Automatic Emotion Recognition for Cultural Heritage

Claudio Baecchi, Andrea Ferracani, Alberto Del Bimbo
MICC, University of Florence, Viale Morgagni 65, Florence, IT
E-mail: [name.surname]@unifi.it

Abstract. In this work we present an automatic emotion recognition system for the re-use of multimedia content and storytelling for cultural heritage. A huge amount of heterogeneous multimedia data on cultural heritage is available in online and offline databases that can be used and adapted to produce new content. In the real world, human video editors may want to select the video sequences composing the final video with the intention to induce an emotional reaction in the viewer (e.g. happiness, excitement, sadness). Usually they try to achieve this result following their personal judgement. However, this task of video selection could benefit a lot from the exploitation of an automatic sentiment classification system. Our system can help the editor in choosing the video sequences that best fit the desired emotion to be induced. First-of-all the system splits the video in scenes. Then it classifies them using a multimodal classifier which combines temporal features extracted form LSTM, sentiment-related features obtained through a DNN, audio features and motion-related features. The system learns which features are more important and exploits them to classify the scenes in terms of valence and arousal which are well known to correlate with induced emotions. Finally it provides an online video composer which allows the editor to search, filter and compose the scenes in a new video using sentiment information. To train the classifier we also collected and annotated a small dataset of both users recorded videos and professional ones downloaded from the web.

1. Introduction

Nowadays, the huge diffusion of audiovisual content has made it possible to communicate information of multiple nature to a very large audience. The primary purpose of this diffusion is the vision and appreciation of the audiovisual content itself. Just think of content such as advertisements, films, or even more recent user-produced content for social media users. The primary objective is the ‘satisfaction’ of the user who views these contents, since a satisfied user is more likely to buy the sponsored product, continue to watch the film or ‘follow’ who produces the content.

Therefore, it is important for content creators to know how end-users can be ‘stimulated’ to achieve their purpose. In creating a film, for example, it is essential, in addition to the plot, to give the right rhythm and color to the scenes. The choice of the soundtrack is equally essential to correctly convey the sensation you want to convey. Semiotics, which is the science that studies signs and how they are meaningful to some purpose, can be applied to audiovisual content to induce and analyse specific emotional stimuli.

Computer vision systems have exploited sentiment analysis on the basis of results from studies of semiotics [2]. In this context, factors such as the rapidity of changing scenes, the colors and the...
movement of the scene have been analyzed. With the advent of convolutional neural networks, only the color values of the pixels of the scene have been used to extrapolate a possible induced feeling [3].

Our system, which we call the CultMEDIA system (from Culture and Media), provides the possibility of classifying videos, or parts of them, with the feeling that these induce on those who watch them. Given the fundamental importance of content reuse this can make an important contribution. The ability to automatically annotate videos or their parts with information relating to the sentiment induced on users provides important help to human video editors to create new content. The creative process of generating new videos is in fact strongly influenced by the need to provoke emotions. Giving to the creator a tool capable of classifying the videos that he reuses can actually help him in creating videos with a specific emotional target.

2. Video Sentiment Classification

The current methods of video sentiment analysis use only the visual content of the video frames to classify the induced sentiment. What is called sentiment in the literature corresponds to the common term “emotion”, which is defined as a complex response to a relevant stimulus, characterized by certain subjective experiences and a specific physiological reaction.

Emotions are intense and short responses. There are two keywords used in the literature to describe emotions: 1) valuation, which indicates whether an emotion is positive or negative, and 2) arousal, which instead indicates the degree of activation and refers to the intensity of the physiological response of an emotion. Valence and arousal can be negative, neutral or positive. Studies show [4] that the combination of the pairs of these two parameters can be mapped to the basic emotions of the human beings, such as happiness, anger, etc., as shown in Fig. 2. Video sentiment analysis can therefore be provided classifying these two parameters in the combination of their three basic states.

