ISSN 1870-4069

Audio Signal Analysis for Indexing and Rating Movies

Abdullah[1,2], Nida Hafeez[1,2], Muhammad Ateeb Ather[2], José Luis Oropeza-Rodríguez[1], Alexander Gelbukh[1]

[1] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Mexico City, Mexico

[2] Department of Computer Sciences, Bahria University, Lahore, Pakistan

abdullah2025@cic.ipn.mx, nhafeez2024@cic.ipn.mx, joropeza@cic.ipn.mx, gelbukh@cic.ipn.mx, 03-134211-022@student.bahria.edu.pk

Abstract. Identification of the contents of a video sequence is a major factor in determining the context of the video. Video classification has been a popular research area in the fields of Computer Vision, Artificial Intelligence and Natural Language Processing for long now. Classifying videos on the basis of factors like video frames, audio signals, static images and similar key factors have been worked upon. So far, techniques like Hidden Markov Models, Dynamic Time Wrapping and Neural Networks have been incorporated for content-context based classifications of videos. This paper discusses social and technical effects of a specific languages being spoken in a video and focuses on determining the suitable audience for that video sequence i.e. a movie. We have proposed a system that helps classifying the context on the basis of audio track of the video. Natural language processing plays the primary role in the proposed system. Other techniques have also been used according to the different requirements of the system.

Keywords: Natural language, speech recognition, speech synthesis context based classification, natural language processing

1 Introduction

A video sequence is a rich source of information that consists of objects, motion (movement), speech (dialogues), and text (captions), colors (B&W, Greyscale, and Colored) and images. Humans can easily and quickly interpret both the semantic and syntactic context of a video sequence from the information being provided by the video itself. With the advancements in technology, internet is available to anyone, everyone, anywhere and everywhere. With this comes a pressing need for context based

classification since the internet is accessible by children too. There is a need for more efficient and effective tools and methods for dissemination of visual and audible content available on the internet. This simply states that multimedia content needs to be indexed, classified and stored according to their context and be accessible for suitable audience only. This first and foremost requirement for context-based indexing is understanding and interpreting the context before further processing. Other processes include sorting, calculating, storing etc. the subject according to the context.

The key factor in understanding the content and context of a video scene segmentation and audio track of the given sequence. Research has focused on use of information provided by the speech and images. The most common techniques we come across include video segmentation, video frames, static images, identifying objects have been used to interpret the context of a video sequence in terms of visual features. In addition, several researchers have worked on analyzing audio signal for context-based classification. This theory is feasible because audio tracks in different environments are fairly differentiable. For example, the audio of a film is different from that of a documentary.

Although this thought may pop in mind that audio signal alone may not be enough for indexing and categorization of a video and that video analysis would be necessary. However, in this paper, we will be discussing audio-based analysis. Because it's significantly less complex and uses simpler computations. We also propose a technique that will perform the audio analysis of movies and then categorize and rate them according to appropriate audience.

The major highlight of our technique is that it uses Natural Language Processing and the dataset is updateable. Our technique comprises of three steps; first we extract the audio track of any video movie; the system then analyzes and compares the audio, converts it into text using speech recognition and is fed to the system; the system compares the text with the predefined data set; on the basis of the results obtained, simple mathematical calculations are performed and the movie can be rated for different audiences.

2 Motivation

Since the last few decades, there has been a lot of advancement and development in the technological sector. In the late 1990s, internet was released commercially and made available to everyone. Businessmen, Educators, Students all have now easy access to the internet via WIFIs, Edge and now recently released 3G in their smartphones, tablets, laptops and PDAs. Internet plays a major role in everyone's daily life. Whether its social media, news, watching programs online, viewing top charts or looking for a recipe, internet has makes everything you ask for available in the matter of a second. Young or old, everyone needs internet nowadays. But with pros comes cons. Not everything is suitable for everyone. The context of content can vary from one age group to another. What might be appropriate for one age group may not be suitable for the other. In this paper, focus has been laid on classifying movies according to different

age groups. These age groups are classified as PG-13, PG-18, and Rated-R in terms of classification with respect to the context. Following this, audience can review the rating and therefore interpret whether the movie is suitable or them or not.

3 Related Work

In the recent past, researchers have focused towards investigating the potential of analyzing an audio signal for video classification [5-6]. Saunders proposed a method for separating speech from music [9]. Nam and Tewfik presented to detect sharp temporal variations [8]. Lie et al. proposed a method to detect change in the feature vectors [6].

