Audio-Textual Classification System Design for Arabic News Videos
Amal Dandashi, Jihad Al Jaam, and Sebti Foufou
Pages 205 – 212
DOI: http://dx.doi.org/10.21700/ijcis.2016.124

Audio-Textual Classification System Design for Arabic News Videos

Amal Dandashi, Jihad Al Ja'am, and Sebti Foufou
Department of Computer Science, Qatar University
Doha, Qatar
[amal.dandashi, jaam, sfoufou]@qu.edu.qa

Abstract: *The large scale release of raw Arabic news related videos from many sources over the internet has only been increasing. The videos released are uncategorized, and unused. The majority of work aimed at classficiation of Arabic videos is based on textual annotation or closed caption text extraction and processing. We propose a system design that implements multimodal video classification such that annotations and caption processing is excluded. The domain targetted is the news domain. The system consists of audio features extraction and classification, combined with speech-to-text conversion and processing, by utilizing Arabic Named Entity Recognition tools. We also propose to develop a new Arabic dataset based on news channel videos as well as raw videos from various online sources for testing and evaluation. Results are to be documented and graphed.*

Keywords: *Multimodal video classification, audio feature extraction, named entity recognition, news videos.*

Received: July 30, 2016| ***Revised:*** August 10, 2016 | ***Accepted:*** August 25, 2016

1. Introduction

There has been vast proliferation in Digital Arabic Content (DAC) production. However, there have been low advances in developing systems dedicated to the semantic management of DAC. While technologies have been developed attempting to tackle the semantic analysis of DAC, this research aims to do so using multiple modality video classification.

Digital Arabic Content production has been vastly proliferating in the online world. Research work has generated some considerable work pertaining to system-level digital media management and reuse. The predominant features in these systems include the following:

1. Multimedia content exchange between user terminals and home equipment.
2. Semantic search engines provide open source mechanisms that automatically analyse free text available online and build on that content.
3. Open and shared delivery platforms enable digital audio-visual (AV) content between different users.
4. Semantic metadata retrieval and exploitation for the purposes of organisation, content edition, sharing and interactivity.
5. Systems developed address adaptive and personalized content discovery and delivery.

However, none of those technologies support advanced metadata concepts for the retrieval and reuse of multimodal DAC. Multimedia management systems which are semantically-driven concentrate on various types of media and do not cover necessary data mining phases for extraction of semantic meaning from DAC, in a user-conformant style.

TRECVID (http://trecvid.nist.gov/) is a workshop series launched to promote and encourage research in information retrieval and video analysis by providing large test collections, forums, and uniform scoring procedures. Khurana and Chandak [1] have presented a study of various video annotation tools that include automated and semi-automated techniques and employ ontology languages. Zhang et al. [2] have also documented a detailed review on image annotation techniques, highlighting how different researchers have attempted to bridge the semantic gap between low-level image features and high level semantics. Thompson [3] suggested utilizing viewer comments as educational annotation in digital video content in social media. Khoury and others [4] have proposed the Semantic Video Content Annotation Tool (SVCAT) in an effort to address challenges in automated video annotation and usage of models that lack expressiveness. SVCAT is a semi-automatic annotation tool compliant with MPEG-7 standards. The novelty of the SVCAT lies in object localization via automated propagation, and metadata description via contour video tracking, thus alleviating the role of a human annotator.

Chu et al. [5] presented a semantic based content abstraction and annotation method and a semantic patter in an attempt to bridge the semantic gaps of content management. Sanchez et al. [6] introduce a methodology to partially annotate textual Web content in an automatic and unsupervised way. It uses

several well-established learning techniques and heuristics and relies on web information distribution. Jaoua et al. [7] developed a prototype of an Arabic search engine using formal concepts analysis (FCA) and Galois connection. Jaam et al. [8,9] have developed new algorithms for Arabic text summarization, and automatic data classification. Elloumi et al. [10] developed data reduction and redundancy elimination algorithms, and knowledge extraction from Arabic and English news [11]. While technologies have been developed attempting to tackle the semantic analysis of digital content, there exists no global approach or solution to the low-level analysis of Arabic media.