On the other end by its nature sentiment is something that arises and ends quickly. This has to be considered when trying to provide video sentiment analysis. While watching, for example, a movie, the sentiment of the viewer can change several times. Frequently content creators want to create more stimuli, giving the film an alternation or a continuity of emotions that contribute to leaving the viewer more or less satisfied at the end of the vision. Given this, automatic sentiment analysis need to focus on small parts of the video, which by definition have a continuity of sentiment. These parts usually coincide with the scenes.

3. Machine Learning for Automatic Sentiment Classification

The automatic learning of a classifier for video sentiment analysis is carried out through supervised learning of a convolutional neural network. This network receives as input one or more images (i.e. video frames) that are part of the training set, usually a subset of the videos that are annotated, and outputs the probabilities of belonging to each of the three classes of valence and arousal. Exploiting an appropriate loss, the network learns its weights in order to associate the highest probability class to the input image as the output.

To train the network we collected and annotated a small dataset consisting of 157 videos recorded by users from 3 Italian museums in Modena, Firenze and Napoli, plus 30 professional videos downloaded from the famous social media platform YouTube. We first segmented the videos into scenes and annotated each scene with valence and arousal values.

The neural network is not trained from scratch on our dataset. Instead, we perform a fine-tuning of the weights, obtained by training on a different set of data. This is done mainly because the set of videos annotated is not large enough to train satisfactorily, or to give the network the desired generalization properties. So, the network is “pre-trained” on a similar dataset by domain, that is the DeepSentibank dataset. Once convergence is achieved, a new training is
Figure 1. Describing emotions using valence and arousal.

performed on the final dataset using the weights obtained by the training on DeepSentibank as starting point.

4. Multimodal Classification
In addition to visual characteristics, there are many other video features that influence a spectator’s emotional state. The soundtrack, for example, is very important in conveying emotions. The same video edited with different soundtracks can induce completely opposite emotions. Another fundamental aspect is the temporal succession of events. In fact, not only it is important what the content of the images in the video is, but also how they follow each other. In the very short period it can be how fast a scene is, whilst, for the long period, the rhythm of the scenes that form the sequence. For these reasons, it is not enough to only perform an analysis of the visual component, and classification of video sentiment must also take into account audio and temporality.

In literature, audio is analyzed using MFCC (Mel-Frequency Cepstral Coefficients) features extracted from a temporal interval of around one second. The temporal aspect is instead addressed with the use of recurrent neural networks, or RNN [6], such as LSTM [7]. These networks keep an internal state (or memory) which varies according to the sequence of inputs they receive.

By jointly exploiting these three aspects, we trained a classifier that uses all the three typologies of information. The visual component however remains a fundamental part for the sentiment classification. For this reason, there are datasets specially created for the sentiment analysis task. The most famous one, and the one that we used, is DeepSentibank [8]. It is a collection of more than one million images from the social media Flickr, labeled using 2089 Adjective-Noun Pairs (or ANP), such as “happy smile”, or “crying baby”. The neural network trained on DeepSentibank learns to classify the images on the 2089 ANP, searching in the images the concepts expressed by the adjectives that represent emotions.

Therefore, this network can be used as a feature extractor for new images in order to classify sentiments in videos. However, in the classification of sentiment, components such as the structure of the scene and the objects present in it must also be considered. For this reason, in addition to the sentiment features, we combine also features related to the appearance of the
scene itself. For this purpose we exploit the public ImageNET dataset [9]. ImageNet is a dataset made up of over a million images, organized into 1000 classes of everyday objects or animals, such as desks, or cat.

5. Exploiting Attention

In the context of the CultMEDIA project the videos that are going to be used and managed are mostly recorded in museums. In these videos not the whole images (frames) contain information useful for sentiment classification. A video that shows artworks in a museum has its own interesting part in the artworks themselves, while the background does not bring useful information but is rather noise that can degrade the classification accuracy. Considering this, the classifier should focus only on the “important” parts, discarding the useless ones.

