Leveraging Cognitive Computing for Multi-class Classification of E-learning Videos

Danilo Dessì(✉), Gianni Fenu, Mirko Marras, and Diego Reforgiato Recupero

Department of Mathematics and Computer Science, University of Cagliari,
Via Ospedale 72, 09124 Cagliari, Italy
{danilo_dessi,fenu,mirko.marras,diego.reforgiato}@unica.it

Abstract. Multi-class classification aims at assigning each sample to one category chosen among a set of different options. In this paper, we present our work for the development of a novel system for multi-class classification of e-learning videos based on the covered educational subjects. The audio transcripts and the text depicted into visual frames are extracted and analyzed by Cognitive Computing tools, going over the traditional term-based similarity approaches. Preliminary experiments demonstrate effectiveness and capabilities of the system, suggesting that semantic analysis improves the performance of multi-class classification.

Keywords: Cognitive computing · Multi-class classification · E-learning video classification · Semantic classification

1 Introduction

Digital videos have become one of the most important e-learning formats. The growing popularity of online course providers, such as Coursera[1] and edX[2], has enabled learners to experience smart video-based lectures which are rapidly increasing in number. They mainly provide knowledge through the teacher's voice and the content is usually depicted by presentation slides or digital whiteboards. This has led to specific approaches for the analysis of their educational content.

The maturity of Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) services has made possible the extraction of text from audio and visual frames. As a result, several studies have tried to address content-based video lecture analysis as text analysis for various purposes (e.g. clustering, classification, retrieval). For instance, [1] applied topic modeling to cluster videos from their audio transcripts and [2] extracted key-phrases and topic-based segments that effectively summarize the content of a video lecture. In [3], ASR and OCR results were subsequently analyzed to detect keywords based on Term Frequency Inverse Document Frequency (TF-IDF) scores. Similar approaches were

[1] https://www.coursera.org/.

[2] http://edx.org/.

E. Blomqvist et al. (Eds.): ESWC 2017 Satellite Events, LNCS 10577, pp. 21–25, 2017.
https://doi.org/10.1007/978-3-319-70407-4_5

integrated in [4,5] for video lecture retrieval. However, they tend to adopt traditional term-based similarity approaches. In contrast, knowledge extraction from natural language text can detect insights out of the video data. State-of-the-art cognitive systems, such as IBM Watson[3] and Microsoft Cognitive Services[4], have the ability to infer semantic information rather than simple word frequencies and they can enable systems to better learn about resources.

In this paper, we introduce a supervised multi-class classification system for e-learning videos which uses semantic content together with textual data extracted from audio transcripts and text depicted in visual frames. The goal is to assign each sample to one category chosen from a predefined list according to the covered educational subjects. The text derived from videos is processed to extract semantic content pertaining to concepts. It is the first attempt of mixing text features and semantics for performing multi-class classification of video lectures following the methodology and the tools stated below. For this purpose, we developed a prototype and performed a first evaluation, showing that enriching textual data with semantic content improves classification performances.

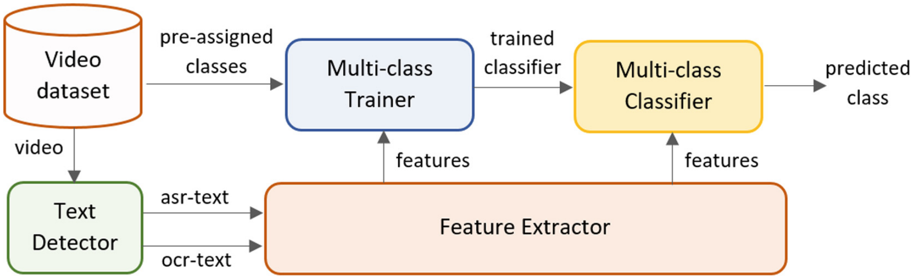

Fig. 1. A reference schema for the proposed system.

2 System Overview

The proposed system is depicted in Fig. 1. It is built on top of three main foundations: (i) the use of textual data derived from videos, (ii) its content-based semantic analysis through cognitive computing services, and (iii) the high modularity for easy customization in terms of feature types and classification algorithms. The modules work in a pipelined work-flow. At this stage, we have implemented a Python prototype following the design described as follows.

Text Detector. The module takes a *video* as input and returns two texts namely *asr-texts* and *ocr-texts* extracted from its audio and its visual frames respectively. It gets the video transcript from ASR computation and the text depicted in the images from OCR. Both raw texts are spell-checked and eventually corrected.

[3] https://www.ibm.com/watson/.
[4] https://www.microsoft.com/cognitive-services.

The module employs IBM Speech-to-Text API[5] and Google Text Recognition API[6] for text detection, WordNet[7] for spell-checking.

