Video Categorisation Mimicking Text Mining

Cristian Ortega-León[1], Pedro A. Marín-Reyes[1], Javier Lorenzo-Navarro[1], Modesto Castrillón-Santana[1], and Elena Sánchez-Nielsen[2]

[1] Instituto Universitario SIANI, Universidad de las Palmas de Gran Canaria, Las Palmas de G.C., Spain pedro.marin102@alu.ulpgc.es
[2] Departamento de Ingeniera Informática y de Sistemas, Universidad de la Laguna, Santa Cruz de Tenerife, Spain

Abstract. With the rapid growth of online videos on the Web, there is an increasing research interest in automatic categorisation of videos. It is essential for multimedia tasks in order to facilitate indexing, search and retrieval of available video files on the Web. In this paper, we propose a different technique for the video categorisation problem using only visual information. Entity labels extracted from each frame using a deep learning network, mimic words giving rise to manage the video classification task as a text mining problem. Experimental evaluation on two widely used datasets confirms that the proposing approach fits perfectly to video classification problems. Our approach achieves 64.30% in terms of Mean Average Precision (mAP) in CCV dataset, above other approaches that make use of both visual and audio information.

Keywords: Video Classification · Text Classification · Text Mining · Semantic Video Tagging.

1 Introduction

Online video is responsible for more than 58% of Internet traffic, with an upward trend, as suggested in a recent report [5]. The main reasons for this trend are the growth and consolidation of social networks and streaming platforms, such as YouTube, Vimeo, Viddler or Netflix. For these platforms, due to the huge amounts of uploaded videos, the problem of classifying them into genres/categories is an essential issue. At the high structure level, film or video sets are categorised into different subcategories, such as fiction, horror, thriller, etc.

Currently, most of the digital content uploaded to web pages comprises meta-information which is assigned manually by the user when s/he shares the video. However, this approach is time-consuming for users; being also affected by the different users' understanding on video categories. Therefore, a reviewing process is needed to verify video labels. In this context, automatic video categorisation is a key approach to solve these limitations. Intelligent systems can in fact help users to tag videos, suggesting the most accurate labels. Thus, resulting categories will be more coherently assigned than allowing users to choose manually the categories for the uploaded videos.

Commonly, video categorisation is developed mainly by using models based on meta-textual, visual Region of Interest (ROI) features, audio or their combination, as presented by [23, 31, 24, 25, 28]. Nowadays, authors keep on using the same cues, but making use of Deep Neural Networks (DNNs) as [9, 32]. However, any focus based on low level features is not robust enough for online video categorisation because videos from the same category are frequently diverse in term of features. Based on this limitation, this work aims to propose a methodology that allows automatic video categorisation making use of text mining based strategy to extract features. The adopted approach is evaluated using different models over various public datasets.

The paper is organised as follows. Section 2 presents the related work. Section 3 describes the proposed approach. The experimental design is illustrated in Section 4. Section 5 presents the obtained results of the experiments. Finally, Section 6 concludes with a summary.

2 Related work

2.1 Non Deep Learning based approaches

Video categorisation has received the attention of the computer vision community since years. In [3] an extended survey about the literature of video classification is presented where the authors highlighted three types of modalities: text, audio and visual features. The authors explored an extensive number of approaches using single modality and combinations of them. Another work [4], proposed a systematic study of automatic categorisation of consumer videos, dividing them into a set of classes with diverse semantic concepts, which have been carefully selected based on user studies. In this sense, they manually annotated over approximately 1,300 recordings from real users. Their goals are summarised in: (i) evaluation of the state of the art in multimedia analytics, (ii) evaluation of different approaches in consumer video classification; and (iii) to discover new research opportunities. The work summarised in [13] described a generic video classification algorithm that detects object of interest. They used online user-submitted recordings and aimed to categorise videos into six broad categories. Recently, the approach described in [33] proposed an improved K-means algorithm to categorise video fragments.