For this purpose, an attention mechanism has been developed, see Fig. 2, which allows the network to focus only on the relevant parts. The attention mechanism makes use of an Objectness Detector [1] to identify the salient part of the input image. The coordinates of the related bounding box containing the salient part are then used to feed the neural network. In this way, noisy features are discarded and are taken into account only those with greater information contribution.

6. System Architecture

System architecture is shown in Fig. 3 which summarizes its main modules. For each video shot we first extract the frames. The frames are then grouped into sets of 24 frames to extrapolate the temporal information. Each frame group is processed to extract the features related to audio, sentiment and the scene. For the audio, the relevant MFCCs are extracted from the soundtrack corresponding to the 24 frames under analysis. For the sentiment, the central frame is used as the reference frame for the group and processed using a CNN previously trained on DeepSentibank. For the visual part of the scene, the 24 frames are processed, processing them first through the attention mechanism to detect the salient area, the same frames are then processed by a CNN trained on ImageNET to extract 24 related features. The information relating to the salient zone is used to obtain the most significant features.

MFCC and scene features are further processed by a recurring LSTM neural network to extrapolate the temporal component as well. The relative outputs, together with the sentiment
features, are concatenated and used as inputs for a linear classifier, which has the task of classifying them in the three basic classes of valence and arousal.

The system is divided into two main modules: 1) the feature extraction module, where the MFCC, sentiment and scene features are generated, and 2) the classification module, where the features are first processed by the LSTM networks and then by the classifier. The first module is used as a feature generator and its weights are never changed. The second module is in a recurrent neural network that must be trained in order to correctly classify the sentiment.

7. Experiments
We evaluated our Video Sentiment Analysis classifier on our Dataset, introduced in Sec. 3. We tested the system with and without our attention mechanism module (see Sec. 5).

We splitted the Dataset into training and validation set, respectively of 80% and 20% of the videos.

In Table 1 we show our results in terms of accuracy (valence and arousal). It can be noted that when including the attention mechanism in the system both valence and arousal accuracy improves by 15% and 11% respectively.

8. REST API for Video Sentiment Classification
The video sentiment classifier has been embedded in a web framework for remote use by other systems. The framework exposes REST-type APIs that allow to request the classification of one

| Method          | Valence | Arousal |
|-----------------|---------|---------|
| Without attention | 42%     | 48%     |
| With attention  | 57%     | 59%     |

Table 1. Accuracy for valence and arousal classification.
or more video shots, obtaining their estimated valence and arousal values, and to monitor the processing status of the service.

The REST API allows to submit a video and all the scenes (shots) in order to have them emotionally classified according to the valence and arousal paradigm. The REST service creates a queue of tasks that run synchronously. The results and the status of the processing are verifiable via specific calls to the REST API. Messages are encoded using the JSON format.

9. Video Composer
To demonstrate the potentiality of the Video Sentiment Analysis service we developed a web interface for video editing and compositing. The application allows video editors to search for museum and cultural heritage video sequences, automatically and/or manually annotated with information on sentiment, and to compose them in a new video (see Fig. 4). In particular, the application allows:

- to upload a video;
- user registration;
- multi-user annotation of uploaded video shots using valence/arousal parameters. This information serves as ground truth for the Video Sentiment Analysis system;
- to search and filter video shots by concept and on the basis of sentiment (arousal/valence), see Fig. 5;
- the composition of a new video through the interactive arrangement of shots in a timeline;
- the choice of a transition effect between the scenes;
- adding an audio file to the video;
- recording audio from the user microphone and associating it with the video;
- to add subtitles and to synchronize the text to the video sequence, see Fig. 6;
- to export the composition as an Adobe Premiere project, allowing more advanced editing capabilities;
- the generation and download of a new video directly from the web platform.

