ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 5; Issue 2; March-April: 2017; Page No. 32-40

A Proposed Framework for Knowledge Extraction from Digital Videos

M.E. EL Alami [1], A. F. Elgamal [1], M.Hussien[2]

[1] Professor of Computer Science, Faculty of Specific, Education, Mansoura University, Egypt

Moh_elalmi@mans.edu.eg

[1] Professor of Computer Science, Faculty of Specific, Education, Mansoura University, Egypt

Amany_elgamal@hotmail.com

[2] Lecturers, Computer Teacher Preparation Department, Faculty of Specific Education, Mansoura University, Egypt

Marwahussien@ mans.edu.eg

Abstract

Video contains a huge amount of data which include complex interaction between its elements. Using manual techniques to get video content description is complex process, time consuming and have a lot of limitation as a result of wrong understanding therefore, this paper presents a proposed framework for knowledge extraction from digital video. The framework consists of two phases, the first phase deals with the audio channel and the second phase deals with the video channel. The audio channel passes through three components namely speech recognition, sentence boundary detection and summarization consequently. The second phase also passes through three components namely video segmentation, feature extraction and key frame extraction.

The proposed system was applied on online course from Lynda.com (excel data mining fundamentals).The evaluation of the proposed framework is compared with other systems and shows that the proposed framework is efficient.

Keywords: speech recognition, sentence boundary detection, summarization, machine learning, pause duration

1. Introduction

Artificial intelligence (AI) is typically defined as "the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines" [1]. The different artificial intelligence techniques gives rise to important application area where human processing ability weakness [2]. One of the most exciting recent technologies in artificial intelligence is machine learning. Because it enables the machine to gain human like intelligence without explicit programming [3].Machine learning is the study of computational method for improving performance by mechanizing the acquisition of knowledge through experience. Machine learning aims to provide increasing levels of automation in knowledge process [4]. It emphases how to design new methods to extract knowledge from data which may be meaningless in the real world [5]. Knowledge extraction and representation is known problem in AI –machine learning techniques have been used in different kinds of data [6].The past few years have seen explosive growth in multimedia data such as image, video, audio the new possibilities offered by information highway have made a large amount of video data publicity available [7]. Of all media type's video is the most challenging as it combines all other media information into single data stream [8]. Digital videos collection are growing rapidly in both the professional and consumer environment and are characterized by steal increasing capacity and content variety [9]. Therefore the need for intelligent system to extract knowledge from any video to provide rapid browsing, retrieve the relevant video, summarization of video, providing the user with a quick idea about the content. Examine the most important or useful video in flexible and efficient way has become challenge task. Many intelligent related systems have been implemented in this field, Edward Jorge. Presents proposed method using local descriptors, temporal video segmentation and visual words to extract the semantic information expressed by the video's visual entities to obtain video summarization that can produce meaningful and informative video

Corresponding author: M.E. EL Alami

summaries[10].Youness TABII Developed novel algorithms for sports video processing and analysis for segmentation of video into shots and classification of these shots in the case of soccer video and provide a new algorithm for score box detection in soccer match video based on motion vector computation to generate summaries and highlights[11].Danila Potapov ,et al developed a novel approach to produce higher quality video summaries based on category-specification that delivers short and highly-informative summaries, First efficiently performs an automatic kernel-based temporal segmentation then, equipped with an Support Victor Machine (SVM) classifier for importance scoring that was trained on videos for the category at hand and score each segment in terms of the importance. Finally, produce a video summary composed of the segments with the highest predicted importance scores [12]. Omar U. Florez proposed novel algorithm for automatic and semi-automatic analysis of activities in video, scene understanding based on interactions between activities, and the predicting of labels for new scenes [13]. Y. Gao, et al proposed a New algorithm to generate video summarization based on two-level redundancy detection first video segment into shots using color histogram and optical-flow motion features, the cast indexing procedure is employed to generate the storyboards of cast in the video. Then similar key frames are removed using HAC in each scene [14].