2.2 Deep learning based approaches

An extensive empirical evaluation of Convolutional Neural Networks (CNNs) on large-scale video classification is described in [12], using a dataset of one million YouTube videos belonging to 487 categories. Another work [32] proposed and evaluated different Deep Neural Network (DNN) models to combine image information across a video over longer time periods. The authors proposed two methods capable of handling full length videos. One of the proposed methods explored various convolutional temporal feature pooling architectures. The second proposed method explicitly modeled the video as an ordered sequence of

frames, applying Long Short Term Memory (LSTM) units which are connected to the output of the underlying CNN. Moreover, [22] proposed to leverage high-level semantic features to improve the state of the art of temporal model in video categorisation. A LSTM network was used with the aim to understand what is learned by the network. Firstly, object features were extracted from a CNN model that was trained to recognise $20K$ objects. These features feeded the LSTM to capture the video temporal dynamics. The work described in [29] presented a novel approach to combine multiple layers and modalities of DNNs for video classification. In [10], the authors studied the challenging problem of categorising videos according to high-level semantics such as the existence of a particular human action or a complex event. The authors proposed a novel unified approach that jointly exploits the feature and class relationships to improved categorisation performance. Particularly, these two types of relationships are estimated and applied by rigorously imposing regularisation in the learning process of a DNN.

All the above mentioned works share a similar point of view to extract features, which are later used to train a classifier. In this paper, we propose a different focus, i.e. to apply a text-mining approach to the problem of video categorisation, deriving high-quality information from semantic information extracted over visual objects of the video.

3 Methodology

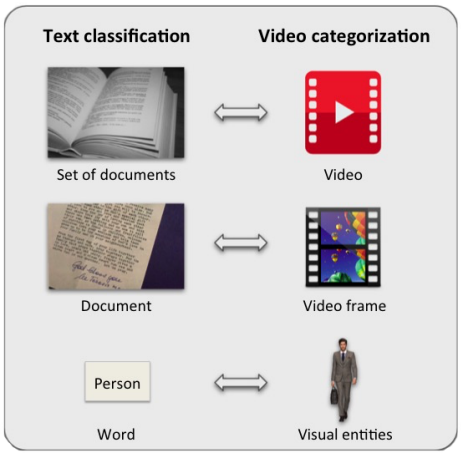


Fig. 1. Equivalence among the elements used in text classification and those used in video categorisation.

The proposal of this paper is based on mimicking the text mining pipeline for the video categorisation problem. Firstly, it is necessary to establish the

analogies between the elements in which a video can be broken down; and the elements that make up a set of documents. In fact, a focus to find a equivalence between a video, which is composed of frames, and a document, which is made up of paragraphs or pages, may be text classification. Thus, a frame is formed by a set of visual entities, that are equivalent to the set of words that appear in the paragraph or page. The above analogies are clarified in Figure 1. As a result, a relationship between the concepts used in text classification and those used in video classification can be established.

The first process in text mining is called tokenization. Similarly, in video categorisation firstly it is needed to extract the visual entities from the frames. Visual entities are extracted using an object detector based on a deep neural network. In this way, labels such as *person*, *dog*, *car* are obtained for the video frames. The semantic information of each visual entity is then considered as a word in our approach. But, only those words that the confidence of the visual entities are larger than a confidence (α). This information is extracted from the corresponding class of the detected entity. These are the basics to manage video categorisation as text classification.

Fig. 2. Approach Overview. Semantic visual entities are extracted from the video frames to apply text mining. Then, a classification process is carry out to categorise the video.

At this point, it is necessary to transform a video, $V = \{f_1, f_2, ..., f_j, ..., f_{|V|}\}$, into a feature vector of fixed dimension for each video. This problem is similar to transform a document into a feature vector. For this purpose, we employ the concept Bag of Words (BoW), which allows to encode documents, in our case a video, in features vectors regardless of the number of frames of the video. Applying a BoW technique consists in converting the token sets of each video frame into sparse vectors of the size of the vocabulary ($Voc = \{t_1, t_2, ..., t_i, ..., t_{|Voc|}\}$) built from the tokens that compose each frame. Formally, $\mathbb{R}^l \rightarrow \mathbb{R}^{|Voc|}$ where l is the number of tokens in the video. Thus, we obtain as many vectors as videos in which the values that are not 0 depend on the chosen method to weight the presence of vocabulary tokens in each frame (w_{ij}). The simplest technique is the Boolean model, where 0 is assigned if a token do not appears in a frame, otherwise, 1. However, the boolean model does not take into account the relevance of each token. In order to consider the relevance of each token, the model, proposed by [21], Term Frequency - Inverse Document Frequency (TF-IDF) is used; and it is defined as follows:

$$w_{ij} = \begin{cases} 0 & \rightarrow \quad tf_{ij} = 0 \\ tf_{ij} \times idf_i & \rightarrow \quad tf_{ij} \geq 1 \end{cases} \tag{1}$$

where

$$idf_i = \log \frac{|V|}{|f_j \in V : t_i \in f_j|} \tag{2}$$

being idf_i the inverse document frequency. In our case, it corresponds to the inverse video frequency of the token i; and t_f is the Term Frequency of the token i in a frame j.

