JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Thabet, M., Ellouze, M., & Zaied, M. (2021). A new approach for video concept detection based on user comments. *Journal of Information and Communication Technology, 20*(4), 629-649. https://doi.org/10.32890/jict2021.20.4.7

A New Approach for Video Concept Detection Based on User Comments

[1]Maha Thabet, [2]Mehdi Ellouze & [3]Mourad Zaied
[1]ISITCom, University of Sousse, Tunisia
[2]Faculty of Economics and Management, Sfax University,Tunisia
[3]Research Team in Intelligent Machines, Gabes University, Tunisia

maha.thabet@fsm.rnu.tn
mehdi.ellouze, mourad.zaied@ieee.org

Received:1/11/2020 Revised: 27/1/2021 Accepted: 7/3/2021 Published: 27/9/2021

ABSTRACT

Video concept detection means describing a video with semantic concepts that correspond to the content of the video. The concepts help to retrieve video quickly. These semantic concepts describe high-level elements that depict the key information present in the content. In recent years, many efforts have been done to automate this task because the manual solution is time-consuming. Nowadays, videos come with comments. Therefore, in addition to the content of the videos, the comments should be analyzed because they contain valuable data that help to retrieve videos. This paper focused especially on videos shared on social media. The specificity of these videos was the presence of massive comments. This paper attempted to exploit comments by extracting concepts from them. This would support the

research effort that works only on the visual content. Natural language processing techniques were used to analyze comments and to filter words to retain only the ones that could be considered as concepts. The proposed approach was tested on YouTube videos. The results demonstrated that the proposed approach was able to extract accurate data and concepts from the comments that could be used to ease the retrieval of videos. The findings supported the research effort of working on the visual and audio contents of the videos.

Keywords: Keywords-based video retrieval, social media tagging, natural language processing, video concept detection.

INTRODUCTION

In video processing, the semantic concepts consist of labels like objects or people, which can be depicted by a video or can comprise a situation with a complex interaction of different entities. In the last few years, many efforts have been done in this research field. TREC Video Retrieval Evaluation (TRECVID) is the most known evaluation campaign for video concept detection systems. It is a contest organized by the National Institute of Standards and Technology (NIST) in the United States of America (USA) (Awad et al., 2016). It provides hours of video clips manually annotated by the participants involved in this competition. The annotated database is divided into two parts: development and testing. Participants will train their systems on a development database. After that, the systems will be tested on the other part to evaluate their performance in detecting targeted visual concepts. The performance of video concept detection systems is affected by the size and quality of the training database. The more important and varied the database is, the better the result. Making a training database becomes a challenging problem. Thousands of hours of human efforts are required.

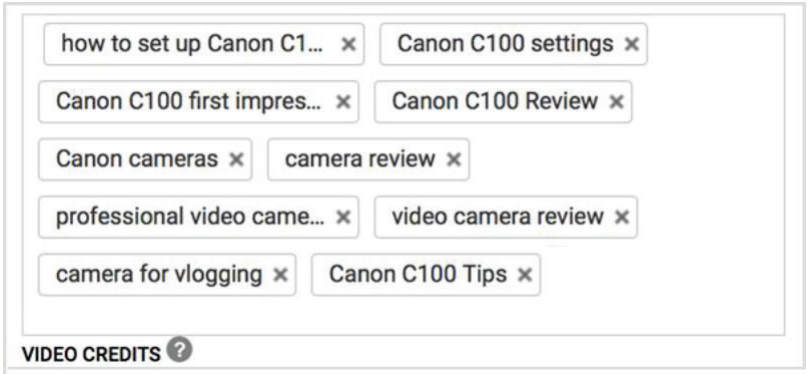
In TREVID for instance, the participants share the effort of tagging the database manually to prepare the training database. However, since 2012, a new task has appeared in TRECVID. It is called “no annotation semantic indexing”. The idea is to promote the development of methods that permit the indexing of concepts in video shots using only data from the web or archives without the need of additional

annotations (manual or automatic). The training data could be, for instance, images retrieved by a general-purpose search engine (e.g. Google).

In this paper, the problem of automatic video concept detection on social media is addressed. On social media, people do not simply share the multimedia content, they also give comments.

