ISSN(Online): 2320-801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Survey on Speech Emotion Detection Using SVM and HMM

[1]Vijayalakshmi B, [2]Dr.Sugumar Rajendran

[1]Research Scholar, Department of computer science, Bharathiar University, Coimbatore, India

[2]Associate Professor, Department of Computer Science and Engineering, Veltech Mutitech Dr.Ranagarajan Dr.Sakunthala Engg College, Chennai, India.

ABSTRACT: Emotion detection in speech processing is one of the burning arenas in data mining field. Detecting the motion of the speech is not that easy as it seems to be. Many different researchers have tried their approach ing this field but accuracy is the major factor of the processing. Our basic problem is to detect the kind of emotion gets detected from a pitch file. This would be done with the help of the HMM algorithm which would identify the frequency parameters. Then after finding the exact length of the file, we will have to get into the predefined clusters. Mugging into the predefined clusters would be achieved by the SVM algorithm and each cluster will rollback to a result value. The exact cluster which would give us the maximum probilitical analysis of the file would be our target cluster. This work is done previously with the help of ANN algorithm and they have provided an accuracy of about 92.1%.Our problem would be increasing this accuracy ratio, in comparison to the ANN module.

KEYWORDS: Audio and Video Segmentation, Audio and video classification, Support Vector Machine (SVM), Auto associate Neural Network (AANN), Hidden Markov model (HMM)

I. INTRODUCTION

Audio and video are used for enhancing the experience with Web pages (e.g. audio background) to serve music, family videos, presentations etc. The Web content accessibility guidelines recommend to always providing alternatives for time-based media, such as captions, descriptions, or sign language.

II. AUDIO AND VIDEO SEGMENTATION

The main objective of audio and video segmentation is to distinguish between the audio and video data. Various segmentation techniques are used to distinguish between the audio and video data.

III. AUDIO AND VIDEO CLASSIFICATION

The main objective of audio and video classification is to detect the category of audio and video data. The categories include news, advertisement, sports, serial and movies.

A. Methods used for audio and video based segmentation and classification:

- Support Vector Machine (SVM).
- Auto associate Neural Network (AANN).

Audio exists at everywhere, but is often out-of-order. It is necessary to arrange them into regularized classes in order to use them more easily. It is also useful, especially in video content analysis, to segment an audio stream according to audio types. In this paper, we present our work in applying support vector machines (SVMs) in audio segmentation and classification. Five audio classes are considered: silence, music, background sound, pure speech, and non-pure speech which includes speech over music and speech over noise. A SVM learns optimal

ISSN(Online): 2320-801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

class boundaries from training data to best separate between two classes. A sound clip is segmented by classifying each sub-clip of one second into one of these five classes. Experiments on a database composed of clips of 14870 seconds in total length show that the average accuracy rate for the SVM method is much better than that of the traditional Euclidean distance based (nearest neighbour) method.

B. Use of AANN for Audio and Video Segmentation and Classification:

Recent study shows that the approach to automatic audio classification uses several features. To classify speech/music element in audio data stream plays an important role in automatic audio classification. The method described in [1] uses SVM and Mel frequency spectral coefficients, to accomplish multi group audio classification and categorization. The method gives in [11] uses audio classification algorithm that is based on conventional and widely accepted approach namely signal parameters by MFCC followed by GMM classification. In [6] a generic audio classification and segmentation approach for multimedia indexing and retrieval is described. Musical classification of audio signal in cultural style like timber, rhythm, and wavelet confident based musicology feature is explained in [5]. An approach given in [8] uses support vector machine (SVM) for audio scene classification, which classifies audio clips into one of five classes: pure speech, non pure speech, music, environment sound, and silence.