In summary, our approach is composed by three modules as summarised in Figure 2. First, a deep neural network object detector is used to extract the visual entities (words) for each video frame. Second, text mining techniques are used to obtain a feature vector, which is composed from the extracted visual entities of the whole set of processed videos. Third, different models can be adopted to obtain the category of the video.

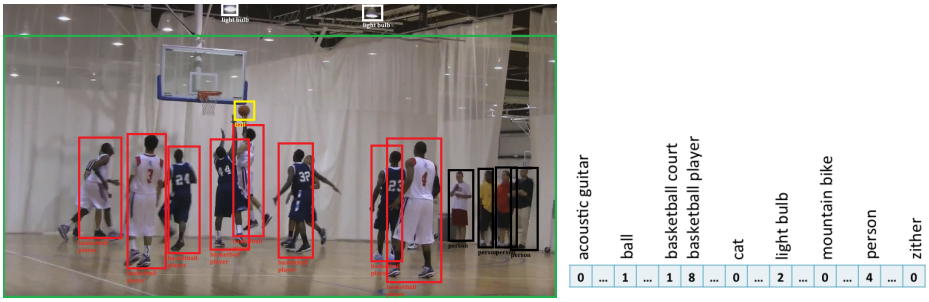
4 Experiments

The experimental evaluation is performed on two widely used benchmarks. First, a subset of videos from YouTube-8M, see [1], is selected. This dataset consists of 200 videos corresponding to ten different categories. These are related to sports (Basketball, Bowling, Cycling, Football, Jumping, Parachuting, Rallying, Surfing, Tennis and WinterSport). Second, Columbia Consumer Video (CCV) dataset is selected, see [11]. This collection consists of 9,317 videos. However, just 7,578 of them could be downloaded due to broken links. Either the user who uploaded the video has deleted or blocked it, or YouTube has deleted the video due to copyright infringement. Once both datasets were obtained, we followed a protocol regarding training and testing. A repeated holdout validation was adopted, including 10 repetitions with re-sampling of the samples, making use of two third of the samples to train and the rest to test.

Table 1. Results in terms of mAP for the subsets built from YouTube-8M and CCV with different classifiers.

	NB	**SVM**	KNN	C4.5	RF
YouTube-8M	94.00	**98.50**	90.60	82.40	94.00
CCV	61.20	**64.30**	53.10	46.30	58.60

As general purpose object detector to get the visual entities, YOLO9000 [20] is used. The number of different object that this detector is able to detect is 9,000. For each video, a vector $\mathbf{X}_v$ of 9,000 elements is obtained as follows. Each frame of the video is processed with YOLO, making a vector $\mathbf{X}_f$ where each position represents how many entities of each type appear in the frame (Fig. 3), considering those entities with an $\alpha \geq 0.1$. Then, $\mathbf{X}_v$ is computed as the sum of all the $\mathbf{X}_f$ of the video. So, $\mathbf{X}_v$ gives how many object of each class are detected in the video. After that, a matrix M where each row corresponds to



a) b)

Fig. 3. a) Example of yolo9000 detections. b) Resulting translation from a frame into a vector where each column represents the number of occurrences of the entities.

the $\mathbf{X}_v$ of each video is obtained. To note that M is the TF matrix of the video but as it was explained before in this work TF-IDF is going to be used. So, a matrix M_{tf-idf} is computed for M using (2) with $|V| = 9,000$. From now on, the problem of video classification can be considered as a traditional classification problem because each video is represented with a 9,000 element feature vector independently of the number of frames the video comprises, being the M_{tf-idf} the dataset for training and testing the different classifiers under consideration using a holdout approach.

To classify the samples, classifiers with different kinds of heuristics were used: Naïve Bayes (NB) proposed by [2]; Decision Tree (C4.5) with pruning is used to avoid overfitting, it is introduced by [19]; the ensemble method proposed by [15], denominated as Random Forest (RF) is used with 10 trees; Support Vector Machine (SVM) presented by [6] with polynomial kernel (C=1); and K-Nearest Neighbors (KNN), proposed by [7], is applied with K set to 10.

