IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

A Framework for Semantic Video Content Indexing Using Textual Information

Sadek Mansouri
LATICE Laboratory
Higher institute of computer science
Mednine, Tunisia
mansouri sadek@hotmail.fr

Mbarek Charhad
Al Madina Al Mounawra, (KSA) 41411
Kingdom of Saudi Arabia.
Taiba university
mbarek.charhad@gmail.com

Ali Rekik
Higher institute of computer science
Gabes Unversity
Mednine, Tunisia
alirekik1@yahoo.com

Mounir Zrigui
LATICE Laboratory FSM of Monastir
Monastir, Tunisia
mounir.zrigui@fsm.rnu.tn.

***Abstract*—In these last years, many works have been published in the video indexing and retrieval field. H owever, e xcept some specific c ases s uch a s s port v ideo w here i ts p ossible t o estimate the set of important events and concepts in the document, this research is generally limited to analyzing low level content. In this paper, we introduce an approach for semantic video indexing that combines two levels of descriptions. First, we extract automatically textuel information from video frames.The second part of our approach consists to exploit linguistic techniques and semantic network in order to extract semantic concepts such as person identity, location name, event type etc .These informations are then used for semantic description of video content .Our proposed approach was tested on video collection of Arabic TV news and experimental results have been satisfying.**

***Index Terms*—Arabic news video,semantic indexing,text detection, conceptual network.**

I. Introduction

The quantity of audiovisual information has increased dramatically with the emergence of the high-speed Internet and TV channels. In addition, the technological advances in recent years in the field o f i nformatics (storage a reas m ore and more considerable, digitization of data, etc.) have helped to simplify the use of data videos in various areas by the public. The complexity of video data at the level structure and heterogeneity has been the source of various research work. The major challenge of the latter is the establishment of systems to allow the user, even casual, to access and interpret easily the video data. In this context, the description of the content of a document video through the indexing process is a decisive step. In effect, the indexing is present upstream of any treatment approach of video data. The indexing is the operation that is to extract a digital signature or text, which describes the content accurately and concisely. The success of this step depends, as well, the success of any process of access to video data. Text embedded especially the artificial t ext i n v ideo f rames i s o ne o f t he important semantic features of the video content analysis. This type of text is artificially added to the video at the time of editing and provides highlevel information of video content that seems to be a useful clue in the multimedia indexing system. Usually, it provides information about when, where and who elements of the news video events. However, text detection and localization in the video frame is still a challenging problem due to the numerous difficulties resulting from the variety of text features (size, color, and style),the presence of complex background and conditions of video acquisition. The second problem concerns the extraction of knowledge from textual data in order to provide relevant and accurate information. This poses a challenge to the scientific community that must be able to propose effective systems for the extraction of information in particular with the diversity of fields applications and the peculiarity of the studied language .

To treat these various problems, we propose in this article an approach of video indexing using the semantic contents of document. This approach is based on a conceptual description of the contents. Each video document is described by list of concepts (person, localities, etc.). This description makes possible to abstract the semantic content resulting from various sub-media (image, audio, text). . The main challenge is how extract semantic information from text signal in order to provide a high description of video content .

The rest of this paper is organized as follows: In Section 2,we presented state-of-the-art of semantic video indexing systems . Section 3 presents an overview of video indexings levels. In section 4 details the experimentation of proposed approach , followed by conclusions in Section 5.

II. STATE OF THE ART

In this part, we present a categorization of approaches and methods proposed in the literature for video modeling and retrieval. There are two basic classes.

The first class focuses on low-level features extraction [1] [2] from audiovisual information such as color, shape, texture or motion that characterize visual low level content. The major disadvantage of these approaches is the lack of

semantic description. However, users can't express their query to retrieve video segment using semantic description. These kinds of systems dont efficiently resolve the problem of video parsing that exploits semantic content.

Second, the semantic information that makes physical information comprehensible by user. This second level makes possible to support the "interface" between the user and the machine and to exploit thus the video contents more easily. To make possible the complementarity between the two points of views, it will be necessary to design an approach that exploits in the same time the semantic and the signal content [3].