Saraceno and Leonardi presented a technique for differentiating silence, speech music and voice clips from an audio sequence [5] and so did Pfeiffer [7]. Lie Lu presented a theory for audio content analysis on the basis of nature of the audio using an algorithm based on the KNN method [2].

Wold, Blum and Wheaton proposed a method to classify audio content on the basis of its characteristics such as pitch, sharpness, beat and rhythm [3].

Mahedero, Martinez and Cano have proposed a theory on the natural language processing of music lyrics using language identification, structure extraction and thematic analysis [16].

Gunsel and Tekalp have discussed automatic scene change detection and key-frame selection for content-based video abstraction using automatic threshold selection techniques [17].

Ferman, Tekalp, Mehrotra have proposed a methodology for effective video content representation using temporal segmentation and domain-specific implementation such a sports, dramas/movie and news etc. [18].

Tsekeridou and Pitas have done a research on audio-visual analysis that analyzes audio and video information and interprets their relationship [19].

Liu et al. used automatic audio classification and speaker identification techniques for content-based video analysis and classification [20].

Chunneng Huang et al. proposed a framework that uses text-based content classification of videos on online video sharing sites using user-generated data. Huang et al. discussed about recent changes in performance of speech recognition systems for larger data sets. Reduction in word error rate and improvements were observed for a large corpus [22].

Wei Jiang et al. investigated the incorporation of Short-term Audio Visual Atoms for video concept classification [23]. Yuichi Nakamura and Takeo Kanade highlighted the association between the images clues and language clues to conclude corresponding video segments using natural language processing and scene segmentation [24].

Finally David J. White et al. studied the role video recognition in behavioral research and developed a system that uses voice recognition to collect and accumulate data [25].

4 Types of Speech Recognition Systems

Several speech recognition systems available on the market. A strong speech recognition system can trace over thousand words. These systems usually require training before they can be put to practical and professional use. Normally, we come across two types of speech recognition systems. Speaker-Dependent and Speaker-Independent. The third type, a more recent and therefore a less common type, known as Speaker-Adaptive.

4.1 Speaker Dependent

This system works by learning characteristics, unique attributes of the user like his voice, his accent, etc. so that the system becomes accustomed to a specific person. Users train the system by speaking to the system so that it can analyze how the user talks. Usually the user has to read a few texts pages to the system before they can use it. Speaker-dependent systems are usually easy to develop, cheap and more efficient to use, but not as dynamic as speaker independent or speaker adaptive systems.

4.2 Speaker-Independent

A speaker-independent system is developed to recognize anyone's voice and is operable for any number of speakers (of a particular dialect e.g. American English). Since it has been designed to interpret anyone's voice, it doesn't require any prior training like the speaker-dependent system. Such systems are more feasible to use in interactive environments where users can just use the system without having to read any text to the system to train it.

Having said that, speaker-independent systems in interactive environments where users can just use the system without having to read any text to the system to train it. Having said that, speaker-independent systems in contrast to speaker dependent systems are less accurate and not as efficient.

4.3 Speaker-Adaptive

Another trend in speech recognition systems is emerging nowadays, known by the name of speaker-adaptive systems. The unique feature these systems offer is that these system start off as speaker-independent systems. And later on adjust themselves according to individuals in a brief training period.

5 Speech Recognition Input Types

Different speech recognition systems have different input mechanisms. On the basis of the input types, speech recognition systems are distinguished as follow.

5.1 Isolated Word Recognition

As the name suggests, this input type system requires the user to input one utterance at a single time. This means that there needs to have quite at the beginning and after termination. It may seem that the system accepts only one word inputs but this is not the case. Rather the system requires pauses in between so that it can process the input during these pauses. Such systems have "Listened/Not Listened" states that define if the system heard the input word clearly or not.

5.2 Connected Word Recognition

Similar to Isolated Word Recognition systems, Connected Word Recognition systems require accept a combination of words to be input once at a time with a short pause in between inputting each chunk.

5.3 Continuous Speech Recognition

Continuous speech recognizers are more advanced form of speech recognition systems. That is why they are one of the more difficult to develop because they require special methods to interpret the input signal. These systems allow the user to speak naturally while the system process the content.

5.4 Spontaneous Speech Recognition

There is a number of explanations as to what is spontaneous speech. Basically, spontaneity of a speech is determined by its naturalness. A speech recognition system with spontaneous speech recognition has the ability to interpret and process natural speech expressions such as "ums" and "ahs" or other similar expressions.