A first experiment is carried out to assess the validity of the proposal in the subset of YouTube-8M with only 10 video categories as it was explained before. After this first experiment, the CCV dataset is used to evaluate the performance of the proposal that is a real challenge dataset with 20 video categories.

The approaches taken for comparison purposes are: [11], [30], [27], [16], [14], [17], [8], [18], [26] and [10]. These are the state of the art in the last years in CCV dataset.

The software/hardware requirements to carry out the experiments comprise an i7 core with a Nvidia GeForce GTX 960 graphic card, where YOLO9000 was executed with Tensorflow and Keras library. The classification methods used Weka library running over Java programming language.

5 Results

In this section, we evaluate the validity of our approach. First, we describe the evaluation metric used to establish a comparison. Then, we detail the obtained

results in a subset of YouTube-8M. Furthermore, we present the results related to CCV. Finally, we compare the performance of our approach and other methods in CCV dataset. We evaluate our methodology using a well known in video categorisation problems, as it is Mean Average Precision (mAP).

The obtained comparison results with for YouTube-8M subset are presented in the first row of Table 1. The proposed approach performs very well, achieving mAPs larger than 80% for all models. SVM yields the best performance, 98.5%, which is 16 percentage points better than the worst classifier (C4.5) and four over the second classifiers (NB and RF). With this toy example, we verify the validity of our methodology for video categorisation.

In relation to CCV, the achieved results are summarizes in the second row of Table 1. Again the best mAP is obtained for SVM, but just 64.3%, three percentage points larger than NB, the second best classifier, and above 18 points larger than C4.5.

Table 2. Results obtaining by different methods in CCV.

Method	mAP
$\propto$ SVM [14]	43.60
SIFT + STIP + MFCC [11]	59.50
Feature Weighting via Optimal Thresholding (FWOT) [27]	60.30
Reduced Analytic Dependency Model (RADM) [17]	63.04
Robust Late Fusion (RLF) [30]	64.00
Regularized Multi-modality Auto-Encoders (RMAE) [8]	64.00
Our (SVM)	**64.30**
Sample Specific Late Fusion (SSLF) [16]	68.20
Students-t Mixture Model + Temporal Pyramids (StMM+TP) [18]	71.70
regularized Deep Neural Network (rDNN) [10]	73.50
Multi-Stream Multi-class Fusion (MSMF) [**26**]	84.90
Labelling by humans	77.40

Finally, Table 2 presents the results obtained from different authors over CCV, including the results reported by our approach using a SVM classifier. Rank 5 is reached with respect to the state of the art methods in CCV. Most of the approaches that have a lower score make use of Scale-Invariant Feature Transform (SIFT), Spatial-Temporal Interest Points (STIP) as features. These features are not significant to extract independent features from different categories. Our result is achieved using just BoW features, contrary to other approaches which combines multiple features and modalities. Furthermore, it is interesting to compare with the reported human performance, 77.40%. The latter suggests the complexity of categorising the CCV dataset. It is awe-inspiring that some methods over-match the understanding of the humans about video categorisation. For instance, in the case of [26] is 7.5% over labelling by humans. Labelling by humans is presented by [11].

6 Conclusions

This paper has focused on the use of a novel technique for the problem of video categorisation. The adopted approach, based on text mining, extract visual entities from videos, which are represented textually. This gives rise to manage such textual information following the standard text mining protocol, applying tokenization and BoW.

Results are evaluated over two datasets, a subset of YouTube-8M to evaluate at first step whether our hypothesis is feasible. Reaching a 98.5% in terms of mAP with SVM classifier. Moreover, CCV dataset is used to test in a challenging categorisation set where humans are not capable to classify correctly the different categories, reaching a 77.40% of mAP. Our approach achieves 64.30% using SVM as classifier.

Those results suggest the possibility to make use of *forgotten* features in video categorisation, which may in a close future be combined with state of the art approaches to boost the overall performance.