The Video Composer interface is organized in three main horizontal sections: 1) the header; 2) the video carousel; 3) the interactive timeline for video composition. In the header there are tools which allow users (e.g. video editors) to search video scenes matching with particular keywords or combination of keywords. Furthermore they can filter the results using sentiments (i.e. arousal/valence). The video carousel shows the result of the search, i.e. the video scenes found for the user query. The video editor can grab video scenes in the carousel and drop them in the video timeline in order to combine them in the final video. For each scene in the timeline the user can add a title, an audio file, he can register an audio or add and synchronize captions. He can also choose a transition effect between scenes. The composition can be downloaded as it is or can be exported as an Adobe Premiere project for more advanced editing.

10. Conclusions
In this article we presented a system for sentiment annotation and classification of video scenes. The classifier is multi-modal and takes into account visual features as well as temporal, audio and motion-related characteristics of video streams (i.e. our attention mechanism). The scenes are classified in terms of valence and arousal, a two dimensional space often used in literature to characterize the human emotional spectrum. The classifier feeds a web-based video composer which enables video editors to ‘pre-filter’ video scenes on the basis of the results of video sentiment annotation. In this way they can easily find video scenes that can be more relevant.
Figure 4. Video Composer: the search and composition interface. The user can drag the sequences of interest from the upper list to the lower timeline, arranging as desired the timeline of the final video.

Figure 5. Video Composer: video shots can be searched by concept and sentiment through specific emotional filters, both content and stylistic, in particular by using the valence / arousal paradigm.

for the mood they intend to convey in their final product. As future research we are going to improve the way we represent ‘mood’ adding to the videos in the system annotations on visual features and video stylistic characteristics correlated with sentiment (e.g. colorfulness, presence of happy faces, dialogue, indoor vs outdoor, night vs day). A usability study of the web interface will be also conducted.

10.1. Acknowledgments
This system has been developed in the context of the project CultMEDIA “Machine learning-based services for harvesting multimedia documents to support low-cost video post-production and cross-media storytelling” financed by Italian Ministry of Education, University and Research, and in the context of the project MORPHCAST (UNIFI_FSE2017) “Real time video creation according to your emotions” financed by Regione Toscana (POR FSE 2014-2020) Giovanisi.
Figure 6. Video Composer: adding and synchronizing subtitles for a video shot. The contextual buttons below each shot in the timeline respectively allow 1) to add a title; 2) to add and synchronize the subtitles; 3) to add an audio file and / or an audio recording through the computer microphone.

References
[1] Galteri L, Seidenari L, Bertini M and Del Bimbo A 2017 Spatio-Temporal Closed-Loop Object Detection Transaction on Image Processing 26 1253-1263.
[2] Colombo C, Del Bimbo A and Pala P 1999 Semantics in visual information retrieval IEEE Multimedia 6 38-53.
[3] Baecchi C, Uricchio T, Bertini M and Del Bimbo A 2017 Deep sentiment features of context and faces for affective video analysis Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval 72-77.
[4] Thayer R E 1989 The biopsychology of mood and arousal Oxford University Press.
[5] Sulistijono I A, Urrosyda R C and Darojah Z 2016 Mel-Frequency Cepstral Coefficient MFCC for Music Feature Extraction for the Dancing Robot Movement Decision in Proceedings, Part II, of the 9th International Conference on Intelligent Robotics and Applications 9835 283-294.
[6] Elman J L 1990 Finding structure in time Cognitive science 14, 179–211.
[7] Hochreiter S and Schmidhuber J 1997 Long Short-Term Memory Neural Computation 9 1735-1780.
[8] Borth D, Ji R, Chen T, Breuel T and Chang S-H 2013 Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs ACM Multimedia Conference.
[9] Deng J, Dong W, Socher R, Li L-J, Li K and Fei-Fei L 2009 ImageNet: A Large-Scale Hierarchical Image Database IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 248-255.