In this paper a proposed framework is presented to automatic extract knowledge from digital video. We depend on Speech recognition to recognize speech into text. We depend also on extractive summarization to present the most important information in a shorter version of the original text while keeping its main content and help the user to quickly understand large volumes of information [15]. Extractive text summarization requires high accuracy sentence boundary detection methods because even a single incorrectly separated incomplete sentence will greatly reduce the coherence and readability of a summary. Previous studies on sentence segmentation suggest that pause duration is an effective sentence boundary indicator [16]. Pauses in speech are extremely complex phenomena may have physiological functions (Breathing, swallowing) [17]. So we depend also on linguistic approach. To extract knowledge from video channel we depend on the extraction of key frames. Extracting a small number of key frames that can abstract the content of video is very important to enable a quick browsing of a large collection of video data and to achieve efficient content representation and access [18].

2. The Proposed Framework for Knowledge Extraction

We construct the knowledge extraction framework for automatically extracting knowledge from video based on two phases of video (audio, visual). The audio phase comprises several tasks, such as speech recognition, sentence boundary detection and summarization while the visual phase comprises several tasks, such as video segmentation, feature extraction and key frame extraction. The proposed framework exemplified in Figure 1

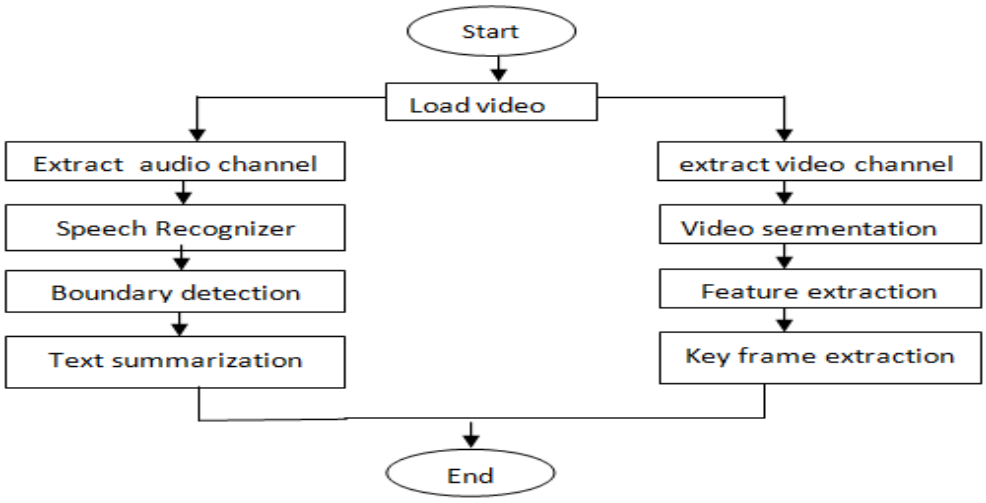

Figure 1: a general overview of the proposed framework

2.1 Extract Knowledge from Audio Channel

In order to extract knowledge from audio channel we need to convert speech signals into a sequence of words that is meaningful for users. The audio signal is transcribed using an Automatic Speech Recognition then Automatic summarization is used to extract the most relevant information as follow:

A. Speech Recognition

We have used Microsoft speech SDK (system development kit) version 5.1 tool. It's one of many tools that enable developer to speech and recognize spoken words. The tool is compatible with number of programming languages such as c# .The proposed system is based on speaker- independent automatic speech recognition.

B. Sentence Boundary Detection (SBD)

The output of automatic speech recognition generates a stream of words which pauses between sentences are not found. SBD is done using combination of linguistic and acoustic approaches in order to increase the performance. We apply acoustic approach based on pause duration. In speech technology, information on pauses is used in majority of algorithms of automatic punctuation detection .It has been shown that 95 % of silent pauses longer than (350 ms) are the sentence boundaries [19]. Therefore, the duration that is shorter than (350 ms) are not considered as sentence boundaries. The first step is split the input audio file into a group of audio files; each audio is recognizing to represent a sentence. A linguistic Approach based on supervised machine learning is applied to make the machine determine the full stop according to the results of pause duration. This process for sentence boundary detection goes through many stages, the first stage is tokenization. We implemented tokenization to separate the input document into individual words. The second stage is part of speech tagging (POS).Every word in sentence is passed to the (POS) tagger to assign grammatical tags to each word such as noun, verb, adjective, etc in order to recognize all the words in the sentence. The third stage is parsing, the task of natural language parser is to take sentence as input and return syntactic representation. Stanford parser is used, once we parsed the sentence, the parser extracts the relations and build parsing tree for each sentence. A parser is trained on a training set to generate a machine learning model. To construct the training dataset we apply the following algorithm:

```
For All sentence, Do
        Apply segmentation and tokenization
        Call the part of speech tagger
        Return syntactic representation
        Extract (EOF) and (BOS) features
IF
        Extracted features are matching with existing sentence features (EOF, BOS) in dataset.
THEN
        Ignore the sentence to prevent the duplication
ELSE
        Update dataset with the new record
```

