Overview of MediaEval 2012 Genre Tagging Task

Sebastian Schmiedeke
Technische Universität Berlin,
Germany
schmiedeke@nue.tu-berlin.de

Christoph Kofler
Delft University of Technology,
The Netherlands
c.kofler@tudelft.nl

Isabelle Ferrané
University of Toulouse, France
Isabelle.Ferrane@irit.fr

ABSTRACT

The MediaEval 2012 Genre Tagging Task is a follow-up task of the MediaEval 2011 Genre Tagging Task and the MediaEval 2010 Wild Wild Web Tagging Task to test and evaluate retrieval techniques for video content as it occurs on the Internet, i.e., for semi-professional user generated content that is associated with annotations existing on the Social Web. The task uses the MediaEval 2012 Tagging Task (ME12TT) dataset which is based on the whole blip10,000 collection, in contrast to the MediaEval 2010 Wild Wild Web (ME10WWW) set used in previous tasks. In this task overview paper, we describe the principal characteristics of the dataset, the task itself, and the evaluation metrics used to test the particpants' results.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]

1. THE ME12TT DATASET

The MediaEval 2012 Tagging Task dataset (ME12TT) contains video episodes from `blip.tv`. These videos were collected for shows for which the link to one of their episodes has been mentioned in Twitter messages of users tweeting about them. Topsy[1] was used to collect links to `blip.tv` videos from tweets. The discovered videos were checked that they were licensed under Creative Commons, downloaded from `blip.tv`, and converted into the container format `ogg` that is unrestricted by software patents using Theora as video codec and Vorbis as audio codec, respectively. The ME12TT dataset is based on the blip10,000 dataset [3]. In previous tagging tasks a subset of this set was used as ME10WWW dataset [4].

The dataset contains 14,838 episodes comprising a total of ca. 3,288 hours of data. These episodes were separated in a development and test set, containing 5,288 videos (having a runtime of 1,143 hours) and 9,550 videos (having a runtime of 2,145 hours), respectively. The proportion of videos in the development and test set is ca. 1:3. Compared to the original dataset, the separation between development and test set is more balanced, enabling the direct application of both retrieval and classification approaches to address the task. These episodes were taken from 2,349 different shows. It was ensured that genres are most equally distributed among both sets.

Each video is associated with metadata (e.g., title, description, tags, ID of uploader), social network information (i.e., Twitter messages), automatic speech recognition transcripts (ASR transcripts), and shot information including key frames. Each video is associated with only one genre label. The following sections describe these different parts of the dataset in more detail.

1.1 Videos and Keyframes

Each video is associated with exactly one of 26 genre labels[2]. These genre labels were determined by querying the `blip.tv` web API[3]. The genre label of each video is represented by the field `categoryName` in the `JSON` output provided by the API. Subsequently, the genre labels were normalized by replacing whitespaces with underscores ('_') and ampersands ('&') by the word 'and'. Some videos are associated with rare genre categories (i.e., `Friends` and `Science`), which were merged to the default category.

For each episode, the shot boundaries were provided by TU Berlin [1]. For each shot segment, a keyframe is extracted from the middle of the shot. In total, this dataset includes approximately 420,000 shots/keyframse concluding an average shot length of about 30 seconds.

1.2 Metadata

The metadata for each video (stored in UTF-8-formatted XML files) include information about the *title, description, uploader id, license, duration, and tags* that were assigned by the uploaders of videos. We performed a normalization to the tags: they are formatted to have no special characters or whitespaces. We only kept tags that occur 10 or more times in the whole dataset.

1.3 Speech Transcripts

Audio was extracted from all videos using a combination of the `ffmpeg` and `sox` software (sample rate = 16,000 Hz, number of channels = 1). The automatic speech recognition

[1] `http://topsy.com`

[2] `Art`, `Autos and Vehicles`, `Business`, `Citizen Journalism`, `Comedy`, `Conferences and Other Events`, `Documentary`, `Educational`, `Food and Drink`, `Gaming`, `Health`, `Literature`, `Movies and Television`, `Music and Entertainment`, `Personal or Auto-biographical`, `Politics`, `Religion`, `School and Education`, `Sports`, `Technology`, `The Environment`, `The Mainstream Media`, `Travel`, `Videoblogging`, `Web Development and Sites`, `Default Category`

[3] `http://wiki.blip.tv/`

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

transcripts were generously provided by LIMSI/Vocapia[4] and LIUM Research team (LST)[5]. Videos are predominantly in English, but several of them are in French, Spanish or Dutch. Depending on the source, (LIMSI/Vocapia or LIUM), the transcripts are accompanied by sets of complementary information or scores.