The vast majority of digital Arabic AV content found on social media platforms, information databases, professional archives and such, is available in "raw" form, uncategorized and unclassified. This type of "raw" DAC data is steadily increasing in availability and number day by day. Large portions of this data remain unused and are constantly replaced or edited with new data. The reason that this digital material is not being reused and that large amounts of content are unaccounted for, is that the production chain of such data does not include the vital step of content structuring. The lack of audio-visual search and retrieval tools for capturing unannotated digital media assets is the cause of the lack of reuse of this data. To address these and several other shortcomings of current multimedia systems, the research community has launched initiatives such as the MPEG-7 and MPEG-21 international standards to facilitate content-based representation and description of audio-visual data. However, the problem of inferring the semantics of the content is still an open research issue. Moreover, these standards are not language-oriented and do not accommodate linguistic, social and cultural specifications of Arabic digital media scenes.

In this context, the objective of this study is to design, implement and assess a multimodal system for automated classification of Arabic videos. The proposed system will target both raw and general purpose Arabic content found on a variety of platforms on the Web involving multimodal video processing. Raw material has some distinct constraints that must be considered during development: camera settings, shot-boundary detection, soundtrack irrelevance, redundancy detection, irrelevant annotations, and isolated fragments.

The proposed system will be composed of the following components: utilization of Arabic Named Entity Recognition (NER) for text-based video classification, combined with audio-based analysis to extract patterns for domain-based video classification. The classification domain targeted is "news" videos, pertaining to "shooting" and "explosion".

2. Background

As the proposed approach involves the use of Arabic Named Entity Recognition to extract entities from speech, combined with audio-based pattern extraction for event detection, the background in this section will elaborate on the respective fields.

2.1 Named Entity Recognition

Named Entity Recognition (NER) was initially introduced as an information extraction technique. NER is a task that locates, extracts and automatically classifies named entities into predefined classes in unstructured texts [12]. It covers proper names, temporal expressions and numerical expressions. Proper names are classified into three main groups: persons, locations and organizations. A class can be divided into sub-classes to form an entire hierarchy, i.e., location can be classified into city, state and country. The majority of NER studies have been focused on the English language, as it is the internationally dominant language, while research on other languages for the NER task has been limited.

Arabic is a richly morphological language of complex syntax. The lack of simplicity in the characteristics and specifications of the Arabic language make it a challenging task for NER techniques. Arabic can be classified into three types: Classical Arabic, Modern Standard Arabic and Colloquial Arabic. It is imperative for the task of NER to be able to distinguish between those three types. Classical Arabic is the formal version of Arabic used for over 1,500 years in religious scripts. Modern Standard Arabic is that used in today's newspapers, magazines, books, etc. Colloquial Arabic is the spoken Arabic used by Arabs in their informal day to day speech and differs in dialect for each country and city. There are several specifications of the Arabic language that do not make NER an easy task; lack of capitalization, agglutination, optional short vowels, ambiguity inherent in named entities and lack of uniformity in writing styles. Add to that common spelling mistakes and shortage of technological resources such as tagged corpora and gazetteers, we have several issues to tackle for tasks associated to natural language processing (NLP). The following contains a review and analysis on previous works on Arabic NER.

The impact of using different sets of features in three different machine learning frameworks is investigated by Benajiba et al. [13], for the Arabic NER task. The machine learning frameworks tested are support vector machines (SVM), maximum entropy (ME), and conditional random fields (CRM). Nine different data sets of genres and annotations are explored along with lexical, contextual and morphological features. In order to evaluate robustness to noise of each approach, different feature impacts are measured in isolation and incremental combination. The features that achieved highest impact are: the POS tagger first, and second is the CAP feature, and that is confirmation that the lack of capitalization in languages such as Arabic complicates the NER task considerably. The ME approach is the most sensitive to noise and obtained

significantly lower results than that of the CRF and SVM, specifically when the number of features exceeded six. When the top seven features were used, the SVM approach depicted better performance than that of the CRF approach.

