2009

Automatic Video Classification

Don Jayakody
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Part of the [Computer Sciences Commons](#)

Recommended Citation

Jayakody, Don, "Automatic Video Classification" (2009). *Master's Projects*. 45.
DOI: <https://doi.org/10.31979/etd.4sdc-f2te>
https://scholarworks.sjsu.edu/etd_projects/45

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Automatic Video Classification

A Writing Project

Presented to

The Faculty of the Department of Computer Science

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Don Jayakody

Fall 2009

Copyright © 2009

Don Jayakody

ALL RIGHTS RESERVED
SAN JOSÉ STATE UNIVERSITY

The Undersigned Thesis Committee Approves the Thesis Titled

AUTOMATIC VIDEO CLASSIFICATION

by

Don Jayakody

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

Dr. Robert Chun, Department of Computer Science 12/09/2009

Dr. Teng Moh, Department of Computer Science 12/09/2009

Mr. Tejas Bhandare, Brocade Communications 12/09/2009

APPROVED FOR THE UNIVERSITY

Associate Dean Office of Graduate Studies and Research Date

ABSTRACT

Within the past few years video usage has grown in a multi-fold fashion. One of the major reasons for this explosive video growth is the rising Internet bandwidth speeds. As of today, a significant human effort is needed to categorize these video data files. A successful automatic video classification method can substantially help to reduce the growing amount of cluttered video data on the Internet. This research project is based on finding a successful model for video classification. We have utilized various schemes of visual and audio data analysis methods to build a successful classification model. As far as the classification classes are concerned, we have handpicked News, Animation and Music video classes to carry out the experiments. A total number of 445 video files from all three classes were analyzed to build classification models based on Naïve Bayes and Support Vector Machine classifiers. In order to gather the final results we developed a “weighted voting - meta classifier” model. Our approach attained an average of 90% success rate among all three classification classes.

ACKNOWLEDGMENT

My sincere thanks go to my advisor Dr. Robert Chun for dedicating his valuable time, in some occasions even past midnight, to listen to my thought process. Dr. Chun's valuable advises saw me through this project especially when I thought I had no chance of finding a reasonable solution with such a massive data set.

I also like to thank Dr. Teng Moh for introducing me in to this wonderful world of data mining. Without his CS 274 class and his guidance, I would be nowhere near such a successful completion of this project.

I also thank my other committee member Mr. Tejas Bhandare, for devoting his valuable time to sit in my committee.

Sincerely,

Don Jayakody

Table of Contents

1. Introduction	1
1.1 Online Video Growth	1
1.2 Online Documents to Online Video	2
2. Related Work	6
2.1 Text based video classification.....	8
2.2 Audio based video classification	10
2.3 Visual based video classification	12
2.4 Hybrid approaches for the video classification.....	14
3. Classification Methodologies	17
3.1 Rule Based Classifier	17
3.2 Nearest-Neighbor Classifier	18
3.3 Naïve Bayes Classifier.....	19
3.4 Support Vector Machine (SVM)	20
4. Design Process & Experiments.....	22
4.1 Classification Categories	22
4.2 Video Formats.....	22
4.3 Data Set Information.....	23
4.4 Software Tools, Development Kits Used.....	24
4.5 Experiments	25
4.5.1 Single Mode Classification Approach - Naïve Bayes Classifier	25
4.5.2 Single Mode Classification Approach – Rule Based Classifier	31
4.5.3 Improvements.....	35
4.5.4 Change in Directions	36
5 Final Design Model and Results	41
5.1 Final Design Model.....	41
5.2 Final Results	42
6 Future Work and Conclusion	46
Bibliography	47

List of Figures

Figure 1 - Data Set Information.....	23
Figure 2 - Percentages of Video Categories.....	23
Figure 3 - Single Mode Classification Approach.....	26
Figure 4 - Processing of a News Video	28
Figure 5 - Results Derived From the Single Mode Classification Approach- Naïve Bayes Classifier	30
Figure 6 – Results For the Rule Based Classifier	33
Figure 7- Frequency Domain Feature Analysis	38
Figure 8 - Time Domain Analysis for a Heavy Metal Video	40
Figure 9 - Final Decision Model.....	42
Figure 10 - Final Results based on SVM and Naïve Bayes methods.....	43
Figure 11 - Number of Correct Hits for News, Animation & Music Videos.....	45

1. Introduction

1.1 Online Video Growth

The explosive growth of the video usage within the past two years is a direct result of the huge increase in internet bandwidth speeds. According to the comScore, a leading statistics reporting company for digital media, 150 million U.S. internet users have watched 96 videos in average per viewer in December 2008. (1) This number is a resultant of the 13 percent increase of US online audience in the month of December 2008 compared to the previous month. The underline factor beneath this huge increase in online video usage is the surge in online traffic due to the increasing bandwidth speeds all across the globe. Cisco Visual Networking Index projects global IP traffic will increase at a rate of 46 percent from year 2007 to 2012, which is essentially doubling the traffic every two years. (2) To put this in perspective, in year 2012 internet video traffic alone will be equal to the 400 times the total traffic carried by the U.S. internet backbone in year 2000. (2) Furthermore the number of online video portals has drastically increased within the last two years. Currently there is a vast amount of web sites available on the Internet which includes some kind of streaming video contents. The combination of Google owned sites which include youtube.com, have streamed over ten billion videos in the month of August 2009. (3) The YouTube alone accounts for more than 99 percent of these streaming hits. Altogether a total of twenty five billion online videos have been viewed by the Internet users in the U.S. in August 2009. (3) The following table from comScore, summarizes the top U.S online video portals with regard to the number of videos viewed by the internet users in August 2009.