Figure 1

Tags in Video Clips in YouTube



This paper will study this opportunity. The main contribution of this paper is to show how the comments can be used to extract semantic concepts present in the video. Nevertheless, the present study does not assume that comments can be a perfect solution for extracting semantic concepts. It aims to exploit the comments to support the research effort working on the visual content, motivated by the new working trends of YouTube (Google, 2013).

In this paper, the reliability of comments is investigated to help extract semantic concepts. To the best of the authors' knowledge, this paper is the first to present an approach for extracting semantic concepts from video comments. Works proposed in the literature are based either on analyzing the content of videos or on tags submitted by the video uploaders (see Figure 1). The rest of the paper is organized as follows: Section 2 discusses the related work. In Section 3, the research contribution is presented. Section 4 provides the detail of the

proposed approach. Results of the approach are shown in Section 5. Lastly, the paper is concluded with directions for future research.

BACKGROUND AND RELATED STUDIES

Video Concept Detection

Video concept detection is the task of classifying a video shot to detect the presence of a visual concept. Examples of video concept detection systems can be cited in the pioneering work of Carnegie Melon called the Informedia Project (Li et al., 2012; Yu et al., 2015). In this work, the authors combined techniques from computer vision, speech recognition, natural language processing, and artificial intelligence into an integrated system to develop concept detectors. They started from the assumption that a concept is generally described by visual, textual, and auditory features.

MediaMill (Huurnink et al., 2012; Snoek et al., 2015; Snoek et al., 2017) is a work proposed by the University of Amsterdam. The main contribution of this work is the fact there is no standard way for detecting all concepts inside a video shot. For every concept, specific features should be prioritized. The authors introduced the notion of a semantic pathfinder. It means that to retrieve a concept, a special path should be followed.

LIG lab also proposed an interesting work in video concept detection (Safadi et al., 2014; Safadi et al., 2015). It consists of using six-stage processing pipelines for computing scores for the likelihood of a video shot to contain a target concept. Different kinds of descriptors are used for the three video modalities (visual, textual, auditory); they are optimized and fused before being used by the classifiers.

Actually, all approaches alter their processing and instead of using simple features (descriptors), they use bag-of-words (BOW). They consider the image like a text in which humans have words. In the image, the words are feature vectors that characterize the visual concepts. Every concept will be characterized by a set of vectors. The used features are generally local descriptors like scale-invariant feature transform (SIFT) descriptors.

In TRECVID 2015, Dublin City University proposed an approach for concept detection based on BOW (McGuinness et al., 2015; Mohedano et al., 2016). Features are computed based on convolutional neural network (CNN) activation features, while BOW is generated through K-means clustering. The cluster centroids are the vectors that will represent the concept. All the cited works require an important number of training samples to make their concept detectors be able to detect visual concepts efficiently.

Recent works have become specialized in particular concepts. Video concept detection systems are no longer talked about in general. In particular, work on violence concept detection, video surveillance events detections, etc. are discussed in the research community (Landi et al., 2019; Muchtar et al., 2020; Peixoto et al., 2020). The detectors of these concepts are customized and adapted to these kinds of concepts. The descriptors used to extract these concepts are local descriptors.

When reviewing approaches that work only on the visual part of videos, it is noticed that there are important advances and this is due to important efforts done on image processing and computer vision. However, the community still remain dependent on these techniques and on image processing concept detectors and their performance. Moreover, a video could not be described with only visual indexes. However, the semantics of the video could not be perceived using only the visual concepts.

Tag-Based Annotation

Annotating an image or a video clip means tagging it with keywords that briefly describe its visual content. This action is usually performed by the owner of the image or the video while uploading the multimedia content. However, the main challenge remains on how to surmount the probable irrelevancy of the proposed tags. In the literature, there are many proposed approaches that refine tags to generate the final annotations.

Li et al. (2008) proposed a learning social tag relevance by neighbor voting. Common tags made on images having the same content by different users were considered relevant. Otherwise, they were considered subjective.

In Sawant et al. (2010), the authors used users' local interaction network for automatic image tagging. To annotate an image of a given user in a social network, their interactions on the network were observed. The study was based on the collective tag vocabulary of the users with whom a user interacts to generate the final annotations.

