Evoked Emotion Distribution Learning through Analysis of Temporal User Comments in Social Media Videos

Yiming WANG†, Marc A. KASTNER††, Da HUO†, Takahiro KOMAMIZU†††,†, Takatsugu HIRAYAMA††††,†, and Ichiro IDE†

† Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601 Japan
†† Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
††† Mathematical and Data Science Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601 Japan
†††† Department of Environment Science, University of Human Environments
Motojuku-cho, Okazaki, Aichi 444-3505, Japan
E-mail: † {wangy, huod}@cs.is.i.nagoya-u.ac.jp, ide@i.nagoya-u.ac.jp
†† mkastner@i.kyoto-u.ac.jp ††† taka-coma@acm.org †††† t-hirayama@uhe.ac.jp

Abstract The field of *a ective video content analysis*, which aims to estimate viewers' emotions evoked from a input video, is growing as the amount of online video content increases. However, annotating videos with emotions is challenging due to the subjective and ambiguous nature of emotions. This research introduces the Label Distribution Learning (LDL) paradigm to limit the impact of subjectivity by modeling the label of evoked emotions as a distribution rather than a single dominant emotion. In addition, an approach to automatically annotate the viewers' emotion distribution based on user-generated comments instead of annotating them manually is proposed. A video dataset with emotion distribution annotations is composed using this method. An Evoked Emotion Distribution Learning (EEDL) model is adopted to estimate the emotion distribution evoked from social media videos. Experiments using the proposed EEDL model on the composed dataset show promising prospect for using LDL in this task.

Key words A ective video content analysis, Label Distribution Learning, social media

1. Introduction

Recently, videos on the Internet have become one of the most commonly viewed media. This makes the search, recommendation, and analysis of different kinds of videos in social media increasingly difficult. For example, for users, the abundance of videos can make it hard for them to find videos relevant and interesting to them. They may also feel overwhelmed by the sheer amount of content available [7]. Therefore, research on understanding a viewer's experience or the viewer's perception towards different content is needed, in order to provide some insights that can be applied to address the aforementioned challenges. By analyzing the emotions that viewers experience when watching a video, systems are more likely to recommend content that aligns with users' interests and preferences.

However, measuring the evoked emotion is a complex task as emotion is subjective and often ambiguous. This also leads to a problem that annotating datasets related to human emotions is quite time-consuming and cost-intensive because more annotators are needed to reach agreements on the annotation. Therefore, in this report, the Label Distribution Learning (LDL) paradigm [4] is introduced for the annotation of video frames. User-generated comments are utilized as a prior to automatically annotate viewers' evoked emotion distribution to videos.

Specifically, the LDL paradigm allows models to predict the distribution over different classes rather than a single dominant class label explicitly to limit the subjectivity of human emotions. In order to solve the problem that manual labeling of datasets is infeasible, this report analyzes Danmaku, a special type of comments with timestamps, as a

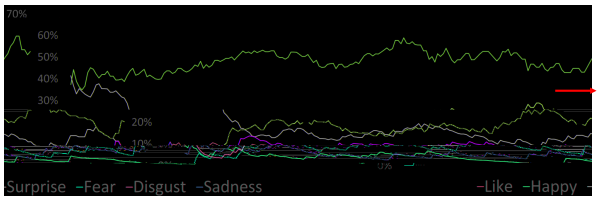

Figure 2 Example of the annotation result.

tation, this work considers using Danmaku, a special kind of comments which contain temporal information, to automatically annotate social media videos to compose a video dataset on LDL evoked emotion as shown in Fig. 1.

The first step is to find videos with a sufficient number of Danmaku available and obtain all Danmaku commented on these videos. Then as the second step, several Danmaku emotion classifiers are applied to detect whether corresponding emotions appear in a Danmaku. As the third step, to automatically annotate the evoked emotions to a video, a Danmaku pool, which contains several Danmaku attached to the video, is created for each frame. Since the emotion classifiers have estimated the emotions contained for all the Danmaku in the second step, the proportion of each emotion contained in the Danmaku pool can be calculated. At last, the frame is annotated with the distribution of these proportions as LDL annotation.

In the following discussions, $P_c(t)$ will be used to represent the label, which is the probability that a viewer will experience the specific emotion c in Parrot's emotion model, where c can take on any value from the set $\{0, 1, 2, 3, 4, 5\}$ corresponding to *like*, *happy*, *surprise*, *fear*, *disgust*, and *sadness*, respectively, while watching the video at timestamp t. Accordingly, it can be defined as:

$$P_c(t) = N_c(t) \Big/ \sum_{c=0}^{5} N_c(t), \tag{1}$$

where $N_c(t)$ is the count of emotion c in the Danmaku pool corresponding to the frame at timestamp t.

Fig. 2 shows the distribution of labels $P_c(t)$ for a video. LDL models will be trained with these labels that contain the distribution of different kinds of evoked emotions to social media videos.

Video Collection. For the proposed method, it is important to identify social media videos with a sufficient number of Danmaku. The videos in the composed dataset are collected from the Chinese video-sharing Website Bilibili(*2), where Danmaku is one of the most iconic features of the service. The number of videos is 1,400 and the average duration is 220 sec.

Danmaku Emotion Estimation. To automatically annotate the emotion distribution for videos using Danmaku, the key aspect of the proposed method is to estimate the emotion for each Danmaku. Since there are over 17 million Danmaku corresponding to the collected videos, it is infeasible to annotate them manually. Therefore, Natural Language Processing (NLP)-based emotion classifiers are applied to estimate emotion expressed by Danmaku automatically. Firstly, a Danmaku emotion estimation corpus is constructed manually. As the data is dominantly Chinese, a pretrained Chinese BERT-wwm [3] is employed and fine-tuned on this corpus. At last, six emotion classifiers corresponding to six emotions in Parrot's emotion model are obtained, which are used to estimate the emotion expressed by all the Danmaku attached to the collected videos.

3.3 Emotion Distribution Generation by Analyzing Danmaku

In the automatic annotation process, a Danmaku pool, which contains several Danmaku, is created for each frame to calculate the proportions of each emotion category in it, represented as $P_c(t)$ for the frame calculated by Eq. 1. To obtain the label distributions calculated for a video over time, it is necessary to decide how the count of emotion $N_c(t)$ is represented, that is, how to create a Danmaku pool for each frame. One intuitive approach is to simply create a Danmaku pool that contains all Danmaku in the following timestamp. This method will be referred to as *Timestamp-Matching Method (TMM)* hereafter, as a baseline. Here, $n_c(t)$ is named to represent the number of times emotion c appears in the following one timestamp t. So if the TMM is applied, $N_c(t)$ in Eq. 1 and $n_c(t)$ are identical and set as:

$$N_c(t) = n_c(t). \tag{2}$$

3.3.1 Sliding Window Method

The proposed *Sliding Window Method (SWM)* aims to take into account not only the Danmaku in the following timestamp, but also those from timestamps before and after when creating a Danmaku pool. Specifically, for each frame, a certain length of clip is cut before and after the current timestamp in a sliding window manner. In this case, $N_c(t)$ will be described as:

$$N_c(t) = \sum_{i=t-L}^{t+L} n_c(i), \tag{3}$$

where $\pm L$ sec. is the duration of the sliding window.

Although the SWM makes use of the context information in the video, the difference in timestamps between Danmaku and video frames is also a concern, as a large difference can reduce the relevance of the Danmaku to the video frame and increase noise. Therefore, after creating a Danmaku pool using the SWM, the idea is to give different weights to each