Our model is constructed based on maximum entropy (Maxent) classifier to test the input sentence. The model learns the contextual features from annotated training dataset and classifies each occurrence of the features End-of Sentence (EOF) or Beginning- of Sentence (BOF).The classification categories are: a valid or invalid sentence boundary using the learned model from the annotated boundaries dataset, the classification algorithm are explained as follows:

```
Initialization:
Set     current sentence    = second sentence.
Set     previous sentence = first sentence.

For all sentences, do

IF
        The begin of current sentence (BOS) is valid AND the end of the previous sentence (EOF) is valid

Then
        It's a right sentence boundary.

ELSE
        Remove full stop from the previous sentence and merge previous sentence with current sentence.

End if
        Set Previous sentence=current sentence.

Loop
```

The model is trained on a collected small test corpus from various websites. The size of the dataset is containing a total of 1000 sentences in different domains. Which were manually annotated with Delimiter full stop (.) Different cases for abbreviations, proposition and Possessive pronouns are placed in the dataset. The performance can be also be improved by training on a bigger corpus. When machine learning is used, the performance increases and we gain correct sentence boundaries.

C. Summarization

After the detection of correct sentences boundaries text summarization was applied. Extractive text summarization process goes through many stages, the first stage is stop words removal .Stop words are the words which appear frequently in document but provide less meaning in identifying the important content such as 'a',' an', 'the', etc .Commonly available stop word lists of about 400 words are used and saved in a file. Stop words are used to eliminate stop words from original text through string comparison method . The second stage is term frequency to find most important words which are repeated multiple .Identify these words is the key to summarize the text .We considered each word as a term and calculate their weight that depends on the number of occurrences of the term in text . The list of terms is stored in descending order. The weight of each term is calculated as follows:

term weight =frequency of the term/Total no. of terms in the text

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Which $N_{i,j}$ is total number of occurrence of the term in the text.

and the denominator $N_{k,j}$ is the total number of occurrences of all terms in text[20].The third stage is Sentence vectorization which used to convert every sentence into vector based on words histogram to rank the individual sentence according to the weight. The weight of the sentence can be calculated by adding the weight of all the terms in the sentence and divided by total number of terms in that sentence(n) as follows:

$$wt_s = \sum_{i=1}^{n} (wt_i)/n$$

Where wt_s = weight of the sentence; wt_1, wt_2, wt_3 ...wt_n are the weights of individual terms in the sentence; n = total number of terms in that sentence [21].

Sentence extraction is based on the sentence weights. Sentences contain higher frequent terms are considered important sentence. All the important sentence are extracted in the order which they appear in the original text.

2.2 Extract knowledge from visual channel

Represents a short summary of original video to give to the user a synthetic and useful visual abstract of video sequence is essential part in video summarization. To extract valid information we first segment the video into frames, then extracted the features for each frame ,finally key frames are elicited .

A. Video segı ..... (1)

In this step, the video sequence was split into multiple shots. The frames in the same shot are very much similar to each other therefore; the frames that reflect the best shot contents are elicited.

B. key frame extraction

The collection of extracted key frame from the original video based on predefined technique using visual features such as color histogram difference and correlation[22]. The first frame is declared as a reference key frame then the color histogram difference and correlation is computed between the current frame and the last extracted key frame. If the frame difference based on the extracted features between the consecutive frames exceeds predefined threshold the current frame is eliciting as key frame. This process is repeated for all frames in the video .The next section we will present an illustrative example showing the proposed framework steps.