Zhu et al. (2010) proposed an approach for image annotation based on the assumptions that on social media, similar images are similarly tagged by users. In every image, tags were divided into two parts: accurate and faulty tags. The problem of the annotation was cast into a decomposition of the user-provided tag, which was further divided into two matrices, a matrix for relevant tags and a matrix of faulty tags. The problem became a constrained optimization problem.

Liu and Forss (2015) developed a system that used rules and term weighting method to extract tags from tweets related to images. The system retrieved tweet-image pairs from public Twitter accounts, analyzed the tweets, and generated labels for the images.

In Ballan et al. (2015), Mettes et al. (2017), and Uricchio et al. (2017), the authors proposed an interesting work for video annotation. The video was different from the image by its temporal aspect. The authors addressed the problem of dispatching tags on the frames of the video clip. Tags made by users were refined and filtered. Every retained tag was searched through web sources to look for associated images. Retrieved images were matched with the original video (visual similarity) to link the tag with the corresponding frame.

Kordumova et al. (2015) investigated the use of tags made on social media to make a database of positive examples that could be used to train visual concept detectors. As stated, the main challenge of visual concept detectors was to have at disposal a good database of positive examples to train concept detectors.

Moreover, other systems like in Mansouri et al. (2018) and Cagliero et al. (2019) worked on overlay text or text embedded in the video in general. Overlay text was segmented and analyzed to extract concepts like locations, persons, events, etc. The overlay text was generally located in scripted videos, such as news or documentaries. In contrast, it was mostly absent in personal videos. For this reason, these approaches are useful only for specific kinds of videos. However, overlay text remains an important information that should be exploited when it is present.

It is noticed that there is an awareness about the importance of data that come with the video content. However, the literature does not have any approaches that conduct a deep analysis of users' comments. Most of them worked on tags made by the uploaders. Certainly, tags are important but they are subjective and reflect the point of view of the uploaders.

METHODOLOGY

Contrary to all proposed approaches, this student analyzed the comments that accompanied videos on social media. The present research was based on the observation that comments could contain information describing the content of the video. Furthermore, motivated by the new trends in YouTube (Google, 2013), natural language processing (NLP) techniques and pattern recognition techniques were applied to retrieve effective concepts that could annotate videos.

The proposed system consisted of four modules. The first module focused on comments acquisition from the social media. This task was known as the crawling step. The targeted social media was YouTube. The second module comprised preprocessing these comments by cleaning them and removing all stop-words. The third module involved extracting candidate words that could be considered as visual concepts. Retained words were grouped into clusters. The present researchers aimed at extracting the cluster of concepts representing the ground truth of the video. Figure 2 shows the overall process of the proposed approach.

Figure 2

Overall Process of the Proposed Approach

Comment Crawling

This paper was interested only on videos with more than one thousand comments. In order to address this task, a focused crawler was implemented. Based on the video URL, it extracted comments of that video using web API through HTTP GET method.

Word Filtering

Comment Preprocessing: The extracted comments were mixed. Therefore, some preprocessing was carried out on these unstructured comments to generate the datasets.

Words that could be used as tags were the only ones kept for this research purpose. Then, the following preprocessing tasks were applied: first, Part-of-Speech tagging and data cleaning (removing stop-words) using Laurence Anthony's Software (Anthony, 2020) were employed. Second, all the irrelevant expressions were removed, like dates, emoticons, links, and special characters.

It was considered that to be a candidate of a real concept present in the video, there had to be a kind of consensus in the comments. This meant that the word should be recurrent in many of the comments. For this reason, all the words in the grammatical class "name" were extracted and sorted according to their redundancy and recurrence. In the literature, term frequency–inverse document frequency (TF-IDF) (Salton & McGill, 1986) is widely used in information retrieval to weigh redundancy and recurrence of the words. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse

document frequency of the word across a set of documents. TF-IDF gives important scores for words redundant in a comment and not recurrent in the other comments. However, it penalizes words that are redundant along the whole comments. For this reason, the present study used another metric known as TF-DF.