Automatic video retrieval requires video classification. In [7], surveys of automatic video classification features like text, visual and large variety of combinations of features have been explored. Video database communication widely uses low-level features, such as color histogram, motion and texture. In many existing video data base management systems content-based queries uses low-level features. At the highest level of hierarchy, video database can be categorized into different genres such as cartoon, sports, commercials, news and music and are discussed in [13], [14], and [15]. Video data stream can be classified into various sub categories cartoon, sports, commercial; news and serial are analysis in [2], [3], [7] and [16]. The problems of video genre classification for five classes with a set of visual feature and SVM is used for classification is discussed in [16]. For the purpose of video classification, features are drawn from three modalities: text, audio, and visual. Regardless of which of these are used, there are some common approaches to classification. While most of the research on video classification has the intent of classifying an entire video, some authors have focused on classifying segments of video such as identifying violent [4] or scary [5] scenes in a movie or distinguishing between different news segments within an entire news broadcast [6]. Most of the video classification experiments attempt to classify video into one of several broad categories, such as movie genre, but some authors have chosen to focus their efforts on more narrow tasks, such as identifying specific types of sports video among all video [7]. Entertainment video, such as movies or sports, is the most popular domain for classification, but some classification efforts have focused on informational video (e.g., news or medical education) [8]. Many of the approaches incorporate cinematic principles or concepts from film theory. For example, horror movies tend to have low light levels while comedies are often well-lit. Motion might be a useful feature for identifying action movies, sports, or music videos; low amounts of motion are often present in drama. The way video segments transition from one to the next can affect mood [9]. Cinematic principles apply to audio as well. For example, certain types of music are chosen to produce specific feelings in the viewer [10] [5]. In a review of the video classification literature, we found many of the standard classifiers, such as Bayesian, support vector machines (SVM), and neural networks. However, two methods for classification are particularly popular: Gaussian mixture models and hidden Markov models. Because of the iniquitousness of these two approaches, we provide some background on the methods here.

IV. HMM

The Hidden Markov model (HMM) is widely used for classifying sequential data. A video is a collection of features in which the order that the features appear is important; many authors chose to use HMMs in order to capture this temporal relationship. An HMM represents a set of states and the probabilities of making a transition from one state to another state [12]. The typical usage in video classification is to train one HMM for each class.

ISSN(Online): 2320-801
ISSN (Print): 2320-9798

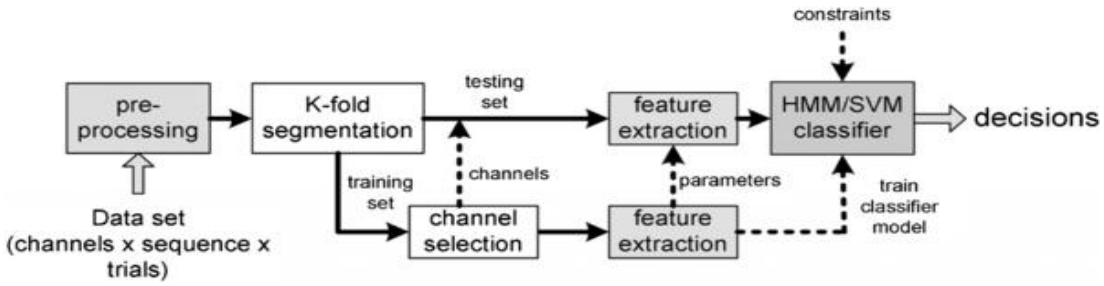
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

When presented with a test sequence of features, the sequence will be assigned to the class whose HMM can reproduce the sequence with the highest probability.

SVM Classifier



V. AUTO ASSOCIATIVE NEURAL NETWORK

Auto associative neural network models are feed forward neural networks performing an identity mapping. The modality would be the ability to solve the scaling problem. The AANN is used to capture the distribution of the input data and learning rule. The processing units in the first and third hidden layers are non-linear, and the units in the second compression/hidden layer can be linear or non-linear. As the error between the actual and the desired output vectors is minimized, the cluster of points in the input space determines the shape of the hyper surface obtained by the projection onto the lower dimensional space. The AANN captures the distribution of the input data depending on the constraints imposed by the structure of the network, just as the number of mixtures and Gaussian functions do in the case of Gaussian mixture model.