3. Illustrative example

The proposed framework is developed using Microsoft visual c#2010 and MATLAB R2012. The proposed system has been applied in online course from Lynda.com (excel data mining fundamentals) as shown in figure 2. The user can upload any video to start splitting it into video channel and audio channel.by clicking select audio folder button the user can choose a folder to save the audio file and by clicking select video folder the video frames are saved.

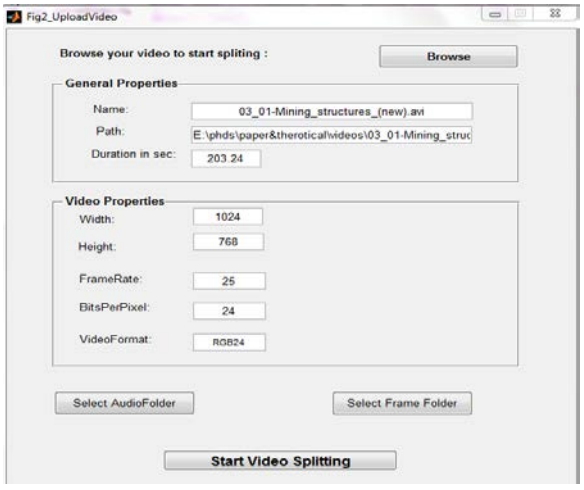

Figure 2: video splitting

Silence detection is done according to fast Fourier transform (FFT) wave as shown in figure 3 then we estimate the length of the pauses between sentences through silence detector .The duration that is shorter than (350 ms) are not considered as sentence boundaries so the audio file is divided into a group of audio files based on long pauses. The results are shown in figure 3 and 4.

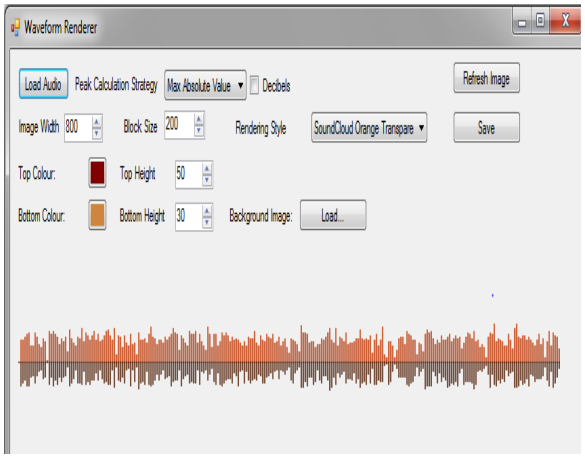

Figure 3: wave form render

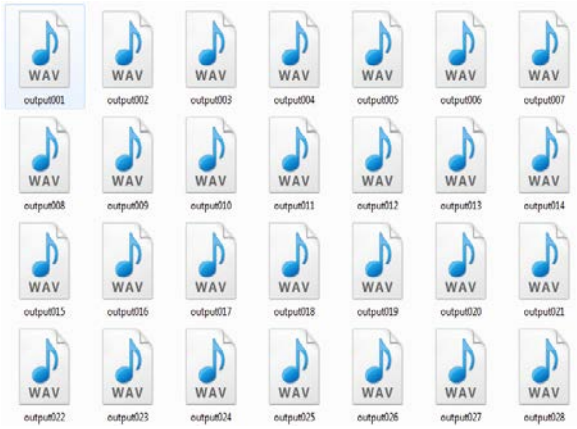

Figure 4: the output wav files

A new instance of speech recognition engine for each audio from the resulted wav file is created and compare with the instance of dictation grammar class, the output is a text file which can be saved as shown in figure 5.