6 Mechanism

The microphone converts the digital signal into an analog on. This is done by the sound card present in the computer. The user inputs the signal, also known as utterance which is a binary sequence consisting of 1's and 0's which also comprise programming languages. Sound-recognition systems comprises of two models, acoustic model and language model.

In the first step, the acoustic model (An acoustic model takes audio recordings and text descriptions to form statistical representation of the sound of each word.) which breaks down the voice signal into speech elements known as phonemes. More recent versions of speech recognition systems have matured so that now they can eliminate noise and other distracting factors in speech recognition process.

After this, in the second step (language model), the input signal is compared to the data set, the "digital dictionary", fed to the computer. This dictionary contains over

100,000 words, a fairly large collection. If the system finds a match from the digital dictionary, it is displayed on the screen.

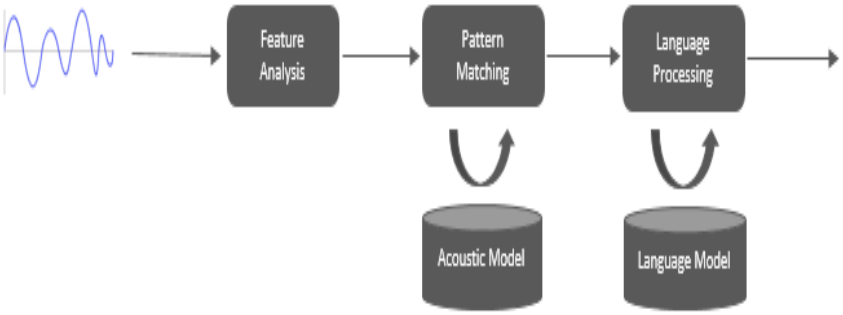

Fig. 1. General processing pipeline.

7 Algorithms used for Speech Recognition

Language modelling is an important part of modern speech recognition algorithms.

7.1 Hidden Markov Models (HMMs)

In HMMs, the state of the system is not visible but the output is visible unlike simple Markov Models in which the states are visible. Using HMMs a speech signal can be viewed as a stationary signal in piecewise. On a short time-scale, speech can be regarded as a stationary process and therefore can be regarded as a Markov Model.

Nowadays general purpose speech recognition systems are based on HMMs. Hidden Markov Models are statistical models that output the given input string in a sequence.

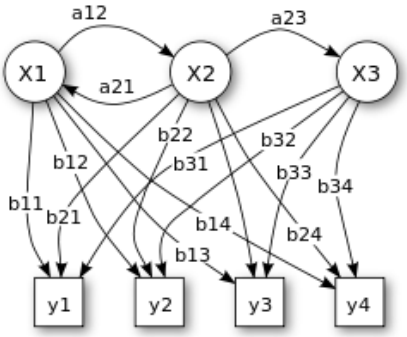

Fig. 2. HMM example.

The system being used is considered to be a Markov method with hidden states. Another reason why HMMs are popular is that they simple, computationally feasible

to incorporate. HMMs are especially recognized for their use in speech, gesture sensors, handwriting recognition, P-O-S tagging and bio-informatics.

In the recent past, HMMs have been generalized to duplets and triplets models to allow more complex data structures to be considered.

7.2 Dynamic Time Wrapping

Dynamic Time Wrapping is used for measuring similarity between two varying temporal sequences. They may vary in time or speed. Dynamic Time wrapping can be applied on any data that can be converted into a linear sequence. A popular application of DWT is speech recognition to deal with different speaking speeds.

Dynamic Time Wrapping based speech recognition systems were traditionally used but have now largely been replaced by Hidden Markov Models. Dynamic Time Wrapping is used for measuring similarity between two input sequences that may vary in time or speed.