In [6], the authors propose a multilingual information extraction (IE) system for annotating sports videos in English, German, and Dutch using ASR (Automatic speech recognition) tools. The IE components of this system include tools for tokenizing, part-of-speech tagging, knowledge extraction, and coreference resolution. [7], the systems aim is to perform automatic knowledge extraction from Italian TV news. This system also utilizes an ASR tool to obtain the video texts and IE techniques (named entities recognition). Another semantic video annotation application called Rich News has been described in [8], where the authors make use of the resources on the web to enhance the indexing process. The overall system contains the following modules: automatic speech recognition, key-phrase extraction from the speech transcripts and searching the video using key phrases. Moreover, the proposed system allows also manual annotation to ameliorate segmentation results. [9]a system has been implemented to annotate Turkish news video using video text as a source of information and IE techniques including named entity recognition, person entity extraction, co-reference resolution, and semantic event interpretation. For better knowledge, our work presents the first attempt for semantic Arabic news video indexing based on text analysis and information extraction (IE) techniques that subsume low and conceptual features of video content.

III. PROPOSED SYSTEM

In this part, we present an overview of our semantic video indexing system. Fig.1 illustrates the framework of the proposed system, which are based on three levels . The first level puts a focus on low-level processing such as video segmentation, text detection and recognition. The second level seeks for extracting the semantic concepts including named entity such as a name of person, organization, location and event. In the final step, our work is based on the construction and a semantic network that addresses the taxonomic and contextual relations between concepts. This step aims to enhance the semantic content in terms of indexes generated by the second step. We detail the different stages of our proposed system and their goals in the following sub-paragraphs.

A. Level 1:Low-level processing

Key-frames extraction: In this work, we have applied a temporal segmentation based on the following assumption the text in the image requires at least two seconds to be readable by the user, to generate shots. Then for each video shot, the middle image will be selected as a key-frame.

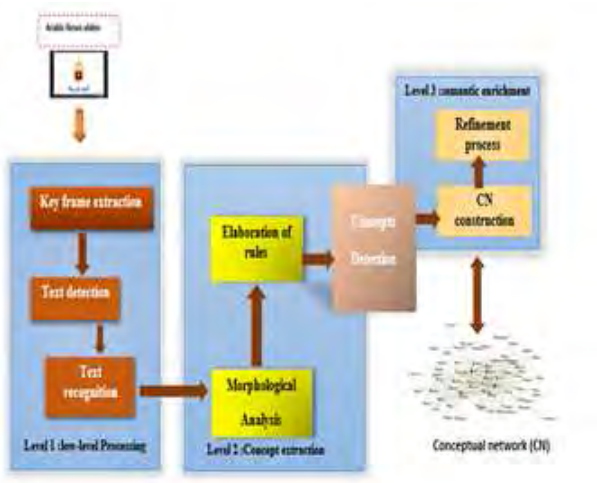

Fig. 1. Proposed system of Arabic news video indexing.

Text detection and localization : After key-frame extraction, text information is detected and extracted from each key-frame. Our text detection method relies on two necessary steps: text detection and text validation. The first step detects connected components (CC) using a hybrid method which combines MSER and edge information. These CC are then grouped by mathematical morphology operators to form candidate text regions. The second stage aims to remove non-text region using geometric constraints and specific signature of Arabic script called baseline (see Fig.2).

Text recognition: After text detection in the video frame, the next step target is to segment and binarize text region in order to separate it from the rest of the frame using Otsus global thresholding method. An optimal threshold is calculated on the tributions of text pixels and non-text pixels. The method abasis of the grey level histogram by assuming Gaussian disims to maximize the interclass variance. In the last stage, commercial OCR engine ABBYY FineReader has been applied for the recognition of text news. More details are shown in our papers [10] [11].

B. Level 2:Conceptual level

A person who is watching a video summarizes it in general by using concepts (identity of person, name of place, etc.), the subject (politic, sport, business, etc.) and sometimes actions to specify these descriptions. This constitutes a way to video content representation. The target of such representation is mainly to get for video document a list of marked points that facilitate access and re-use of content. Considering the heterogeneity of the content from a point of view data (image, audio, and text) and semantic. Indeed, for each video segment, we can associate multiple possibilities of interpretation that can be assorted by specific / generic relationship. We consider that when we have a description issued from a specific media, it consists a way for categorizing the content. For example,

Fig. 2. Text detection :(a) original image ,(b) MSER extraction in image(a) ,(c) Image mask integrating MSERs and canny edges (d) open result ,(e)candidate text regions ,(f) final result.

when we use the action speak or speaking about we suppose that the description is related to audio content. This makes easier the distinction between the multiple possibilities of interpreting the content of the same video segment. In our work, concepts such as person name, location, organization, etc are extracted automatically using NLP techniques such as Morphological Analysis and linguistic rules.