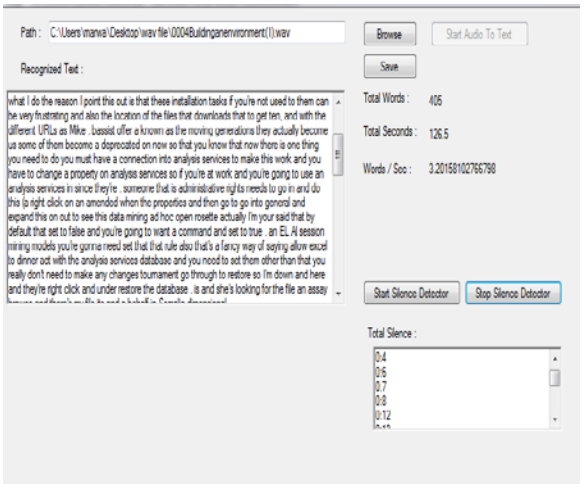

Figure 5: silence detector

To apply linguistic approach, the tokenization process is implemented into each sentence of the entered text.For example: "Data mining techniques are the result of a long process of research and product development" is converted to tokens as follows: [data] [mining] [techniques] [are] [the] [result] [of] [a] [long] [process] [of] [research] [and] [product] [development].Then we apply part of speech tagging (POS) for the previous example the sentence is converted to

POS Tagging: (NNS data) (NN mining) (NNS techniques) (VBP are) (DT the) (NN result) (IN of) (DT a) (JJ long) (NN process) (IN of) (NN research) (CC and) (NN product) (NN development). A sample of tag sets is shown in table (2).

Table 1: sample of used tag sets

word	Tag	description
data	NN	Noun, singular or mass
mining	NN	Noun, singular or mass
techniques	NNS	Noun, plural
are	VBP	Verb, non-3rd person singular present
the	DT	Determiner
result	NN	Noun, singular or mass
of	IN	Preposition or subordinating conjunction
a	DT	Determiner
long	JJ	Adjective
process	NN	Noun, singular or mass
of	IN	Preposition or subordinating conjunction
research	NN	Noun, singular or mass
and	CC	Coordinating conjunction
product	NN	Noun, singular or mass
developme nt	NN	Noun, singular or mass

The parser extracts the relations and build parsing tree for each sentence as shown in figure 6. The previous example sentence is converted to:-
(TOP (S (NP (NNS Data) (NN mining) (NNS techniques)) (VP (VBP are) (NP (NP (DT the) (NN result)) (PP (IN of) (NP (NP (DT a) (JJ long) (NN process)) (PP (IN of) (NP (NN research) (CC and) (NN product) (NN development))))))) (. .)))

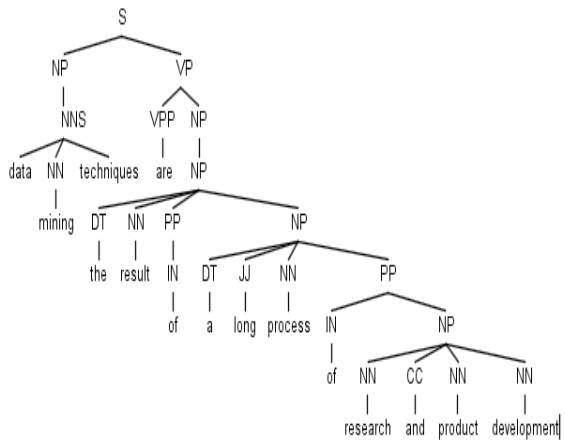

Figure 6: parsing tree

Applying the sentence boundaries detection algorithm to improve the performance shown in figure 7 which illustrates The GUI of text before and after enhancement .

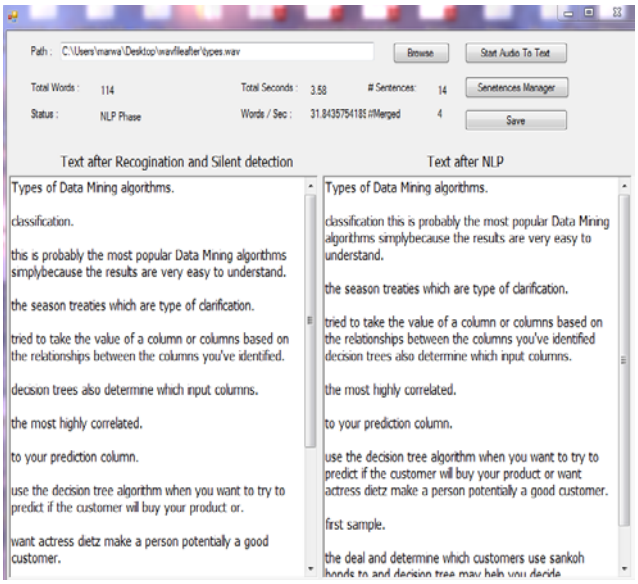

Figure 2: GUI of sentence boundary detection

To apply summarization each word in the text is represented as a term and calculates its weight, then the terms are ranked in descending order as shown in table 2.