Zitouni et al. [14] present a statistical approach to Arabic mention detection and chaining (MDC) systems. The approach is based on the Maximum Entropy (ME) principle. The system first detects mentions in an input document and then chains the identified mentions into entities. The ME framework allows for a large range of feature types, including lexical, morphological, syntactic and semantic features. The Arabic main-type mention detection system obtained an F-measure of 80.0 in the ACE evaluation. The authors started the system evaluation by only allowing access to the lexical features and gradually increasing features with each increment. A system using only lexical, stem and gazetteer features achieves a measure of 76.5. Syntactic and other classifier feature outputs add more than 3 F-measure points to overall performance (80 vs. 76.5).

We can safely conclude that the use of different combinations of features along with testing of several classifiers may impact the results of an NER system greatly. In another study, Zitouni and Benajiba [15] proposed a semi-supervised approach that utilizes multilingual parallel data (English-Arabic) in an effort to enhance the mention detection (MD) task. The challenge with MD is directly related to the complexity of the morphology of a given language. For this reason, the authors explore the idea of using a MD system designed to suit the needs of a resource rich language (RRL), namely English, to improve the performance of a MD system for a target language (TL), namely Arabic. The impact of features obtained through cross-lingual system proved to be far more effective in enhancing system performance.

Hand-aligned data results: This model achieved a 57.7 F-measure. The low performance is an indication of the level of noise, and we may understand that better performance can be achieved when there is no human annotated data. However, when the Arabic MD system is poor in resources, significant improvements are obtained. When only Lexical features are used, there is an improvement of approximately1.9 points. When the Arabic MD system uses a rich feature-set Syntac, a 1.5 point improvement is obtained. The model based feature (En-Model) depicted the greatest performance (76.22) when the Arabic MD system uses a feature-set that includes more than just the Lexical features. Comb features achieved higher performance than Baseline features (77.18 and 75. 68, respectively).

MADAMIRA is a system designed by Pasha et al. [16] that can be utilized for morphological analysis and disambiguation of Arabic text. It combines aspects of two previously used systems for NLP; MADA [17] and AMIRA [18]. MADAMIRA optimizes both previously mentioned systems with a more streamlined, robust, portable, extensible and faster Java-based implementation. It includes several tasks useful for NLP processes: part-of-speech tagging, tokenized forms of words, diacritization, lemma stemming, base phrases, and NER.

2.2 Audio-based Classification

Audio-only approaches [37] are more commonly utilized than text only approaches for video classification. Audio approaches require fewer computational resources than that of visual methods. When features are stored, they also require less space. Another advantage is that segmented audio clips tend be very short (average 1-2 seconds), so the processing of the audio clips would be easier.

Audio features can lead to three layers of audio understanding: low-level acoustics, such as the average frequency for a frame, midlevel sound objects, such as the audio signature of the sound a ball makes while bouncing, and high-level scene classes, such as background music playing in certain types of video scenes.

Two main techniques are using either time domain features or frequency domain features. Using time domain means plotting amplitude of a signal with respect to time, while frequency domain means plotting amplitude with respect to frequency, which pertains to the spectrum of signal.

The volume standard deviation and volume dynamic range measure may be utilized for time domain features, i.e., sports have a nearly constant level of noise. Different classes of sounds may be categorized by setting certain thresholds. The zero-crossing rate (ZCR) is the number of signal amplitude sign changes per frame. A high ZCR indicates high frequency, i.e., speech has a higher ZCR variability than music does. Silence ratio is the proportion of a frame with amplitude values measured with respect to some threshold, i.e., news has a higher silence ratio than commercials, and speech has a higher silence ratio than music.