1) Morphological Analysis : Firstly, we segment text in words based on spaces delimiter. Then, we proceed to a In the second step, we parse transcriptions files to extract named entities by comparing each item to the three concepts classes (person identity, the name of a city and organization). This procedure is based on the projection of each news text on the list of keywords called gazetteers. Gazetteers are of a varied nature: lists of first names for the recognition of person names, cities names for the detection of location, etc. Each list is associated with a semantic label which shall be the type of annotation

2) Elaboration of rules : Due to Arabic language complexity and specific characteristics, we also exploit a set of Lexical triggers to extract the name of the person, location and organization not covered by the gazetteer resources (see Table1). To do this, we have used three kinds of rules to improve concepts detections process. This task is object of this publication [12].

TABLE I
A SAMPLE SET OF LEXICAL TRIGGER

Named entity	lexical triggers
Person	وزير , قائد , العقيد , أستاذ , دكتور , مستشار , نائب
Organisation	منظمة , حزب , مؤسسة , جمعية , وزارة , شركة
Location	منطقة , قرية , مدينة , ريف, بلدة , قطاع

The extracted semantic information such as name of person, location, organization and event class is used to annotate the video text and to improve the searching using metadata. The original description are attached to the news video as xml file.

C. Level 3:Semantic enrichment

The semantic enrichment process aims to enrich the semantic interpretation and further enhance the performance of semantic indexing and multimedia retrieval content systems. This task consists of two steps:

1) Construction of conceptual network:: This network consists of set concepts which refers to the politic domain and linked by arcs. The latter denote semantic and contextual relations between concepts nodes.

2) Refinement process: Given an initial set of indexes C= c1, . . . ,cn, the refinement process consists in selecting the most related concepts among the conceptual network(CN) . In the remainder of this work, we will try to propose a measure which we use for the calculation of the relatedness between a candidates concepts in CN and a given set of indexes C.

IV. Experimentation and result evaluation

A. corpus

In order to evaluate the performance of our proposed system in terms of robustness and effectiveness, we used a set of 20 video news (10,000 images) that have been collected from different Arabic TV channels: Aljazeera, Alarabiya, Wataniya 1, Elmayadeen, RT-arabe over the period of September 15 ,2017 until the 5th of December, 2017 and they have a total duration of about two hours. The videos have been automatically transcribed leading to a transcription text of 9704 words. Besides, the named entities extraction phase is done with Farassa [1]platform using Gazetteers and lexical triggers as linguistic resources.

B. Results

1) text detection: .A comparative study with previous systems is performed using precision, recall as the evaluation measures. We applied the evaluation method that has been proposed for the AcTiV-DB Test set, together with evaluation results reported in [13]especially many-to-one matches method. Table II shows that the proposed system achieves excellent results for Aljazeera channel and it is able to outperform the other methods .We can notice the excellent precision rate of

[1]http://qatsdemo.cloudapp.net/farasa/

TABLE II
RESULTS OF THE TEXT DETECTION METHOD

Channel	Method	Precision	Recall
HD(Aljazeera)	Chen [14]	0.67	0.56
	Zayene [4]	0.85	0.83
	our system	0.90	0.87
SD(france 24)	Chen [14]	0.45	0.52
	Zayene [4]	0.75	0.73
	our system	0.71	0.70
SD(RTArabic)	Chen [14]	0.63	0.52
	Zayene [4]	0.73	0.73
	our system	0.75	0.74

our method. This is due to the good rejection ability of false alarms using baseline descriptor. However, this higher score has been decreased For SD channels. This is explained by the fact that the text in these channels is not clearer and the poor quality of graphic text as shown in Fig.3.