For this reason, we used another metric. We called it Frequency Score (FS) that made up of TF and DF. TF, DF and FS are computed using Equations 1, 2, and 3.

$$TF(W) = \frac{Number\ of\ occurrences\ of\ W\ in\ comments}{Total\ Number\ of\ Words\ in\ the\ comments} \quad (1)$$

$$DF(W) = \frac{Number\ of\ comments\ containing\ W}{Total\ Number\ of\ comments} \quad (2)$$

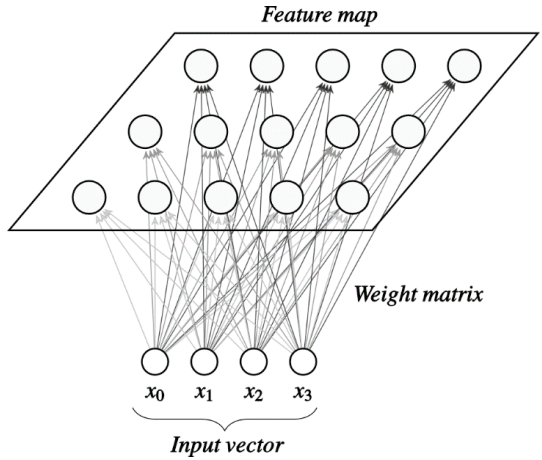
$$FS(W) = TF\ (W)\ X\ DF\ (W) \quad (3)$$

Word Clustering

It was believed that concepts evoked in the video were semantically related. For this reason, this study made a clustering of all remaining words and attempted to discover the predominant cluster. It was considered as the cluster that enclosed the main visual concepts displayed in the video. Self-organizing maps (SOM) have been already used for text clustering (Klami & Lagus, 2006; Pacella et al., 2016). They are well adapted for doing such kind of tasks. The main advantage of SOM is their ability in providing a data visualization technique that helps to understand high dimensional data by reducing the dimension of data to the map. SOM clusters similar data together without needing to know the number of clusters in advance like K-means or K-nearest neighbor (KNN) clustering algorithms.

Figure 3

Overall Process of the Proposed Approach



SOM was proposed in the 1980s by Kohonen (1982). The proposed algorithm produces a topological map organization. The process depends only on the inputs and does not require the intervention of a supervisor.

SOM tends to cluster the set of observations in similar groups. It is composed of two layers: an input layer and an output layer. The input layer is made up of nodes, whereby each of them is represented by a vector of dimension n and connected to the m output nodes through weights (See Figure 3).

Algorithm 1 demonstrates Kohonen map as in Equations 4,5, and 6.

Algorithm 1 : Kohonen Map

Step 1 : Random initialization of weight links W_{ij}

Step 2 : Determination of the winner neuron by selecting the closest neuron j^* to the input:

$$j^* = \arg min_j \sum \left(X_i(e) - W_{ij}(t)\right)^2 \quad (4)$$

(continued)

Algorithm 1 : Kohonen Map

Step 3 : The weights of the wining neuron and their neighbors are updated at every iteration as follows in Equation 5:

$$W_{ij}(t) = W_{ij}(t-1) + a(t)h_j(j*,t)\left[X_i(t) - W_{ij}(t-1)\right] \quad (5)$$

Where ***t*** represents the time, *a(t)* is a variable decreasing with time, and *h(t)* represents a neighborhood function. The idea is to reduce the influence when the neighborhood radius increases.

During the learning of SOM, Steps 2 and 3 are repeated many times until the weights become stable.

Computing Features: In this study's context, the input of SOM is a similarity matrix. If there are *N* candidate words, then the matrix will have a size of $N \times N$, where the entry at position *(i, j)* is the distance computed between the word *(i)* and the word *(j)*. The word *(i)* will be represented by a vector, the i^{th} column of the similarity matrix, which contains *N* component that represents the distances between the word *(i)* and the *N-1* remaining words. The output will be the map in which every word is mapped into a given node (Figure 4). Two similarity measures were tested in this study. All of them made their calculation by using WordNet (Miller, 1995), which is a large lexical database of English. In WordNet, nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

The researchers distinguished two important categories of measures: path-based measures and information content measures. The main idea of path-based measures was that the similarity between two words was a function of the length of the path linking the word and the position of the words in the taxonomy. However, in information content measures, the more common information two words share, the more similar the words are.