VI. TEXT BASED APPROACH

Text-only approaches are the least common in the video classification literature. Text produced from a video falls into two categories. The first category is viewable text. This could be text on objects that are filmed (scene text), such as an athlete's name on a jersey or the address on a building, or it could be text placed on-screen (graphic text), such as the score for a sports event or subtitles [7]. Text features are produced from this viewable text by identifying text objects followed by the use of optical character recognition (OCR) [2] to convert these objects to usable text. The text objects can become features themselves, which we discuss in the section on visual features. The second category is the transcript of the dialog, which is extracted from speech using speech recognition methods [13] or is provided in the form of closed captions or subtitles. Closed captioning is a method of letting hearing-impaired people knows what is being said in a video by displaying text of the speech on the screen. Closed captions are found in Line 21 of the vertical blanking interval of a television transmission and require a decoder to be seen on a television [14]. In addition to representing the dialog occurring in the video, closed captioning also displays information about other types of sounds such as sound effects (e.g., [BEAR GROWLS]), onomatopoeias (e.g., grrrr), and music lyrics (enclosed in music note symbols, _). At times, the closed captions may also include the marks >> to indicate a change of speaker or >>> to indicate a change of topic [15]. In addition to closed captioning, text can be placed on the television screen with open captioning or subtitling. Open captioning serves the same purpose as closed captioning, but the text is actually part of the video and would need to be extracted using text detection methods and OCR. Subtitles are also part of the video in television broadcasts, although this isn't necessarily the case for DVDs. However, subtitles are intended for people who can hear the audio of a video but can't understand it because it is in another language or because the audio is unclear; therefore, subtitles typically won't include references to non-dialog sounds.

ISSN(Online): 2320-801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

VII. AUDIO BASED APPROACH

Audio-only approaches are found slightly more often in the video classification literature than text-only approaches. One advantage of audio approaches is that they typically require fewer computational resources than visual methods. Also, if the features need to be stored, audio features require less space. Another advantage of audio approaches is that the audio clips can be very short; many of the papers we reviewed used clips in the range of 1-2 seconds in length.

This paper proposed an automatic audio-video based segmentation and classification using AANN. Mel frequency spectral coefficients are used as features to characterize audio content. Colour Histogram coefficients are used as features to characterize the video content. A non linear support vector machine learning algorithm is applied to obtain the optimal class boundary between the various classes namely advertisement, cartoon, sports, songs by learning from training data. An experimental result shows that proposed audio-video segmentation and classification gives an effective and efficient result obtained.

VIII. VIDEO BASED APPROACH

Most of the approaches to video classification that we surveyed rely on visual elements in some way, either alone or in combination with text or audio features. This corresponds with the fact that humans receive much of their information of the world through their sense of vision. Of the approaches that utilize visual features, most extract features on a per frame or per shot basis. A video is a collection of images known as frames. All of the frames within a single camera action are called a shot. A scene is one or more shots that form a semantic unit. For example, a conversation between two people may be filmed such that only one person is shown at a time. Each time the camera appears to stop and move to the other person represents a shot change, but the collection of shots that represent the entire conversation is a scene. While some authors use the terms shots and scenes interchangeably, typically when they use the term scene they are really referring to a shot. Many visual-based approaches use shots since a shot is a natural way to segment a video and each of these segments may represent a higher-level concept to humans, such as “two people talking” or “car driving down road”. Also, a shot can be represented by a single frame, known as the key frame. Typically the key frame is the first frame of a shot, although some authors use the term to refer to any single frame that represents a shot. Shots are also associated with some cinematic principles. For example, movies that focus on action tend to have shots of shorter duration than those that focus on character development. One problem with using shot-based methods is that the methods for automatically identifying shot boundaries don’t always perform well. Identifying scenes is even more difficult and there are few video classification approaches that do so. The use of features that correspond to cinematic principles is popular in the visual-based approaches, more so than in text-based and audio-based approaches. These include using colours as a proxy for light levels, motion to measure action, and average shot length to measure the pace of the video. One difficulty in using visual-based features is the huge amount of potential data. This problem can be alleviated by using key frames to represent shots or with dimensionality reduction techniques, such as the application of wavelet transforms.