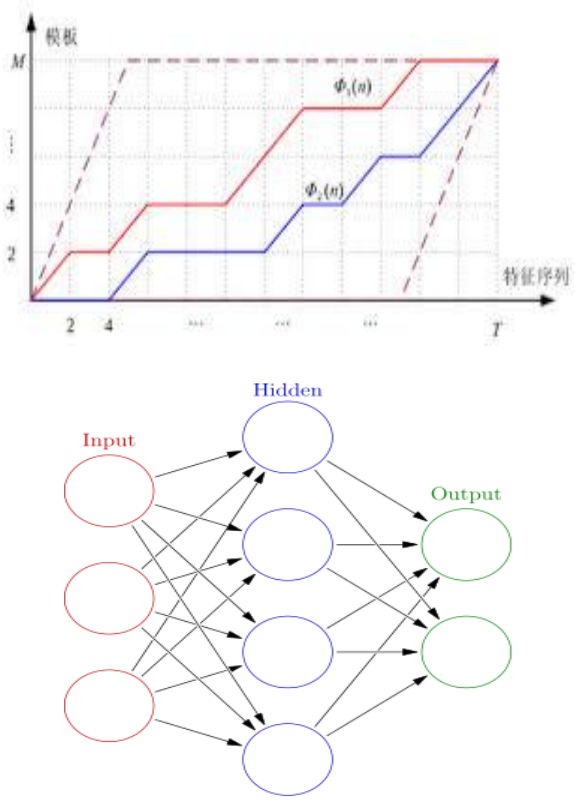

Fig. 3. Dynamic type wrappig algorithm.

One well known application of Dynamic Time Wrapping algorithm is automatic speech recognition which allows a computer to find an optimal match between two given sequences within certain bounds.

7.3 Neural Networks

Inspired from the biological neural networks, Neural Networks are computational algorithms used approximate or estimate functions that depend on a large number of unknown inputs. Neural networks emerged in the late 1980. Since then, they have been used in many different aspects of speech recognition including phoneme classification, isolated word recognition and speaker adaptation.

Unlike HMMs, these do not assume about statistical properties of the given input and possess several other qualities which makes them a better choice for speech recognition. But despite their effectiveness in classifying short-time units, they seldom succeed in continuous recognition tasks.

Like other systems that learn from the input data, Neural Networks have been applied in a number of hard tasks including speech recognition and computer vision using simple rule-based programming.

8 Audio Basic Properties

Audio just like any other entity possess properties. We will focus on two types of properties. Physical; and cognitive. Physical properties include measureable properties such as amplitude and phase. Whereas cognitive properties include properties associated with cognitive senses such as loudness and pitch.

8.1 Physical Properties

Sound is produced when the air pressure changes which is represented in the form of a wave which in turn is made up of sine waves having different frequencies, amplitude and phase. From experiments, it has been concluded that human ear does not sense change in the phase but does recognize change in the amplitude such as loudness and also any changes in the frequency such as change in the pitch of the sound. But changes in the phase are still important e.g. locating sound source on the basis of phase difference. This shows that human acoustical mechanism can analyze waveforms directly.

8.2 Cognitive Properties

Upon hearing a certain sound, humans perceive certain information on the basis of physical information but not amplitude or frequency. The extracted information can be general to specific e.g. hearing someone talking or hearing what specifically someone is talking about. This sound comprises of physical information only. But it is still very

difficult to derive information from this physical information such as distinguishing between silence, music, speech or noise from the audio sine wave.

9 Proposed System for Content-Based Rating of Movies

A movie consists of characters, a certain dimension, dialogues and music. We can regard characters as objects, dimension as the subject and dialogues and music as the sound. Here we will focus on the latter features. We will take into account the audio track of the movie to give a certain criterion and ratings based on our algorithm.

As has been mentioned earlier, this paper proposes a method for content-based classification movies. Our methodology is very basic and somewhat limited (partly because there are still a few limitations in ASRs) since our research hasn't matured much. It uses a very simple algorithm to compute the suitable audience for a movie by extracting the audio track of the movie via speech recognition and calculating the percentage of swear words present in the script. The speech-turned-text description is run and compared with the predefined dataset fed to the system. These swear words if they match with the data present in the dataset are recorded. On the basis of this count, the computer performs simple average calculation to derive a percentage of how many words matched the dataset (Fig. 4)..

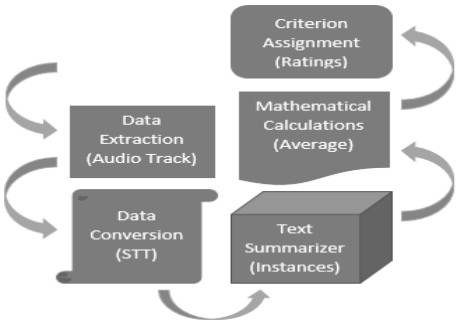

Fig. 4. General pipilien of the system.

Our system shall have the following features:

– Be speaker-independent.
– Be a continuous speech recognizer.
– Supports a text summarizer.
– Be based on a HMM.