Two measures were used:

Resnik's measure (RES): Resnik (1999) defined the similarity between two words by comparing the two synsets to which they belong. The comparison of the two synsets is made by computing the information content of their lowest superordinate (most specific common subsumer).

Wu and Palmer measure (WUP): This measure was proposed by Wu and Palmer (1994). It calculates the similarity by considering the depths of the two synsets in the WordNet taxonomies. Let C1 and C2 be two concepts in the taxonomy; this similarity measure considers the position of C1 and C2 relatively to the most specific common concept C. Several parents can be shared by C1 and C2 through multiple paths. The most specific common concept is the closest common ancestor C.

Concept Extraction

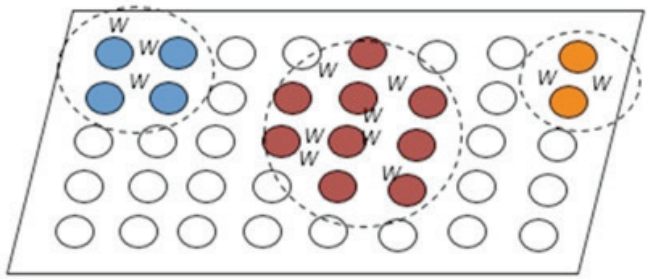
At the end of the training phase and when each word was matched to a node (neuron) of SOM, the present researchers extracted the cluster of effective concepts. The process of extracting the cluster was based on two assumptions:

(i) The cluster of effective concepts included important number of words.

(ii) The cluster of effective concepts was dense. It meant that nodes of this cluster were close to each other.

Figure 4

Clusters in the Self-organizing Map



This study computed the hit histogram for the map. The hit histogram was made by counting the number of words attached to each node. The U-matrix (Ultsch & Siemon, 1990) was also calculated. The U-matrix was obtained by calculating the average of distances between each node and its neighbors. The U-matrix showed clearly similar zones on the map. It was considered that the cluster of effective concepts was occupying an important zone in the map and not in a particular node. The extractions steps are as follows:

(i) Extract the node having the most important value in the hit histogram.

(ii) Based on the U-Matrix, check if that node is in a dense zone. If yes, select this node.

(iii) Else, go the next important node in the hit histogram until a node is found satisfying the two criteria.

(iv) From the selected nodes and from the direct neighbors, retrieve all the words matched to them. These words are considered as the concepts for the video.

RESULTS AND DISCUSSION

Dataset

To show the efficiency of the proposed approach, the current study selected from YouTube 50 video clips. They dealt with different contexts of the human life, for instance “food”, “phone and technology”, “auto vehicles”, “health”, “games”, “animals”, etc. The study was only interested on videos with more than one thousand comments. The average duration of the videos was 9m25s. The total duration was about 8 hours.

Four assessors were invited to make the ground truth of the 50 YouTube video clips. The assessors were not familiar with the proposed approach and knew nothing about it. However, they were regular video consumers. They belonged to different age categories and different backgrounds. For every video clip, they were asked to

write the concepts they identified in the videos. To ensure that they understood their task, a series of blank tests were inserted at the end to ascertain their comprehension.

For instance, when watching the video of Steve Jobs launching an iPhone in 2007, concepts that could be identified by the assessors were "Steve Jobs", "iPhone", "Phone", "Apple", "MacWorld", "touchscreen", etc. These derived concepts represented the ground truth. The concepts that were extracted from the comments would be evaluated against them.

The proposed approach was evaluated by computing the recall and precision between the concepts extracted by the approach and the concepts given by assessors. They are computed as follows in Equations 6 and 7:

$$Recall = \frac{\#correct\ concepts}{\#\ ground\ truth} \tag{6}$$

$$Precision = \frac{\#correct\ concepts}{\#correct\ concepts + \#false\ concepts} \tag{7}$$

Where "#correct concepts" refer to the number of concepts correctly detected, "#ground truth" means the total number of concepts considered as ground truth, and "#false concepts" signify the number of incorrect concepts. The distribution of the recall and precision rates of each assessor's assessments were plotted (see Figures 5 and 6). They were presented as Tukey-style boxplots. Tukey Boxplots (also known as Box and Whisker plots) are used to create a graphic image of the several key measures of distribution including the minimum, maximum, median, and 25th and 75th percentiles (the middle 50%). For every assessor, the results for RES and WUP measures were plotted. The results were generally interesting.