Acknowledgements

This research work has been partially supported by the Spanish Ministry of Economy and Competitiveness (TIN2015-64395-R MINECO/FEDER), by the Office of Economy, Industry, Commerce and Knowledge of the Canary Islands Government (CEI2018-4), and the Computer Science Department at the Universidad de Las Palmas de Gran Canaria.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. ArXiv Preprint arXiv:1609.08675 (2016)
2. Bayes, T., Price, R., Canton, J.: An essay towards solving a problem in the doctrine of chances (1763)
3. Brezeale, D., Cook, D.J.: Automatic video classification: A survey of the literature. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **38**(3), 416–430 (2008)
4. Chang, S.F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A.C., Luo, J.: Large-scale multimodal semantic concept detection for consumer video. In: Proceedings of the International Workshop on Multimedia Information Retrieval. pp. 255–264. ACM (2007)
5. Convivia: https://www.conviva.com/ (2018), online; accessed 13 December 2018
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
7. Fix, E., Hodges Jr, J.L.: Discriminatory analysis-nonparametric discrimination: consistency properties. Tech. rep., California Univ Berkeley (1951)
8. Jhuo, I.H., Lee, D.: Video event detection via multi-modality deep learning. In: International Conference on Pattern Recognition. pp. 666–671. IEEE (2014)

9. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 352–364 (2018)
10. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 352–364 (2018)
11. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: Proceedings of the ACM International Conference on Multimedia Retrieval. p. 29. ACM (2011)
12. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
13. Kowdle, A., Chang, K.W., Chen, T.: Video categorization using object of interest detection. In: IEEE International Conference on Image Processing. pp. 4569–4572. IEEE (2010)
14. Lai, K.T., Felix, X.Y., Chen, M.S., Chang, S.F.: Video event detection by inferring temporal instance labels. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2251–2258. IEEE (2014)
15. Liaw, A., Wiener, M., et al.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)
16. Liu, D., Lai, K.T., Ye, G., Chen, M.S., Chang, S.F.: Sample-specific late fusion for visual category recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 803–810. IEEE (2013)
17. Ma, A.J., Yuen, P.C.: Reduced analytic dependency modeling: Robust fusion for visual recognition. International journal of computer vision **109**(3), 233–251 (2014)
18. Nagel, M., Mensink, T., Snoek, C.G., et al.: Event fisher vectors: Robust encoding visual diversity of visual streams. In: British Machine Vision Conference. vol. 2, p. 6 (2015)
19. Quinlan, J.R.: Induction of decision trees. Machine learning **1**(1), 81–106 (1986)
20. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger (2016)
21. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
22. Sun, Y., Wu, Z., Wang, X., Arai, H., Kinebuchi, T., Jiang, Y.G.: Exploiting objects with lstms for video categorization. In: Proceedings of the ACM on Multimedia Conference. pp. 142–146. ACM (2016)
23. Truong, B.T., Dorai, C.: Automatic genre identification for content-based video categorization. In: Proceedings of the International Conference on Pattern Recognition. vol. 4, pp. 230–233. IEEE (2000)
24. Wang, J., Duan, L., Xu, L., Lu, H., Jin, J.S.: Tv ad video categorization with probabilistic latent concept learning. In: Proceedings of the International Workshop on Multimedia Information Retrieval. pp. 217–226. ACM (2007)
25. Wu, X., Zhao, W.L., Ngo, C.W.: Towards google challenge: Combining contextual and social information for web video categorization. In: Proceedings of the ACM International Conference on Multimedia. pp. 1109–1110. ACM (2009)
26. Wu, Z., Jiang, Y.G., Wang, X., Ye, H., Xue, X.: Multi-stream multi-class fusion of deep networks for video classification. In: Proceedings of the ACM on Multimedia Conference. pp. 791–800. ACM (2016)

27. Xu, Z., Yang, Y., Tsang, I., Sebe, N., Hauptmann, A.G.: Feature weighting via optimal thresholding for video analysis. In: IEEE International Conference on Computer Vision. pp. 3440–3447. IEEE (2013)
28. Yang, L., Liu, J., Yang, X., Hua, X.S.: Multi-modality web video categorization. In: Proceedings of the International Workshop on Multimedia Information Retrieval. pp. 265–274. ACM (2007)
29. Yang, X., Molchanov, P., Kautz, J.: Multilayer and multimodal fusion of deep neural networks for video classification. In: Proceedings of the ACM on Multimedia Conference. pp. 978–987. ACM (2016)
30. Ye, G., Liu, D., Jhuo, I.H., Chang, S.F.: Robust late fusion with rank minimization. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 3021–3028. IEEE (2012)
31. Yuan, X., Lai, W., Mei, T., Hua, X.S., Wu, X.Q., Li, S.: Automatic video genre categorization using hierarchical svm. In: IEEE International Conference on Image Processing. pp. 2905–2908. IEEE (2006)
32. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4694–4702 (2015)
33. Zhou, Y., Song, W.: Video classification algorithm based on improved k-means. Technical Bulletin **55**(1), 138–144 (2017), www.scopus.com