Frequency domain suggests an energy (signal) distribution across frequency components. The frequency centroid approximates brightness, and is the midpoint of the spectral energy distribution, i.e., brightness is lower in speech than in music. Bandwidth is the measure of the frequency range of a signal, i.e., speech has lower bandwidth than music. The lowest frequency in a sample is the fundamental frequency, which approximates pitch, and may be used to distinguish between speaker genders, or to identify parts of speech such as introduction of a new topic. A frame that is not silent and does not have a pitch represents noise.

There are several studies that have attempted video classification using only audio signals. Lui et al. [20] used sample audio signals at a specific frequency, and after segmenting and subdividing into overlapping frames, utilized the following audio features: nonsilence ratio, volume standard deviation, volume dynamic range, pitch standard deviation, and others. Results depicted that the features with the highest discriminatory power are frequency centroid,

frequency bandwidth, and energy ratio. Classification was then performed using one-class-one-network structure. The audio samples were then classed into commercial, basketball, football, news report, and weather forecast categories.

Roach and Mason [21] have utilized audio from video for the purpose of genre classification. They used Mel-frequency cepstral coefficients which are coefficients derived from a cepstral representation of an audio clip. This approach was utilized due to its success with speech recognition. The authors find that best results are achieved with 10-12 coefficients. Classification is performed with the Gaussian mixture model due to its effectivity for speaker recognition. The genres studied are fast-moving sports, cartoons, news, commercials and music.

Dinh et al. [22] use a Daubechies 4 wavelet to seven sub-bands of TV show audio clips. Wavelet transforms are useful for reducing dimensionality and have good energy compaction. The audio features used are subband energy, subband variance, zero crossing rate, as well as two customized features; centroid and bandwidth. Classifiers used are the C4.5 decision tree, K nearest neighbor, and support vector machine. Clips of different lengths not higher than 2 seconds were tested, and depicted no significant difference in performance. The genres tested were vocal music shows, news, commercials, cartoons and motor racing sports.

Moncrief et al. [23] utilize audio-based cinematic principles to distinguish between horror and nonhorror films. Variations in energy intensity were used to detect sound energy levels, which in this study are associated with feelings of surprise, alarm, apprehension, surprise followed by alarm, and apprehension progression to climax. These four types of sound were found to be effective to distinguish horror movies and even to distinguish scenes within a horror movie.

2.3 Combination Approaches

Many studies incorporate the use of several combinations [24] of text, audio and visual features in order to complement each technique and overcome weaknesses of each. The main challenge of utilizing features from different modalities is knowing how and when to combine these features [25].

Qi et al. [26] use audio, visual and textual features to classify news streams into genres of news stories. Audio and visual features are utilized to segment and group video shots into scenes. Text processing is used after detection of text through closed captions or scene text detection. Support vector machine classifier is used to classify the news stories.

Jasinschi and Louie [27] classify TV shows using audio, visual and textual features. The audio features are used to classify six categories; noise, speech, music, speech and noise, speech and speech, speech and music. Visual features are utilized to detect commercials. Textual features segment noncommercial parts of the TV program via annotations in closed captions. Finally, all audio categories are combined to classify the TV program as financial news or talk show.

Roach et al. [28] extend on their previous work in [21] which consisted of classifying videos using audio features, to include using visual features. Adapted Gaussian Models for Image Classification (AGMM) is the classifier used for linear combination of the conditional probabilities of visual and audio features. The video classes studied are news, commercial, sports, cartoons and music videos.

Rasheed and Shah [29] utilize cinematic principles with accordance to audio and visual features to classify movies by analyzing the movie previews. Intersection of hue, saturation and value (HSV) color histograms are used to segment previews into shots. Motion per preview is then calculated by using the ratio of moving pixels to total pixels per frame (visual disturbance), for each frame per preview. After visual disturbance is plotted against average shot length, a linear classifier is used to distinguish action and nonaction movies. Then audio energy variation analysis is used to categorize action movies into those with fire or explosions, or without. Light intensity thresholds are used to classify movies as comedy, drama or horror. Horror movies have low levels of light intensity while comedies have the highest light levels, and dramas are in the middle.