First, they showed that concepts enclosed in the videos could be found in the comments. Second, the proposed approach could detect these concepts. They displayed the ability of SOM to isolate these concepts.

However, there was a certain sparsity in the obtained recall rates. For certain videos, the score was about 90 percent, while for others, it was under 65 percent.

Figure 5

Distribution of Recall Rates for the Assessors

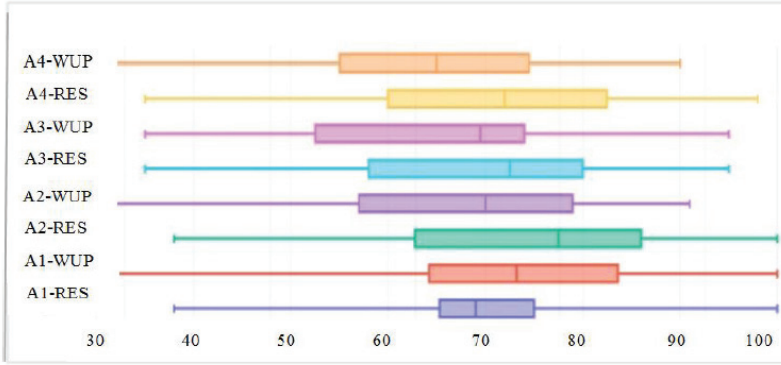
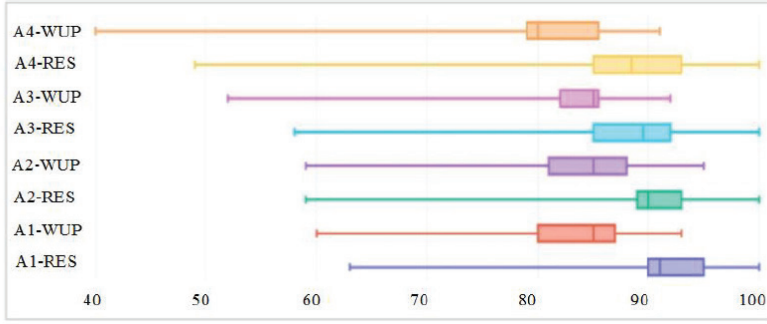


Figure 6

Distribution of Precision Rate for the Assessors



For videos in the context of "phone and technology", "auto vehicles", and "games", the results were better. This was due the enormous number of comments made on these kinds of videos and the large public targeting them. This was not surprising because according to YouTube Academy (Youtube, 2020), these categories are the most popular ones. For the other kinds of videos, the target audience was centered around a very informed and less talkative audience.

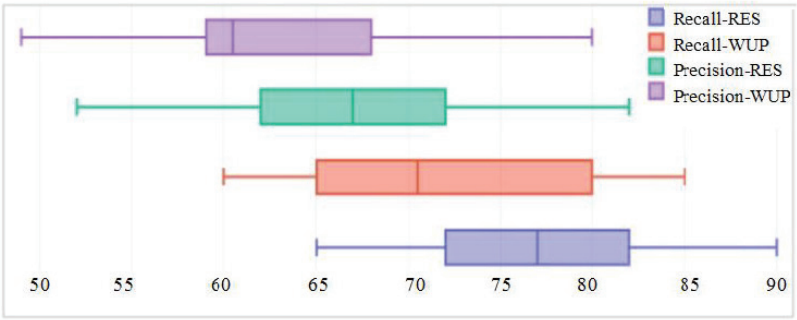
In the obtained results, the RES distance was slightly better than WUP. This confirmed that the semantic distance between two words should be measured by analyzing the context of the word and not by computing the path from one word to another in WordNet.

Evaluation According to Uploader Tags

In YouTube, tags are submitted when the video is uploaded. Through these tags, uploaders aim to ease the search task. The present study attempted to measure the effectiveness of the extracted concepts according to these tags. Regarding the recall rates, many extracted concepts were commonly used as tags by uploaders (Figure 7). In fact, in many of the videos, the number of extracted concepts was more important than those proposed by the uploaders. Therefore, many of the tags proposed by the uploaders that were cited in the comments could be retrieved. However, in some videos and due to the low quality of the comments, many of the tags were not retrieved. This had caused the precision rate to become relatively low. Extracted concepts were effective and semantically correct but they were not evoked as tags in the YouTube videos. This comparison also confirmed the advantage of the RES distance as compared to WUP.