IX. VISUAL FEATURES

1) Color-Based Features: A video frame is composed of a set of dots known as pixels and the color of each pixel is represented by a set of values from a color space [4][8]. Many color spaces exist for representing the colors in a frame. Two of the most popular are the red-green-blue (RGB) and hue saturation- value (HSV) color spaces. In the RGB color space, the color of each pixel is represented by some combination of the individual colors red, green and blue. In the HSV color space, colors are represented by hue (i.e., the wavelength of the color percept), saturation (i.e., the amount of white light present in the color), and value (also known as the brightness, value is the intensity of the color) [4][9]. The distribution of colors in a video frame is often represented using a color histogram, that is, a count of how many pixels in the frame exist for each possible color. Color histograms are often used for comparing two frames with the assumption that similar frames will have similar counts even

ISSN(Online): 2320-801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

though object motion or camera motion will mean that they don't match on a per pixel basis. It is impossible to determine ,in rare cases a single shot may contain more than one scene. From a color histogram the positions of pixels with specific colors, so some authors will divide a frame into regions and apply a color histogram to each region to capture some spatial information.

Another problem with color-based approaches is that the images represented in frames may have been produced under different lighting conditions and therefore comparisons of frames may not be correct. The solution proposed by Drew and Au [5] is to normalize the color channel bands of each frame and then move them into a chromaticity color space. After more processing, including the application of both wavelet and discrete cosine transforms, each frame is now in the same lighting conditions.

2) MPEG: One of the more popular video formats is MPEG (Motion Pictures Expert Group), of which there are several versions. We provide a high-level and somewhat simplified description of MPEG-1; for more complete details, consult the MPEG-1 standard[5][1].During the encoding of MPEG-1 video, each pixel in each frame is transformed from the RGB color space to the Y Cb Cr color space, which consists of one luminance (Y) and two chrominance (Cb and Cr) values. The values in the new color space are then transformed in blocks of 8 × 8 pixels using the discrete cosine transform (DCT). Much of the MPEG-1 encoding process deals with macro blocks (MB), which consist of four blocks of 8 × 8 pixels arranged in a 2 × 2 pattern. Consecutive frames within the same shot are often very similar and this temporal redundancy can be exploited as a means of compressing the video. If a macro block from a previous frame can be found in the current frame, then encoding the macro block can be avoided by projecting the position of this macro block from the previous frame to the current frame by way of a motion vector [5][2]. Much research has been conducted on extracting features directly from MPEG video, primarily for the purpose of indexing video [5][3]. For video classification the primary features extracted from MPEG videos are the DCT coefficients and motion vectors. These can improve the performance of the classification system because the features have already been calculated and can be extracted without decoding the video.

3) Shot-Based Features: In order to make use of shots, they first must be detected. This has proven to be a difficult task to automate, in part because of the various ways of making transitions from one shot to the next. Lienhart [4] states that some video editing systems provide more than 100 different types of edits and no current method can correctly identify all types. Most types of shot transitions fall into one of the following categories: hard cuts, fades, and dissolves. Hard cuts are those in which one shot abruptly stops and another begins [5]. Fades are of two types: a fade-out consists of a shot gradually fading out of existence to a monochrome frame while a fade-in occurs when a shot gradually fades into existence from a monochrome frame. A dissolve consists of one shot fading out while another shot fades in; features from both shots can be seen during this process. While it is important to understand shot transition types in order to correctly identify shot changes, the shot transition types themselves can be useful features for categorization [6].