3. Proposed Approach

Several studies have addressed the issue of multimodal fusion for video classification [25]. Some chose to combine all the features into a single vector, while others trained classifiers for each modality, combined them at the end, and used another classifier to make a final decision. Identifying which modalities to use in combination, and the technique to combine them is an issue that largely depends on the domain being studied.

In this study, since the domain being tackled is in the Arabic news domain, specifically targeting the shooting and explosion categories, we choose to combine the audio based approach with the textual based approach to improve accuracy of results. As there are Arabic transcription tools, such as the IBM 2011 GALE Arabic speech transcription system presented in [30] that may successfully automatically transcribe Arabic videos in the news domain, we propose to extract the text from speech for the video classification purpose using this tool. We will utilize Named Entity Recognition (NER) tools, such as the MADAMIRA system presented in [16] to process and classify the text retrieved from transcription of speech in Arabic news videos, in order to retrieve elements vital to the classification data we need, namely; location, date, persons involved, events happening, such as the events we aim to classify; involving shootings or explosions. Next, we will complement the results of textual classification of videos by utilizing audio-based classification techniques, for event classification.

Finally, the results from the text based classification and audio based classification are to be combined and another classifier with set weights for each modality will be utilized with a preset threshold to classify data in a precise manner. Results are to be compared with baseline results obtained from large scale experiments such as TRECVid [31].

The output of this system is intended to satisfy the following scenario in Figure 1.

Samar is a journalist. She was tasked with producing a documentary on the "Arab Spring" events 2010-2013 within a maximum period of 2 weeks. To complete her task by the prescribed deadline, she had a quick look at the Web information as well as the YouTube clips relevant to the topic but found them **sparse** and rather **incohesive**, hence not fit for purpose. She then went to Al-Jazeera archives which have all been automatically annotated using the proposed technology. Samar could then find the relevant audio-visual material which has been classified, annotated and categorized based on the audio-visual content of key events in the footages e.g. breaking news, atrocities, key speeches and reporting, interviews with key politicians, major state representatives and peace mediators, etc. Samar then retrieved the necessary content and used it for completing her task by the enforced deadline.

Figure 1. Scenario

4. Video Dataset

The proposed system must be evaluated using baseline results achieved from similar systems that have utilized multimodal video classification techniques to classify Arabic news videos. However, most studies aimed at Arabic video classification have focused on classifying via textual techniques based on retrieving closed caption text or annotations, such as shown in [32,33,34,35] and many others.

The Activ video dataset developed by Zayene et al. [36] has been developed to include 80 videos consisting of more than 850,000 frames, from four different Arabic news channels. However this dataset has been designed specifically to retrieve text from closed captions and assess the performance of text detection, tracking and recognition systems. The data is accompanied by detailed annotations for each text box. While this dataset would be useful for textual-based video classification, for this study it may not serve as an optimal base for evaluation as the objective is to classify raw as well as broadcast news videos utilizing text-from-speech, and audio classification.

As TRECVid has also not released an Arabic news dataset since 2006, and neither has any other source, we propose two main solutions: 1) Design a new Arabic dataset consisting of Arabic news broadcast videos as well as raw news-related videos, to set as a new baseline evaluation tool for future classification attempts. 2) Adopt a mutli-lingual approach, such that we test results using Arabic as well as English Named Entity Recognition for speech-to-text processing, combined with audio-based event classification which is by default language-independent.

5. System Design

The proposed system design is composed of the following components as illustrated in Figure 2:

1. Dataset of videos obtained from: Arabic news channel (i.e., AlJazeera) and social media (raw videos). This dataset is to be designed by the authors specifically to test multimodal video classification systems for the Arabic language. A different dataset is to be used for the English video testing, from the TRECVid database. The domain to be detected is "news" with categories "explosion" or "shooting".
2. Audio Processing component: This step involves the utilization of audio processing tools to classify videos based on frequency domain features. The open source audio feature extraction toolbox, Yaafe [37] is to be utilized for audio classification pertaining to specific noises such as explosions or shootings. The Yaafe toolbox includes several intermediate representations such as spectrum, envelope and autocorrelation. It also includes options for temporal integration. There are several studies which have utilized the Yaafe toolset for audio-based component to classify videos and music[38,39]. There are other audio processing libraries built on the base of Yaafe features such as Essentia [40].