Table 1 - Top U.S. Online Video Content Properties* by Videos Viewed - August 2009

Property	Videos (000)	Share (%) of videos
<i>Total Internet*</i>	<i>25,366,195</i>	<i>100.0</i>
Google Sites	10,051,924	39.6
Microsoft Sites	546,547	2.2
Viacom Digital	539,471	2.1
Hulu	488,255	1.9
Fox Interactive Media	380,115	1.5
Yahoo! Sites	355,226	1.4
Turner Network	298,991	1.2
CBS Interactive	168,993	0.7
Disney Online	162,934	0.6
AOL LLC	156,871	0.

*Total Internet -U.S. Home/Work/University Locations

Source: comScore Video Metrix (3)

1.2 Online Documents to Online Video

All these statistics pinpoint one absolute truth: the usage of online videos is growing in exponential numbers. We can compare this exact phenomenon to the text based documents when the Internet was at its infant age in the late 90's and early 2000 time period. The amount of online text based documents were increasing in astonishing rates during the early days of the Internet resulting in a vast amount of cluttered data, scattered around all across the net. If it's not

for the web crawlers and the search engines, these data would have been of little or no importance, simply because no one can find it and use it when there is a need. The same argument can be applied to the growing amount of online videos. If these online videos aren't properly tagged and categorized the benefit of having these videos online would be greatly reduced. However there's a fundamental difference when it comes to automatic object classification of these two worlds, the text documents world and the video world. With the text based documents, any simple web crawler carries the competency to go through those documents and extract the words and basic meanings. In essence, given a search query, a search engine would be able to utilize the extracted words to output reasonably accurate results. Unfortunately, the case is not so simple and straightforward with online videos. It is impossible to go through a huge amount of videos, such as the videos available on the Internet, and extract meanings out of it, so that a search engine would use those results to respond to a query. For example, a practical use case scenario would be an internet user uploading a sports video clip to the YouTube. After uploading the video to the YouTube website that user has the option to categorize the video to appropriate sections, in this case the "Sports" category. Also, that user can define the appropriate keywords and tags such as: "Baseball", "Red sox", "Word Series". If the user decided not to manually categorize and define the tags and the keywords, any other users would not have been able to retrieve this video.

Hence, the automatic video classification problem is fundamentally different from the classical document classification problem. As mentioned earlier this is mainly due to the semantic differences between a text file and a video file. In easiest terms we can define a text file as a one dimensional file which contains only the text dimension. On the other hand, a video file can be defined as a three dimensional file which contains, all three of the dimensions: audio, visual and

text. There has been a significant progress in document classification research work using various different methods. According to Allamraju & Chun (4) related work section, different authors have proposed various different systems for document classification and clustering problem. Budzik et.al (5) have proposed a keyword extracting system which extracts keywords from a document that are representative of the document's content. These keywords then get fed in to a web search engine and the results are generated according to this extracted keyword. (4) In another type of research that utilizes documents, an automatic summarizer has been built by the authors based on the frequency of the words, cue phrase, location, title and query method. In the word frequency method, each sentence is assigned a score based on the relevant words in that sentence. In the cue phrase method, each sentence was assigned a cue score based on the presence of relevant and important phrases. In the location method, a score is assigned to the sentence based on its location in a paragraph or proximity to headings. In the title method, sentences containing words present in the document's title are given a higher score. In the query method, sentences matching the query words are given more importance. The final decision is based on the weighted sum of the frequency, cue phrase, location, title and query method. (4) By varying and multiplexing the decision among five categories these authors were able to gain a high granularity to the expected results because the weighted final score is a representation of the sentences that are most important and most representative to the content of the original document.

However it is not impossible to generate a video classification solution that will classify videos to different categories based on the content of that particular video. There has been a considerable amount of research and groundwork done by different people and organizations

regarding the video classification problem. In the next section we will discuss the previous work that has been done related to the video classification field.