Figure 7

Distribution of Recall and Precision Rates According to Uploader Tags



DISCUSSION

The obtained results proved that comments were a valuable source of information that could help to understand the visual content. This was confirmed by human assessors. In fact, most of the obtained recall rates were more than 60 percent, and a majority of the precision rate were more than 80 percent. This also corroborated the present researchers' assumption that comments contained valuable data but they could not be the perfect solution for describing the visual content. They simply completed the work of visual concept detectors. It could be a cross-validation between the two sources of information.

Moreover, the obtained results showed that the proposed approach was dependent on the number and the quality of the comments. The greater the number of comments, the better the quality of the concepts becomes.

Nevertheless, the number of comments was not the only parameter. The comments should have a good quality to be informative. Indeed, when comments were of low quality (containing non-clear words and symbols), they could distort the results and increase the processing time. It was discovered that when analyzing the comments for several video categories, the quality of the comments were better and this improved the results for these categories ("phone and technology", "auto vehicles", and "games"). This was not a weakness of the proposed approach, but a parameter that should be considered when analyzing user comments.

CONCLUSION

In this paper, the main interest is in extracting concepts from the comments made on videos shared on social media. This was based on the assumption that users on social media evoke the content of the shared multimedia data in their comments. Nevertheless, it is believed that comments cannot be a perfect solution for extracting concepts. This study investigated how these comments could help to retrieve effective video concepts. NLP techniques and SOM were utilized to filter the user comments and to extract them to achieve this task. The obtained results were encouraging and confirmed that the comments

were a valuable source of information for video retrieval (Google, 2013). In fact, comments could be used as a source of information to help with the retrieval of videos on social media or to enrich existing tags. However, the proposed approach depended enormously on the quality of the comments. Not all videos enclosed comments with good quality. For future research, it is recommended to add another step that analyzes the comments and estimate their quality to decide if they can be used as a source of concept.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support of this work by grants from the General Direction of Scientific Research and Technological Renovation (DGRSRT), Tunisia, under the PEJC program.

REFERENCES

Anthony, L. (2021). *Computer Software*. Waseda University. https://www.laurenceanthony.net/software/tagant/

Awad, G., Fiscus, J., Joy, D., Michel, M., Smeaton, A., Kraaij, W., ... & Larson, M. (2016, November). Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *TREC Video Retrieval Evaluation (TRECVID).*

Ballan, L., Bertini, M., Uricchio, T., & Del Bimbo, A. (2015). Data-driven approaches for social image and video tagging. *Multimedia Tools and Applications*, *74*(4), 1443–1468.

Cagliero, L., Canale, L., & Farinetti, L. (2019, July). VISA: A supervised approach to indexing video lectures with semantic annotations. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 1, pp. 226–235). IEEE.

Google I/O 2013 - Semantic video annotations in the Youtube topics API: Theory and applications [online]. Available: https://www.youtube.com/watch?v=wf_77z1H-vQ

Huurnink, B., Snoek, C. G., de Rijke, M., & Smeulders, A. W. (2012). Content-based analysis improves audiovisual archive retrieval. *IEEE Transactions on Multimedia*, *14*(4), 1166–1178.

Klami, M., & Lagus, K. (2006, September). Unsupervised word categorization using self-organizing maps and automatically extracted morphs. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 912–919). Springer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Kordumova, S., Li, X., & Snoek, C. G. (2015). Best practices for learning video concept detectors from social media examples. *Multimedia Tools and Applications*, *74*(4), 1291–1315.

Landi, F., Snoek, C. G., & Cucchiara, R. (2019). *Anomaly locality in video surveillance*. arXiv preprint arXiv:1901.10364.

Li, H., Shi, Y., Liu, Y., Hauptmann, A. G., & Xiong, Z. (2012). Cross-domain video concept detection: A joint discriminative and generative active learning approach. *Expert Systems with Applications*, *39*(15), 12220–12228.