3. The NER processing component consists of utilizing the NER component of the MADAMIRA system, to extract entities based on the speech-to-text conversion obtained from the videos. The entity categories to extract include: event, person, date and location.

4. The multimodal fusion component is responsible for determining the correlation of the data, the confidence levels of the modalities, and the synchronization technique of between different modalities. There are three main types of fusion: feature level, decision level and hybrid multimodal fusion. Feature level fusion refers to extracting features from multimodalities and combining them and sending them as input to a single analysis unit that performs the classification task. Decision level fusion refers to features being processed individually and a local decision being made before combining them to use a decision fusion unit. The hybrid method consists of performing feature level fusion as well as decision level. There are several rule-based methods for multimodal fusion, the most commonly used being linear weighted fusion, specifically using the support vector machine classifier, as it can be easily used to prioritize different modalities. In this study, we will utilize the linear weighted fusion method, to fuse the results of audio-based and NER-based classification.

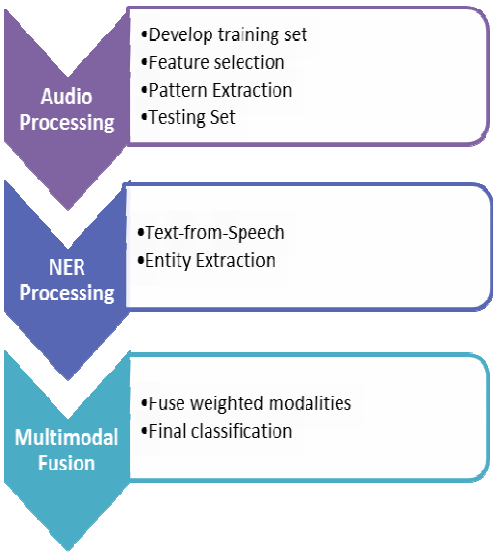

Figure 2. Proposed System Design

5. The final step would be to perform exhaustive evaluation and testing, comparing the effect of using different combinations of features, as well as different classifiers. Testing will be compared on the Arabic dataset as well as the English dataset. Results are to be documented and graphed.

6. Conclusion and Future Work

The amount of Arabic news-related raw videos being released everyday on the internet is increasing, uncategorized and unused. There is a vast need for automated classification of such videos. This system aims to tackle this challenge by proposing a multimodal approach utilizing Arabic NER for processing text retrieved from speech, in order to extract entities related to persons involved, event, location and date. An audio processing component for extracting noise patterns among events like shootings or explosions, is to be used in combination with the NER results. The final step consists of using a multimodal fusion method aimed to achieve optimal results. Issues to consider in our future results pertaining to multimodal fusion include:

- Appropriate synchronization of different modalities
- Optimal weight assignment to different modalities.
- Optimal integration of context into the fusion process.
- Effective utilization of feature vs. decision level correlation
- Optimal modality selection

Acknowledgements

This publication was made possible by GSRA grant # 1-1-1202-13026 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

References

[1] K. Khurana, and M.B. Chandak, Study of Various Video Annotation Techniques, In *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(1), 909-914, 2013.

[2] D. Zhang, M.M. Islam, and G. Lu, A review on automatic image annotation techniques, In *Pattern Recognition*, *45*(1), 346-362, 2012.

[3] P. Thompson, Viewer comments as educational annotation in video content sharing sites, In *International Journal of Social Media and Interactive Learning Environments*, *1*(2), 126-144, 2013.

[4] V. El-Khoury, M. Jergler, D. Coquil, and H. Kosch, Semantic video content annotation at the object level, In *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia* (pp. 179-188). ACM. December 2012.

[5] H. C. Chu, M. Y. Chen, and Y.M. Chen, A semantic-based approach to content abstraction and annotation for content management, In *Expert Systems with Applications*, *36*(2), 2360-2376, 2009.