Feature Extractor. The module receives two texts namely *asr-texts* and *ocr-texts* as input and returns a set of *features*. Each feature is a pair whose first element is the string identifier of that feature and the second element is its relevance score. The relevance value spans in the range [0, 1] where a value closer to 0 represents a low relevance and a value closer to 1 represents a high relevance of the corresponding feature into the text. The module extracts concepts through IBM Alchemy Language APIs[8] in addition to TF-IDF scores. As default, it returns a set of features resulting from their concatenation where the TF-IDF scores are first row-by-row normalized in the range [0, 1] through a min-max technique. However, the returned type of features can be selected as a parameter.

Multi-class Trainer. The module takes a set of *features* together with the *pre-assigned class* for each video in a training set. The features are employed to represent each video as a vector in a N-dimensional space, where N is the number of different features detected from training videos. Using these vectors, the module trains a classifier and returns it. The module can be set to use a subset of videos for validation. The algorithm underling the classifier can be selected from a list of alternatives we have implemented. At this stage, some variants of support vector machines have been integrated.

Multi-class Classifier. The module takes the *features* associated to a no-labeled video together with a *trained classifier* and returns the predicted class from the set of possible classes derived during training. The module works on the same N-dimensional space used for training the classifier; therefore, new unseen features extracted from the no-labeled video are ignored.

The system is designed to be modular and extensible. Each module is independent from the other ones and the addition and the update of feature types, feature fusion methods, or classification algorithms involve almost no changes to the base architecture. Moreover, each module is properly parametrized, with in mind that the system could be used via a graphical user interface in the future.

3 Preliminary Evaluation

We preliminarily evaluated precision, recall and f-measure of the system using support vector machine as classification algorithm and concepts and TF-IDF as feature types (10-fold cross-validation). The system was tested on a Coursera video dataset which is composed by more than 10,000 pre-annotated videos. For each video, the associated class consists of the category assigned to the course in which the video is provided. Due to unequal category distribution, the

[5] https://www.ibm.com/watson/developercloud/speech-to-text.html.

[6] https://developers.google.com/vision/text-overview.

[7] https://wordnet.princeton.edu/.

[8] https://www.ibm.com/watson/developercloud/alchemy-language.html.

metrics are locally calculated for each category, then their average is obtained by weighting each category metric with the number of instances of the category in the dataset. In Table 1, the preliminary evaluation we conducted shows that the combination of TF-IDF and concepts obtains the highest F-measure.

Table 1. System performance using weighted average computation of metrics.

Features	Precision	Recall	F-Measure
TF-IDF	0.6852	0.6817	0.6741
Concepts	0.6320	0.6205	0.6138
TF-IDF + Concepts	0.6984	0.6951	0.6873

4 Conclusion and Future Work

In this paper, we described our work on developing a system for assigning content-based categories to educational videos from a pre-defined taxonomy based on audio transcripts and text in visual frames. Preliminary results suggest semantic analysis can improve the performance over using textual data only.

In next steps, we would investigate new approaches for assigning relevant scores depending on additional features (e.g. text fonts size), the use of other semantic analysis tools (e.g. frame semantic) and classification algorithms (e.g. neural networks), and larger datasets where to test our system. Moreover, we plan to employ Big Data architectures to support large-scale fast computations. Our system can be applied to other domains where the extraction of content-based categories from videos is essential.

Acknowledgments. Danilo Dessì and Mirko Marras gratefully acknowledge Sardinia Regional Government for the financial support of their PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014–2020 - Axis III Education and Training, Thematic Goal 10, Priority of Investment 10ii, Specific Goal 10.5).

References

1. Basu, S., Yu, Y., Zimmermann, R.: Fuzzy clustering of lecture videos based on topic modeling. In: 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–6. IEEE (2016)
2. Balasubramanian, V., Doraisamy, S.G., Kanakarajan, N.K.: A multimodal approach for extracting content descriptive metadata from lecture videos. J. Intell. Inf. Syst. **46**(1), 121–145 (2016)
3. Yang, H., Meinel, C.: Content based lecture video retrieval using speech and video text information. IEEE Trans. Learn. Technol. **7**(2), 142–154 (2014)

4. Kothawade, A.Y., Patil, D.R.: Retrieving instructional video content from speech and text information. In: Satapathy, S.C., Bhatt, Y.C., Joshi, A., Mishra, D.K. (eds.) Proceedings of the International Congress on Information and Communication Technology. AISC, vol. 439, pp. 311–322. Springer, Singapore (2016). doi:10.1007/978-981-10-0755-2_33
5. Radha, N.: Video retrieval using speech and text in video. In: International Conference on Inventive Computation Technologies, vol. 2, pp. 1–6. IEEE (2016)