X. CONCLUSION AND FUTURE SCOPE

Undergoing the estimated procedure, we expect our conclusion to be a better accurate system for the analysis of the audio files to detect the emotions in the field of clustering. We expect the accuracy to be increased by 2 to 5 percent in comparison with the ANN. SVM combined with HMM is expected to work in better manner because the training set created with the help of SVM and HMM puts a strong emphasis in searching into the inner clusters of the files. In future, work can be done to create more groups into the inner cluster of the files stored so that the searching becomes easy. To perform such task, one can opt the CART algorithm which creates a regression tree which is a substitute of the binary decision trees. The future parameters can add the time slots of the frequencies at which the frequencies are consistent

ISSN(Online): 2320-801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

REFERENCES

[1] Dhanalakshmi. P.; Palanivel. S.; and Ramaligam. V.; (2008), "Classification of audio signals using SVM and RBFNN", In Elsevier, Expert systems with application, Vol. 36, pp. 6069–6075.
[2] Kalaiselvi Geetha. M.; Palanivel. S.; and Ramaligam. V.; (2008), "A novel block intensity comparison code for video classification and retrieval", In Elsevier, Expert systems with application, Vol. 36, pp 6415-6420.
[3] Kalaiselvi Geetha, M.; Palanivel, S.; and Ramaligam, V.; (2007), "HMM based video classification using static and dynamic features", *In proceedings of the IEEE international conference on computational intelligence and multimedia applications.*
[4] Palanivel. S.; (2004)., "Person authentication using speech, face and visual speech", *PhD thesis, I IT, Madras*.
[5] Jing Liu.; and Lingyun Xie.; "SVM-based Automatic classification of musical instruments"*, IEEE Int'l Conf., Intelligent Computation Technology and Automation (2010.),* vol. 3, pp 669–673.
[6] Kiranyaz. S.; Qureshi. A. F.; and Gabbouj. M. ; (2006), "A Generic Audio Classification and Segmentation approach for Multimedia Indexing and Retrieval"., *IEEE Trans. Audi., Speech and Lang Processing,* Vol.14, No.3, pp. 1062–1081.
[7] Darin Brezeale and Diane J. cook, Fellow. IEEE (2008), "Automatic video classification: A Survey of the literature", *IEEE Transactions on systems, man, and cybernetics-part c: application and reviews*, vol. 38, no. 3, pp. 416-430.
[8] Hongchen Jiang. ; Junmei Bai. ; Shuwu .Zhang. ; and BoXu. ; (2005)," SVM - based audio scene classification", *Proceeding of NLP-KE*, pp. 131–136.
[9] V. Vapnik.;"Statistical Learning Theory", *John Wiley and Sons*, New York, 1995.
[10] J.C. Burges Christophe.; "A tutorial on support vector machines for pattern recognition," Data mining and knowledge discovery, No. 2, pp. 121–167, 1998.
[11] Rajapakse. M .; and Wyse. L.; (2005), "Generic audio classification using a hybrid model based on GMMs and HMMs ",*In Proceedings of the IEEE ,pp-1550-1555.*
[12] Jarina. R.; Paralici. M.; Kuba. M.; Olajec. J.; Lukan. A.; and Dzurek. M.; "Development of reference platform for generic audio classification development of reference plat from for generic audio classification", *IEEE Computer society, Work shop on Image Analysis for Multimedia Interactive (2008),* pp-239–242.
[13] Kaabneh,K. ; Abdullah. A.; and Al-Halalemah,A. (2006). , "Video classification using normalized information distance", In *proceedings of the geometric modeling and imaging – new trends* (GMAP06) (pp. 34- 40).
[14] Suresh. V.; Krishna Mohan. C.; Kumaraswamy. R.; and Yegnanarayana. B.; (2004).,"Combining multiple evidence for video classification", *In IEEE international conference Intelligent sensing and information processing (ICISIP-05),* India (pp.187–192).
[15] Gillespie. W. J.; and Nguyen, D.T (2005).; "Hierarchical decision making scheme for sports video categorization with temporal post processing", *In Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (CVPR04) (pp. 908 -913).
[16] Suresh. V.; Krishna Mohan. C.; Kumaraswamy. R.; and Yegnanarayana. B.; (2004).,"Content-based video classification using SVM", *In International conference on neural information processing*, Kolkata (pp. 726– 731).
[17] Subashini, K.; Palanivel, S.; and Ramaligam, V.; (2007), "Combining audio-video based segmentation and classification using SVM", *In International journal of Signal system control and engineering applications,* Vol.14, Issue.4, pp. 69–73.