All these properties have been carefully complied because a speaker-free system will have a wider range to recognize speech (although it may be only one domain e.g. American movies). With the ability to continuously recognize speech, the system won't face much trouble as the characters speak in a natural manner. As a HMM inspects each

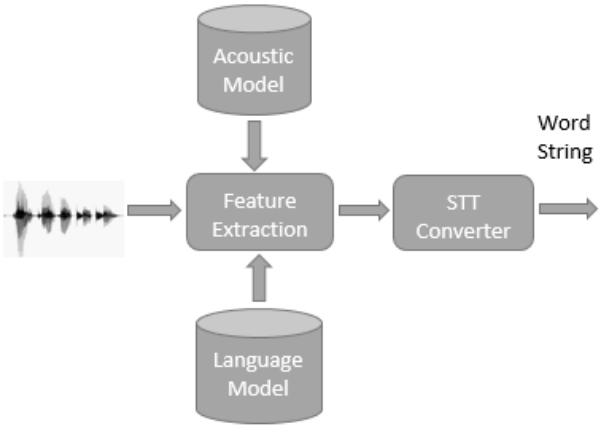

Fig. 5. Signal processing.

Table 1. Features example.

Term	Occurrence	Total
Word 1	07	07
Word 2	05	12
Word 3	09	21
Word 4	11	32
…	…	…
…	…	…
…	…	…

element singularly, therefore it is more reliable as we can assume that the results produced will be more accurate (Fig. 5 and .

Now we move on to how this algorithm actually works. It consists of three steps. Extracting the audio; analyzing and comparing it with the corpus; producing the results and assigning a criterion.

9.1 Audio Track Extraction

First of all, the system extracts the audio track or audio signal of the desired film using speech recognition. The features of the speech recognition system have already been mentioned earlier. After extracting the audio, a textual translation of is prepared via speech to text feature of the system.

9.2 Dataset

The dataset assigned to the system will contain all the information needed to perform the analysis. The problem of large data is resolved as there are only a few abuse or

swear words, may be 500 words in the entire language that are used repeatedly. Therefore the dataset can be made and compiled rather easily. Another interesting feature of the dataset is that it is updateable. If there's a new addition, the dataset can always be updated thereby eliminating any lapses in the analysis.

9.3 Text Analysis of Audio

Once the audio track's text form is achieved, this text form is analyzed by comparing it with the help of a text summarizer by giving in a few key words (in our case swear words) with the dataset that contains all the relevant information. The text summarizer gives all the instances present in the text that match the keyword. The number of occurrences is recorded and stored for further processing.

9.4 Calculating the %age

On the basis of occurrences of swear words recorded in the text, a simple mathematical calculation is performed to compute the %age of swear words present e.g. if in a text of 1000 words, 80 words are swear words, the movie will be given a percentage of 8% on the content scale (Fig. 6).

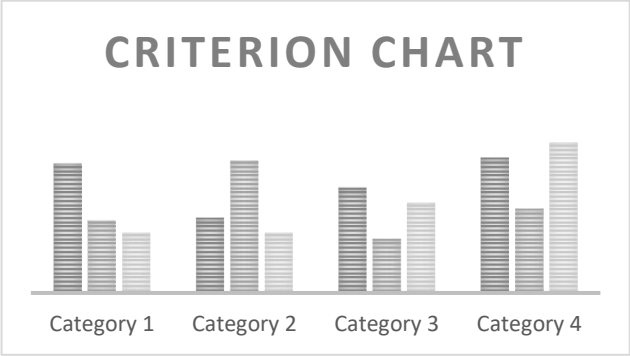

Fig. 6. Classification example.

On the basis of this percentage, the movie shall receive a rating and criterion. Movies can be rated PG-12, PG-13, PG-18+, R-Rated and so on. Each criterion shall be assigned a percentage range e.g. 10%-15% PG-10.

10 Conclusion and Future Work

In this paper, we studied speech recognition, its models, applications, uses and significance in details. We also reviewed audio based classifications and analysis. We came across many proposals ranging from scene segmentation to video frames to audio-visual strategies. At the end, we proposed a theory of our own that indexes online movies according to the percentage of swear words present in the movie's script using speech recognition and other natural language techniques. In the future, we hope to

implement out suggested theory successfully as it can aid a great deal in rating movies on online movies websites such as IMDB and Rotten Tomatoes thereby facilitating the viewers in making a decision and an appropriate choice.