Table 2: term weight

word	frequency	Relative frequency
'data'	74	0.57%
'Mining'	45	0.35%
'techniques'	38	0.29%
'development'	8	0.06%

The first column contains the most frequently used words in the text. The second column contains the frequency of the word (the number of times that word appeared in the document). The last column contains the relative frequency of the word, which is the frequency of the word divided by the total number of words in the document. Then the vectorization was applied for each sentence for examples if we have keywords [‘data’ ‘mining’ ‘techniques’ ‘development’] then the sentence "Data mining techniques are the result of a long process of research and product development “victories to [0.57 0.35 0.29 0.06] then the score of each sentence is calculated and ranked in descending order as shown in table (3) The Sentences with the highest rank are elicited and included in summary.

Table 3: sentence weight

Sentence NO	Sentence score
S1	1.25
S2	1.20
S3	0.93
S4	0.84
S5	0.71

The user can specify the length of summary by input the parentage of original file ten select start summarizer as shown in figure 8.the output is a text file.

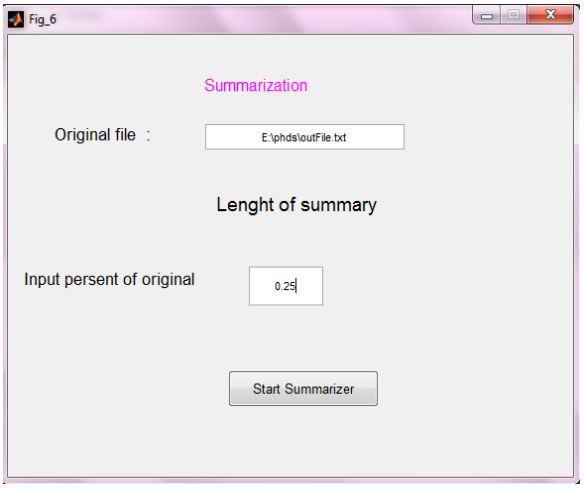

Figure 8: summarization

To extracted knowledge from video channel color histogram difference and correlation difference measure are calculated, then a threshold value is determined through experimental test. The best suitable value closest to manual execrated key frames is equal to 0.6.The two frames is different if the difference measure between them is above the threshold. A sample of extracted key frames is shown in figure 9.

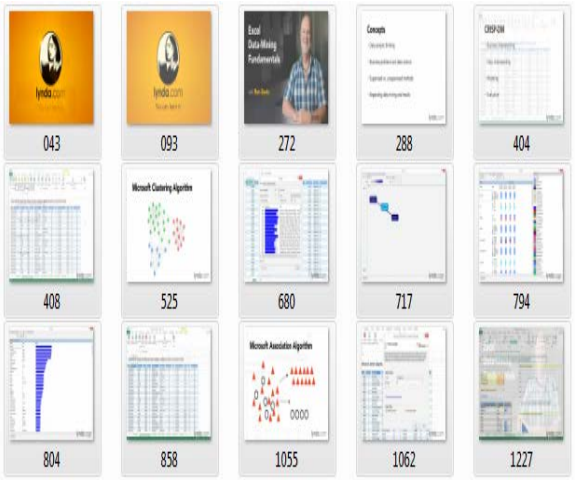

Figure 9: a sample of the extracted Key frames

4. Application and results

Most of related research deals with video channel however a few research deals with audio channel. Integration of knowledge extraction from both channels of video can provide more accurate knowledge. So we keep our comparison focus on each phase of video. To evaluate video channel we use the video ‘‘A New Horizon, segment 02”, selected from the Open Video Project (OV) [23].We depend on the ‘‘Comparison of User Summaries’’ (CUS). It’s a comparison strategy proposed by Avila et al [24] designed to evaluate the quality of the summaries key frames. The summaries generated by automated methods are compared with the summaries created by human users. Each key frame from the automatic summary is compared with the frames of users’ summaries. The number of matching and non-matching key frames of automatic summaries and user summaries are used to compute the Accuracy Rate (CUSA) and Error Rate (CUSE) which are defined as follows:

CUSA = nmAS/nUS …… (4)

CUSE = nm’AS/ nUS …… (5)

Where nmAS = the number of matching key frames from automatic summary (AS), nm’AS = the number of nonmatching key frames from AS and nUS = the number of key frames from user summary (US) [25].

