments show that our descriptors predict significantly well the sum of emotion and attention levels, reaching a coefficient of determination $R^2 = 0.25$. In (García-Faura et al., 2019) we extend the previous work by labeling short video clips according to the audience's emotion (high vs. low) and attention (increasing vs. decreasing), derived from EDA records. Here, we propose a set of low-level audiovisual descriptors and train binary classifiers that predict the emotion and attention with 75% and 80% accuracy, respectively.

Both works reinforce the usefulness of such low-level audiovisual descriptors to model video in terms of the induced affective response and set a promising path for further research on the prediction of emotion and attention from videos using EDA. In addition to the release of new datasets and annotation methods, the team plans to also take advantage of other data sets (Baveye et al., 2015; Demarty et al., 2016) from the literature, so that the advances obtained in CAVIAR can be compared with the state of the art and developments of other researchers.

4 Potential applications

From an applicability point of view, proposed solutions could pave the way for a new generation of recommendation, indexation and summarization systems that could change the way consumers interact with multimedia search engines by allowing them to actively use enhanced search choices inspired on audiovisual, social and/or biometric features that help the retrieved content to be more accurate and related to the affective response and more personalized. In addition, they could also help training inference models that could be far more representative and indicative of the viewers' real feelings and attitudes towards videos.

Anticipating the subjective value perceived by the viewers of any audiovisual content could also enable a more efficient multimedia content production. For instance, it could be a low-cost and reliable alternative to costly conventional test screening processes which are traditionally based on artificial setups that tend to bias the actual feelings or opinions of the participants.

Finally, the use of the proposed technology could also boost advertising by allowing us to anticipate the impact of a marketing campaign on the potential customers before it is released to the media, thus paving the way for the end of trial and error in advertising and saving up many costs. Aside, it could also help pointing out what audiovisual aspects would need to be changed for the campaign to be successful and effective.

5 Technology transfer

As a result of our recent efforts and advances, our automatic aesthetic assessment technology is currently being applied in the "ESITUR-UPM: Interactive Tourist Showcase" project, led by MOVILOK INTERACTIVIDAD MÓVIL S.L., and funded by the Spanish Ministry of Economy and Competitiveness under the "Retos de Colaboración 2016" call, which is oriented to the design and development of useful and efficient solutions for 'Smart tourism' to improve the tourism experience of its users.

Acknowledgments

The work leading to these results has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through the ESITUR (MINECO, RTC-2016-5305-7), CAVIAR (MINECO, TEC2017-84593-C2-1-R), and AMIC (MINECO, TIN2017-85854-C4-4-R) projects (AEI/FEDER, UE).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X Pascal GPU used in this research.

References

Baveye, Y., E. Dellandrea, C. Chamaret, and L. Chen. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*.

Demarty, C.-H., M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefèbvre. 2016. Predicting media interestingness task. In *MediaEval 2016*.

Fernández-Martínez, F., A. H. García, A. Gallardo-Antolín, and F. D. de María. 2014. Combining audio-visual features for viewers' perception classification of youtube car commercials. In *2nd International Workshop on Speech, Language, and Audio in Multimedia*, SLAM'14.

Fernández-Martínez, F., A. Hernández-García, and F. D. de María. 2015. Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials. *Expert Systems with Applications*.

García-Faura, A., A. Hernández-García, F. Fernández-Martínez, F. Díaz-de María, and R. San-Segundo. 2019. Emotion and attention: Audiovisual models for group-level skin response recognition in short movies. *Web Intelligence*, 17.

Hernández-García, A., F. Fernández-Martínez, and F. Díaz-de María. 2016. Comparing visual descriptors and automatic rating strategies for video aesthetics prediction. *Image Communications*.

Hernández-García, A., F. Fernández-Martínez, and F. Díaz-de María. 2017. Emotion and attention: Predicting electrodermal activity through video visual descriptors. In *International Conference on Web Intelligence*, WI '17.

Luo, W., X. Wang, and X. Tang. 2011. Content-based photo quality assessment. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11.