Fig. 3. Some detection results from three different SD channels

2) Concepts Extraction: As shown in table III, the results may be satisfactory achieving 80.52% as overall of F-measure. The main reason for these results is the use of grammars rules, which permit the detection of Named entities more precisely. For event extraction , the conceptual feature improve the classification results compared to other approach which based only on textual feature .

TABLE III
EXPERIMENTAL RESULTS OF THE CONCEPTS EXTRACTION METHOD

Concept	Precision	Recall	F-measure
Person	83.02%	79.56%	81.25%
Location	80.23%	77.62%	78.90%
Organisation	82.5%	80.35%	81.41%
Overall			80.52%
Event	85 %	80.3%	82.78%

V. CONCLUSIONS

In this paper, we have introduced a semantic approach for Arabic videos news based on text analysis process and concepts extraction techniques. The experimentation and the evaluation results are promising.

In future work, we will try to improve our concept extraction tool by implementing other rules that cover all structure of Arabic text. In addition, we plan also to use other visual features to enhance detection task especially for video frames with low resolutions.

REFERENCES

[1] C. Zhu, C.-E. Bichot and Liming Chen. Image region description using orthogonal combination of local binary patterns enhanced with color information. Pattern Recogn., vol. 46, no. 7, pages 19491963, July 2013.

[2] R. Vieux, J. Benois-Pineau and J.-P. Domenger. Content based image retrieval using bag-of-regions. In Proceedings of the 18th international conference on Advances in Multimedia Modeling, MMM, 2012.

[3] Yu. Ye, Xu. Rong, X. Yang, Y. Tian: Region Trajectories for Video Semantic Concept Detection. ICMR 2016: 255-259.

[4] O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, and N. E.“ Text Detection in Arabic news Video Based on SWT Operator and Convolutional Auto-encoders”. In Proc of 12th IAPR Workshop on Document Analysis Systems 2016.

[5] Ch. Lhioui, A. Zouaghi, M. Zrigui“ Realization of Minimum Discursive Units Segmentation of Arab Oral Utterances”. Int. J. Comput. Linguistics Appl. 7(1): 31-50 (2016)

[6] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, and Y. Wilks, “Multi-media indexing through multi-source and multi-language information extraction: the MUMIS project,” Data and Knowledge Engineering, vol. 48, pp. 247264, 2004.

[7] R. Basili, M. Cammisa, and E. Donati,“RitroveRAI: A web application for semantic index-ing and hyperlinking of multimedia news,” in Proceedings of the International Semantic Web Conference (ISWC), 2005.

[8] M. Dowman, V. Tablan, H. Cunningham, and B. Popov, “Web-assisted annotation, semantic indexing and search of television and radio news,”in Proceedings of the International Con-ference on World Wide Web (WWW), 2005.

[9] D. Kucuk, AYazc.“ A semi-automatic text-based semantic video annotation system for Turkish facilitating multilingual retrieval”. Expert Systems with Applications, 40(9), 3398-3411.(2013).

[10] S. Mansouri, M. Charhad, M. Zrigui:“A Heuristic Approach to Detect and Localize Text on Arabic NewsVideo” Computacin y Sistemas Journal (2017)in press.

[11] S. Mansouri, M. Charhad and M. Zrigui.“Arabic Text Detection in News Video based on Line Segment Detector”. International Journal of Research in Computing Science(2017).

[12] S. Mansouri,Ch. lhioui, M. Charhad and M. Zrigui . “Text-to-concept: a semantic indexing framework for Arabic News videos” 18th International Conference, CICLing 2017.

[13] O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, and N. E. BenAmara.“A dataset for arabic text detection, tracking and recognition in news videos-AcTiV”. in Proc. of (ICDAR), Nancy, France, 2015.

[14] C. Huizhong.“Robust Text Detection in Natural Images with Edge Enhanced Maximally Stable Extermal Regions”. IEEE ICPR 2011.

[15] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 39(5):489504, 2009.

[16] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, and Y. Wilks. Multimedia indexing through multi-source and multi-language information extraction: The MUMIS project. Data and Knowledge Engineering, 48(2):247264, 2004.

[17] Y. Zhang, Ch. Xu, Y. Rui, J. Wang, and H. Lu. Semantic event extraction from basketball games using multimodal analysis. In Proceedings of the IEEE Conference on Multimedia and Expo (ICME), pages 21902193, 2007.