Table 4 show the comparison of the results of our technique with OV [26], DT [27] and VISTO [28] based on Accuracy and Error Rates metrics.

Table 4: (CUSA) & (CUSE) by various techniques for the video "A New Horizon, segment 02"

Measure	OV	DT	VISTO	Proposed framework
CUSA	0.33	0.15	0.38	0.33
CUSE	0.09	0.18	0.12	0.14

The result shows that the proposed system is efficient comparing with other systems and the results are closest to the user summary prediction. The performance measures used for the evaluation of audio channel are precision recall, and F-score, as shown in equation (6), equation (7) and equation (8) respectively.

$$precision = \frac{\text{sum manual} \cap \text{sum automated}}{sum\ automated} \dots\dots\dots\dots\dots\dots (6)$$

$$recall = \frac{\text{sum manual} \cap \text{sum automated}}{sum\ manual} \dots\dots\dots\dots\dots\dots (7)$$

$$-measure = 2 * \frac{\text{precision} * \text{recall}}{precision + recall} \dots\dots\dots\dots\dots\dots\dots (8)$$

Where Sum manual ∩ Sum automated is the set of sentences selected by both automated summarizer and manual summarizer ; Sum manual is the set of sentences selected by manual summarizer ;Sum automated are the sentences selected by the automated summarizer[29].

A collection of six documents is prepared and then a comparison between the most commercial summarizer software (MS-word, Text compactor and Intellexer) was applied with reduction ratio of 25%.The results are shown in table (5).

Table (5) the comparison of average results for the reduction ratio set to 25%

SUMMARIZER	Average		
	precision	recall	F-measure
MS-Word-2007	0.46	0.40	0.43
Text-compactor	0.46	0.44	0.45
Intellexer	0.47	0.42	0.44
Proposed system	0.49	0.44	0.45

To compare our proposed summarizer a manual and automated summary are executed according to our software .Our proposed system reaches the average precision of 0.49, recall of 0.44 and f-measure of 0.45. the results have shown that our system has better performance in comparison with MS-word 2007 and Intellexer summarizers and equal to text compactor summarizer.

Conclusion

In this paper, a proposed framework for knowledge extraction from digital videos is presented. The framework extract knowledge from the two video channels (audio, video) .The proposed framework takes into account several problems such as independent speech recognition, sentence boundary detection and identify the value of the threshold for key frame extraction in order to improve the performance. The proposed framework in online course from Lynda.com (excel data mining fundamentals) was applied. The proposed framework can extract knowledge from any other video in different domains. We evaluated and compared our results with different techniques. The experimental result achieves reasonable accuracy in both channels from digital video file.

REFERENCES:

1. JoséHernández-Oralloa , FernandoMartínez Plumeda, UteSchmidb, "Computer models solving intelligence test problems: Progress and implications" journal of Artificial Intelligence" ,pp74:107, 2016, available at www.Science Direct.com.
2. Swapnil Ramesh Kumbhar, "An Overview on Use of Artificial Intelligence Techniques in Effective Security Management " ,International Journal of Innovative Research in Computer and Communication Engineering , Vol(2), No(9) , September 2014.
3. Sumit Das , Aritra Dey ,Akash Pal "Applications of Artificial Intelligence in Machine Learning: Review and Prospect" International Journal of Computer Applications, Vol(115) , No. 9, April 2015.
4. nesma ibrahem "probabilistic machine learning with expert system to disease diagnosis" ,thesis for the master degree on computer science, Mansoura University 2011.
5. Tarek Sobh,"Innovations and Advances in Computer Sciences and Engineering" Publisher: Springer, 2010.
6. Lemonia Ragia, Vladimir Berenzon," Rules Extraction and Representation for Geographic Information Systems", 10th AGILE International Conference on Geographic Information Science, 2007.