LIUM [5]: based on the CMU Sphinx project, the system was developed to participate to the evaluation campaign of the international Workshop on Spoken Language Translation 2011. LIUM provided an English transcription for each audio file 'successfully' processed, that is 5,084 from the development set and 6,879 from the test set. These results consist of: (i) one-best hypotheses under NIST CTM format, (ii) word-lattices under SLF (HTK) format, following a 4-gram topology, and (iii) confusion networks, under a ATT FSM-like format.

LIMSI/Vocapia [2]: an XML file has been provided for each audio file processed (5,237 files for the development set and 7,215 for the test set, respectively). Trancripts were produced for all the above languages, according to the following strategy: the language identification system automatically identified the language spoken in the whole video along with a language confidence score (`lconf`). Each file with a language identification score equal or greater than 0.8 was transcribed with the detected language. The remaining files were transcribed twice, with the detected language as well as with the English system. The average word confidence scores (`tconf`) were compared. The transcription with the higher score was chosen. There were files with other identified language for which there was no transcription system. In such cases, no transcripts were provided. For the remaining files no speech was detected.

1.4 Social Data

The social data was gathered from Twitter using the `topsy` social search engine. It was created to have comments about the `blip.tv` episodes and to have annotations and further information from the Social Web. `Topsy` was searched for all Twitter users who mentioned particular episode in their tweets: 8,856 unique Twitter users (i.e., authors presenting the '0th social level') mentioned the videos in contained in the dataset in their tweets. Based on these Twitter users we then used the Twitter API for crawling seed user profiles using a white-listed IP address. Each seed user's profile includes the list of his 'friends' (persons whom he is following), his followers (persons who are following him), and his interlocutors ('@'). Up to 3200 latest posts were crawled per seed user. The '1st social level' is constituted by each author's contacts and the '2nd social level' by these contacts' own contacts, as the same process was repeated on each author' contacts.

2. TASK

Genre categories or genre tags can support users to more easily discover the desired multimedia content on the Internet they are searching (or browsing) for. Much multimedia content—especially (semi-professional) user generated content—is however not accurately or adequately tagged. The MediaEval 2012 Tagging Task attempts to automatically generate genre labels for Internet videos such as they are used on Internet video platforms such as blip.tv.

The task requires participants to automatically assign genre tags to videos using features derived from speech, audio, visual content or associated textual or social information contained in the provided dataset.

Since we assume that particular information associated with a video already contains explicit information w.r.t. genre information, participants are required to submit up to five *runs* in total to represent their different approaches to the task using different experimental conditions, i.e., different sources of data used in order to predict genre labels for videos. These different experimental conditions include the sole usage of: (i) audio and/or visual information (including information about shots and keyframes), (ii) ASR transcripts, (iii) all data except metadata, (iv) all data except of user uploader ID of videos, and (v) all data.

3. EVALUATION

The ground truth is provided in terms of the genre label which was associated to a video by its uploader. The participants' results will be evaluated in terms of mean average precision (mAP):

$$\text{mAP} = \frac{\sum_{q=1}^{Q} \text{AP(q)}}{Q},$$

where Q stands for the queries for genres. $AP(q)$ is the average precision of the q-th genre, which is also reported.

4. ACKNOWLEDGMENTS

The work presented in this paper was partially supported by SEALINCMedia[6], Quaero[7] and projects funded by the European Commission under contracts FP7-216444 Petamedia, FP7-261743 VideoSense. The organizers would also like to thank Martha Larson for organizing this event.

5. REFERENCES

[1] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. In *10th Workshop on Image Analysis for Multimedia Interactive Services, 2009.*

[2] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 4–15. Springer Berlin Heidelberg, 2008.

[3] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, and G. Jones. The Community and the Crowd: Developing large-scale data collections for multimedia benchmarking. *Multimedia, IEEE*, 2012.

[4] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. Jones. Automatic tagging and geotagging in video collections and communities. In *International Conference on Multimedia Retrieval.* ACM, 2011.

[5] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estèv. Lium's systems for the iwslt 2011 speech translation tasks. In *International Workshop on Spoken Language Translation*, 2011.

[4] `http://www.vocapia.com/news/2011_07_15.html`

[5] `http://www-lium.univ-lemans.fr/en/content/language-and-speech-technology-lst`

[6] `http://www.commit-nl.nl/projects/socially-enriched-access-to-linked-cultural-media`

[7] `http://www.quaero.org/`