[6] D. Sánchez, D. Isern, and M. Millan, Content annotation for the semantic web: an automatic web based approach, In *Knowledge and Information Systems*, *27*(3), 393-418, 2011.

[7] A. Jaoua, W. Labda, and J. Alja'am, Automatic Structuring of Arabic and English Search Engines Results Using Concept Analysis, In *International Journal of Computer Science and Engineering in Arabic. Vol. 3, No 01,* 2009.

[8] J. ALJa'am, A. et al., Text Summarization Based on Conceptual Data Classification, In *International Journal of Information Technology and Web Engineering (IJITWE)*, *1*(4), 22-36, 2006.

[9] A. Hasnah, A. Jaoua, and J. Jaam, Conceptual Data Classification: Application for Knowledge Extraction, In *Computer-Aided Intelligent Recognition Techniques and Applications*, 453-467, 2005.

[10] S. Elloumi, J. Jaam, A. Hasnah, A. Jaoua, and I. Nafkha, A multi-level conceptual data reduction approach based on the Lukasiewicz implication, In *Information Sciences*, *163*(4), 253-262.2004.

[11] S. Elloumi, et al., General learning approach for event extraction: Case of management change event, In *Journal of Information Science*, 0165551512464140, 2012.

[12] D. Nadeau, and S. Sekine, A survey of named entity recognition and classification, In *Lingvisticae Investigationes*, *30*(1), 3-26, 2007.

[13] Y. Benajiba, M. Diab, and P. Rosso, Arabic named entity recognition: A feature-driven study., In *Audio, Speech, and Language Processing, IEEE Transactions on*, *17*(5), 926-934, 2009.

[14] I. Zitouni, X. Luo, and R. Florian, A cascaded approach to mention detection and chaining in Arabic, In *Audio, Speech, and Language*

Processing, IEEE Transactions on, 17(5), 935-944, 2009.

[15] I. Zitouni, and Y. Benajiba, Aligned-Parallel-Corpora Based Semi-Supervised Learning for Arabic Mention Detection, In *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(2), 314-324, 2014.

[16] A. Pasha, et al., Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic, *In Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014.

[17] N. Habash, et al., Morphological Analysis and Disambiguation for Dialectal Arabic. *In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, HLT-NAACL* (pp. 426-432), 2013.

[18] M. Diab, Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. *In 2nd International Conference on Arabic Language Resources and Tools,* pp. 285—288, 2009.

[19] M. H. Lee, S. Nepal, and U. Srinivasan, Edge-based semantic classification of sports video sequences, in *Proceedings of the International Conference on Multimedia and Expo,* vol. 2, pp. 157–160, 2003.

[20] Z. Liu, J. Huang, and Y. Wang, Classification of TV programs based on audio information using hidden Markov model, in *Proceedings of the IEEE Multimedia Signal Processing Workshop,* pp. 27–32, 1998.

[21] M. Roach and J. Mason, Classification of video genre using audio, In *Interspeech*, vol. 4, pp. 2693–2696, 2001.

[22] J.-Y. Pan and C. Faloutsos, Videocube: A novel tool for video mining and classification, In *International Conference on Asian Digital Libraries*, pp. 194-205, Singapore, 2002.

[23] S. Moncrieff, S. Venkatesh, and C. Dorai, Horror film genre typing and scene labeling via audio analysis, In *Proceedings of the International Conference on Multimedia and Expo,* vol. 1, pp. 193–196, 2003.

[24] D. Brezeale, and D. J. Cook, Automatic video classification: A survey of the literature, In *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* 38(3), 416-430, 2008.

[25] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, In *Multimedia systems*, 16(6), 345-379, 2010

[26] W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, Integrating visual, audio and text analysis for news video, In *Proceedings of the 7th IEEE International Conference on Image Processing (ICIP)*, pp. 520–523, September 2000.