7. Deshmukh Bhagyashri," REVIEW ON CONTENT BASED VIDEO LECTURE RETRIEVAL", International Journal of Research in Engineering and Technology, volume(3),2014, Available @ http://www.ijret.org.
8. Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, et al, "A Unified Framework for Video Summarization, Browsing & Retrieval", Elsevier academic press, 2006.
9. Alan Hanjalic, Li-Qun Xu, "Affective Video Content Representation and Modeling", journal of IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 7, NO.(1), 2005.
10. Edward Jorge," A New Method for Static Video Summarization Using Visual Words and Video Temporal Segmentation" MASTER thesis, in Computer Science, University of Ouro Preto , Brazil,2013.
11. Youness TABII," Sports Video Analysis", PhD thesis, in computer science, University of Mohammed V Soussi , RABAT,2014.
12. Danila Potapov,et al ," Category-specific video summarization" European Conference on Computer Vision, Sep 2014, Zurich, Switzerland. Springer,2014.
13. Omar U. Florez ," Knowledge Extraction in Video Through the Interaction Analysis of Activities" PhD thesis, department of computer science , Utah State University, 2013.
14. Y. Gao, et al , "Dynamic video summarization using two-level redundancy detection," international journal of Multimedia Tools and Applications, vol(42) , pp. 233–250, 2009.
15. Marcin Miłkowski, Jarosław Lipski "Using SRX standard for sentence segmentation in Language Tool", Proceedings of the 4th conference on Human language technology,2009.
16. Lei Xie, Chenglin Xu and Xiaoxuan Wang, "PROSODY-BASED SENTENCE BOUNDARY DETECTION IN CHINESE BROADCAST NEWS",Chinese 8th international symposium spoken language processing,kowloon,2012,IEEE.
17. Marcin Miłkowski, Jarosław Lipski "Using SRX standard for sentence segmentation in Language Tool", Proceedings of the 4th conference on Human language technology, 2009.
18. Shruti V Kamath, Mayank Darbari, Rajashree Shettar , "CONTENT BASED INDEXING AND RETRIEVAL FROM VEHICLE SURVEILLANCE VIDEOS USING GAUSSIAN MIXTURE MODEL", International Journal of Computer Engineering and Technology , Vol(4), No(1),2013.
19. Igras-Cybulska et al," Structure of pauses in speech in the context of speaker verification and classification of speech type", EURASIP Journal on Audio, Speech, and Music Processing, , Vol(16),No(2),2016.
20. Muhammad Mahbubur," Intellectual Knowledge Extraction from Online Social Data"International Conference on Informatics, Electronics & Vision, IEEE,2012.
21. R.C. Balabantaray," Text Summarization using Term Weights" International Journal of Computer Applications, Vol(38),No(1), 2012.
22. B. F. Momin1, S. B. Pawar ," Key Frame Extraction Using Features Aggregation", International Journal of Recent Development in Engineering, Vol (2), No(1), 2014.
23. http://www.open-video.org
24. Avila, et al "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method", Pattern Recognition Letters, Vol (32), No(1), pp. 56–68, 2011.
25. Chinh Dang , Hayder Radha,"Key Frame Extraction for Video using Robust Principal Component Analysis", journal of Transactions on Image Processing,Vol(24),No(11),IEEE, 2014.
26. D. DeMenthon, V. Kobla, D. Doermann, "Video summarization by curve Simplification", in: Proceedings of the ACM International Conference on Multimedia, New York, USA, 1998, pp. 211–218.
27. P. Mundur, Y. Rao, Y. Yesha, "Key frame-based video summarization using delaunay clustering", International Journal on Digital Libraries Vol (6),No (2) ,pp. 219–232, 2006.
28. M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "VISTO: visual storyboard for web video browsing," in Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR), pp. 635–642, 2007.
29. Sherif Elfayoumy, Jenny Thoppil ,"A Survey of Unstructured Text Summarization Techniques" International Journal of Advanced Computer Science and Applications, Vol. 5, No. 4, 2014.