References

1. http://www.lumenvox.com/resources/tips/types-of-speech-recognition.aspx.
2. Lu, L.: Content Analysis for Audio Classification and Segmentation. IEEE Signal Processing Society, 10, pp. 504–516, (2002). doi: 10.1109/TSA.2002.804546.
3. Wold, E., Blum, T., Wheaton, J.: Content-Based Classification, Search and Retrieval of Audio. IEEE Multimedia, 3, pp. 27–36 (1996). doi: 10.1109/93.556537.
4. Pal Singh, P.: Speech Recognition as Emerging Revolutionary Technology. International Journal of Advanced Research in Computer Science and Software Eng., 2(10), pp. 410–413 (2012)
5. Saraceno, C., Leonardi, R.: Audio as a Support to Scene Change Detection and Characterization of Video Sequences, Proc. of ICASSP'97, 4, pp. 2597–2600 (1997)
6. Wang, Y., Huang, J., Liu, Z., Chen, T.: Multimedia Content Classification Using Motion and Audio Information. Proc. of IEEE ISCAS'97, 2, pp. 1488–1491 (1997)
7. Pfeiffer, S., Fischer, S., Effelsberg, W.: Automatic Audio Content Analysis. Proc. ACM Multimedia'96, pp. 21–30 (1996). doi: 10.1145/244130.244139.
8. Nam, J., Tewfik, A.H.: Combined Audio and Visual Streams Analysis for Video Sequence Segmentation. Proc. of ICASSP'97, 3, pp. 2665–2668 (1997)
9. Saunders, J.: Real-Time Discrimination of Broadcast Speech/Music. Proc. of ICASSP'96, 2, pp. 993–996 (1996). doi: 10.1109/ICASSP.1996.543290.
10. Liu, Z., Wang, Y., Chen, T.: Audio Feature Extraction and Analysis for Scene Segmentation and Classification. The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, 20, pp. 61–79 (1998). doi: 10.1023/A:1008066223044.
11. http://en.wikipedia.org/wiki/Hidden_Markov_model.
12. http://en.wikipedia.org/wiki/Dynamic_time_warping.
13. http://en.wikipedia.org/wiki/Artficial_neural_network.
14. http://www.ee.columbia.edu/ln/LabROSA/doc/HTKBook21/node5.html.
15. http://ai-depot.com/ska/paper/node9.html.
16. Mahedero, J.P.G., Martínez, A., Cano, P.: Natural Language Processing of Lyrics. doi: 10.1145/1101149.1101255.
17. Gunsel, B., Tekalp, A.M.: Content-Based Video Abstraction. Department of Electrical Engineering and Center for Electronic Imaging Systems University of Rochester. doi: 10.1109/ICIP.1998.727150.
18. Ferman, A.M., Takalp, A.M., Mehrotra, R.: Effective Content Representation for Video. In: International Conference on Image Processing, 3, pp. 521–525 (1998). doi: 10.1109/ICIP.1998.727251.
19. Tsekeridou, S., Pitas, I.: Audio-Visual Content Analysis for Content-Based Video Indexing. In: IEEE International Conference on Multimedia Computing and Systems, 1, pp. 667–672 (1999). doi: 10.1109/MMCS.1999.779279.
20. Liu, S.C., Bi, J., Jia, Z.Q., Chen, R., Chen, J., Zhou, M.M.: Automatic Audio Classification and Speaker Identification for Video Content Analysis. In: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007), pp. 91–96 (2007). doi: 10.1109/SNPD.2007.516.

21. Huang, C., Fu, T., Chen, H.: Text-Based Video Content Classification for Online Video-Sharing Sites. Journal of the American Society for Information Science and Technology, 61(5), pp. 891–906. doi: 10.1002/asi.21291.
22. Huang, J., Kingsbury, B., Mangu, L., Padmanabhan, M., Saon, G., Zwei, G.: Recent Improvements In Speech Recognition Performance On Large Vocabulary Conversational Speech (Voicemail And Switchboard). In: Sixth International Conference on Spoken Language Processing (2000). doi: 10.21437/ICSLP.2000-819.
23. Jiang, W., Cotton, C., Chang, S.F., Ellis, D., Loui, A.C.: Short-Term Audio-Visual Atoms for Generic Video Concept Classification. Electrical Engineering Department
24. Nakamura, Y., Kanade, T.: Semantic Analysis for Video Contents Extraction - Spotting by Association in News Video. In: Conference: Proceedings of the Fifth ACM International Conference on Multimedia '97 (1997). doi: 10.1145/266180.266391.
25. White, D.J., King, A.P., Duncan, S.D.: Voice Recognition Technology as a Tool for Behavioral Research. Behav. Res. Methods Instrum. Comput., 34(1), pp. 1–5 (2002). doi: 10.3758/bf03195418.