[27] R. S. Jasinschi and J. Louie, Automatic TV program genre classification based on audio patterns, In *Proceedings of the IEEE 27th Euromicro Conference*, pp. 370–375, 2001.

[28] M. Roach, J. Mason, and L.-Q. Xu, Video genre verification using both acoustic and visual modes, In *International Workshop of Multimedia Signal Processing*, pp. 157–160, 2002.

[29] Z. Rasheed and M. Shah, Movie genre classification by exploiting audiovisual features of previews, In the *IEEE International Conference of Pattern Recognition*, vol. 2, pp. 1086–1089, 2002.

[30] L. Mangu, et al., The IBM 2011 GALE Arabic speech transcription system, In *Automatic Speech Recognition and Understanding (ASRU)*, 2011 pp. 272-277). IEEE, December 2011.

[31] A. F. Smeaton, P. Over, and W. Kraaij, Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International workshop on Multimedia Information Retrieval* (pp. 321-330). ACM, October 2006.

[32] M. Moradi, S. Mozaffari, and A. Orouji, Farsi/Arabic text extraction from video images by corner detection, In *Machine Vision and Image Processing (MVIP), 2010 6th Iranian*. IEEE, 2010.

[33] M. Halima, H. Karray, and A. Alimi, A comprehensive method for Arabic video text detection, localization, extraction and recognition, In *Advances in Multimedia Information Processing-PCM 2010*. Springer Berlin Heidelberg, 648-659, 2010.

[34] A. Anwar, G. Salama, and M. B. Abdelhalim, Video classification and retrieval using arabic closed caption, In *ICIT 2013 The 6th International Conference on Information Technology VIDEO*. 2013.

[35] M. Halima, A. Alimi, and A. Vila, Nf-savo: Neuro-fuzzy System for Arabic Video OCR, In *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 10, pp. 128-136, 2012.

[36] O. Zayene, et al., A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV, In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015.

[37] B. Mathieu, et al., YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software, In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*. 2010.

[38] C. Frisson, et al., Videocycle: user-friendly navigation by similarity in video databases, In *Advances in Multimedia Modeling*. Springer Berlin Heidelberg, pp. 550-553. 2013.

[39] C. Copeland, and S. Mehrotra, Musical Instrument Modeling and Classification.

[40] D. Bogdanov, et al., ESSENTIA: an open-source library for sound and music analysis,

In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013.

Amal Dandashi is a PhD candidate in the Computer Science and Engineering department of Qatar University, Qatar. She previously graduated with a MSc in Software Engineering from the University of Balamand, Lebanon. She is currently an awarded research fellow, working on multimodal video classification. Her previous research interests include biometrics security, face recognition, natural language processing, information retrieval, artificial intelligence, cognitive systems for children with intellectual disabilities, and graph coloring algorithms.

Prof. Jihad Mohamed Alja'am received the Ph.D. degree, MS. degree and BSc degree in computing from Southern University (The National Council for Scientific Research, CNRS), France. He is currently with the Department of Computer Science and Engineering at Qatar University. His current research interests include multimedia, assistive technology, learning systems, human–computer interaction, stochastic algorithms, artificial intelligence, information retrieval, and natural language processing. Dr. Jihad is leading a research team in multimedia and assistive technology and collaborating in the Financial Watch and Intelligent Document Management System for Automatic Writer Identification projects.

Prof. Sebti Foufou obtained a PhD in computer science in 1997 from the University of Claude Bernard Lyon I, France, for a dissertation on parametric surfaces intersections. He is currently with the Department of Computer Science and Engineering at Qatar University. He previously worked with the Computer Science department at the University of Burgundy, France from 1998 to 2009 as associate professor and then as a full professor. His research interests concern geometric modeling and CAD-CAM topics and include: surfaces blending using Dupin cyclides, subdivision surfaces, geometric constraints solving. He also worked in 2005 and 2006, as a temporary guest researcher, at the National Institute of Standards and Technology, Gaithersburg, MD, USA, where he contributed in product engineering related researches: smart machining systems, tolerances, assembly modeling and PLM.