Li, X., Snoek, C. G., & Worring, M. (2008, October). Learning tag relevance by neighbor voting for social image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval* (pp. 180–187).

Liu, S., & Forss, T. (2015, November). Automatic tag extraction from social media for visual labeling. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)* (Vol. 1, pp. 504–510). IEEE.

Mansouri, S., Charhad, M., Rekik, A., & Zrigui, M. (2018, August). A framework for semantic video content indexing using textual information. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)* (pp. 107–110). IEEE.

McGuinness, K., Mohedano, E., Salvador, A., Zhang, Z., Marsden, M., Wang, P., & Smeaton, A., (2015). Insight Dublin City University at Text Retrieval Conference-Video Track 2015. In *Text Retrieval Conference - Video Track 2015 Overview Papers and Slidespp* (pp. 1–16).

Miller, G. (1995). WordNet: A lexical database for English. *Communications of the Association for Computing Machinery*, *38*(11), 39–41.

Mohedano, E., McGuinness, K., O'Connor, N. E., Salvador, A., Marques, F., & Giró-i-Nieto, X. (2016, June). Bags of local convolutional features for scalable instance search.

In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (327–331).

Mettes, P., Snoek, C. G., & Chang, S. F. (2017). *Localizing actions from video labels and pseudo-annotations*. arXiv preprint arXiv:1707.09143.

Muchtar, N., Afdhal K., Dwiyantoro A., & Prayuda A. (2020). Deep anomaly detection through visual attention in surveillance videos. *Journal of Big Data*, *7*(87).

Pacella, M., Grieco, A., & Blaco, M. (2016). On the use of self-organizing map for text clustering in engineering change process analysis: A case study. *Computational Intelligence And Neuroscience*, 2016.

Peixoto, B., Lavi, B., Bestagini, P., Dias, Z., & Rocha, A. (2020, May). Multimodal violence detection in videos. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2957–2961). IEEE.

Safadi, B., Derbas, N., Hamadi, A., Budnik, M., Mulhem, P., & Quénot, G. (2014, November). LIG at TRECVid 2014: Semantic indexing. In *Proceedings of TRECVID*.

Safadi, B., Derbas, N., & Quénot, G. (2015). Descriptor optimization for multimedia indexing and retrieval. *Multimedia Tools and Applications*, *74*(4), 1267–1290.

Sawant, N., Datta, R., Li, J., & Wang, J. Z. (2010, March). Quest for relevant tags using local interaction networks and visual content. In *Proceedings of the International Conference on Multimedia Information Retrieval* (pp. 231–240).

Salton G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Snoek, C. G., Cappallo, S., Fontijne, D., Julian, D., Koelma, D. C., Mettes, P., ... & Towal, R. B. (2015). Qualcomm Research and University of Amsterdam at TRECVID 2015: Recognizing concepts, objects, and events in video. In *TRECVID*.

Snoek, C. G., Li, X., Xu, C., & Koelma, D. C. (2017). University of Amsterdam and Renmin University at TRECVID 2017: Searching video, detecting events and describing video. In *TRECVID*.

Snoek, C. G., Worring M., Geusebroek J., Koelma D. C., Seinstra F. J., & Smeulders A. W. (2008). The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(2006), 1678–1689.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

Ultsch, A., & Siemon, H. P. (1990, July). Kohonen's self-organizing feature maps for exploratory data analysis. In *Proceedings of the International Neural Network Conference.* Kluwer Academic Press.

Uricchio, T., Ballan, L., Seidenari, L., & Del Bimbo, A. (2017). Automatic image annotation via label transfer in the semantic space. *Pattern Recognition*, *71*, 144–157.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 133–138).

YouTube Academy-Last Visited, April 2020: https://creatoracademy.youtube.com/page/home?hl=fr

Yu, S.I., Jiang, L., Xu, Z., Lan, Z., Xu, S., Chang, X., Li, X., Mao, Z., Gan, C., & Miao, Y. (2015). Informedia@ text retrieval conference - Video track 2015 MED. In *National Institute of Standards and Technology, Text Retrieval Conference - Video Track Workshop*.

Zhu, G., Yan, S., & Ma, Y. (2010, October). Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 461–470).