2. Related Work

The TREC (Text Retrieval Conference) conference, sponsored by National Institute of Standards and Technology (NIST) is an important part of the video classification field. In fact TREC Vid (TREC Video Retrieval Evaluation) was branched off from the original TREC as a result of this growth in the field starting from year 2003. Today TREC Vid has become a benchmarking and evaluation campaign for the automatic video classification field. However, most effort from TREC Vid is focused on retrieving video information and using that information in a search query so that a user would be able to search through a video for a specific content. The automatic video classification problem is a slightly different problem when compared to the video retrieval problem. For instance, an automatic video classifier will define a particular video as a sports video or a news video, whereas a video retrieval machine will focus on indexing each and every portion of the video for future retrieval. One very good example of a video retrieval system is “Gaudi” system (<http://labs.google.com/gaudi>) that has been developed by Google. Gaudi lets the user search through a specific video for a specific keyword and identifies the exact keyword location in a given video(s). But Gaudi only focuses on the audio portion of the video. As of today Gaudi can only retrieve videos using audio indexing.

There has been a significant amount of work done so far to solve the automatic video classification problem. As mentioned earlier, a video file contains three dimensions of data portions: visual, audio and text. According to Brezeale and Cook (6) previous video classification efforts can be classified in to four approaches that go in par with those three dimensions. Namely, the four categories are text-based approaches; audio based approaches, visual based approaches and mixed approaches. Most authors have used a combination of these

approaches because semantics of the video has audio, visual and text components in it. Furthermore, video classification can be applied in different manners. Some authors may choose to classify the video as a whole while others may choose to classify specific feature or a component of a video. For example, while one author tries to classify a whole video segment as a news video, another might focus on identifying and classifying the business news section of that particular video. (6) Continuing on this trend of classification, while most authors may focus on classifying videos in to rather broad categories such as action movies, comedy movies, romantic movies, some authors have attempted a narrow categorization methods. For example instead of classifying videos as sports videos, they have attempted to classify videos as basketball, baseball videos.

Many of these efforts have incorporated cinematic principles or concepts from the film theory. (6) For example, when compared with comedy movies, horror movies have low lighting levels in the scenes. If you get an aggregated number of well-lighted scenes vs. dark scenes there is a higher chance that a horror movie containing a larger number of dark scenes. Also when comparing action movies vs. romantic movies it is apparent that action movies contain higher ratio of fast moving sceneries. Utilizing these kind of cinematic principles tend to yield very accurate results in classifying videos. However, focusing only on cinematic principles in visuals may not be sufficient to get all rounded results. These cinematic theories can easily be applied to audio as well. Audio is a major part of the so called “feeling generation” in a movie. It is common that specific types of audio are integrated with specific scenes to generate certain feelings in the viewers mind. For example horror scenes may contain much more screaming, noisy sounds vs. a romantic movie. Romantic movies tend to have soft audio levels in the music portion of the sceneries.

2.1 Text based video classification

Classification methods that are solely based upon text-based solutions are among the least preferred methods when it comes to video classification. A solution which is based on text based approaches has a higher degree of similarity in comparison with the text document classification methods. However, the extent to which these texts are extracted from the video itself makes the whole effort rather unique compared to the document classification approaches. Methods such as Optical Character Recognition,(OCR) and speech recognitions have been used in a considerable number of researches. These text-extracting options can be categorized in to two different methodologies: the text viewed on screen vs. the text extracted from sounds (6). Text viewable on screen, such as stock quotes from a business news video, or scoreboard results from a baseball game can be extracted using OCR solutions. Also OCR can be very helpful if a video contains on screen subtitles. Capturing the subtitles has its own advantages and disadvantages. For example, an obvious advantage would be if the subtitles were in the language of which the classification is being done. If so, capturing the subtitles would be similar to getting a transcript of the whole video. However this is not the norm in most cases. Subtitles are mostly used to bridge the language barriers. A disadvantage of capturing only the subtitles is that it would not give any record of non-dialog sounds. Speech recognition techniques can also be applied to capture the transcript of a video.

Many of the text based video classification approaches have employed the vast amount of research that has already been done for the document text classification problem. This is one of the main advantages when it comes to text only based video classification research. The rate of successful results for this stream of research depends upon how much appropriate text can be

extracted from a video. If it's closed caption text, which use text to describe every single sound including dialogs and non-dialogs (such as sound of a vehicle by use of text that states "vehicle starting" etc.) the final results would produce a higher hit rate. W. Zhu et al. (7) have used closed-captioned text in their research to categorized news videos. Authors have used 425 news stories from CNN and compared the categorizing performance with different classification methods. News video stories were initially segmented based on the demarcation in the closed-caption text, such as the symbol ">>>", which indicates the switch of topics. (7) They have used a natural language processor to process the captured text. Using that language processor, keywords from noun phrases and proper nouns were extracted. From this point on, they were able to utilize the text categorization techniques, and the problem has thus become a classic text categorization problem. Those authors have defined eight categories: Politics, Daily Events, Sports, Weather, Entertainment, Health, Business and Science & Technology. The 425 stories were manually segmented and labeled for both training and testing purposes. The system was trained with a randomly selected ten to eighty percent of the stories and tested with the remaining ones. The authors have achieved a high precision and recall rates for the categories with salient language feature such as Sports, Weather, Health and Business. However, as expected, categories such as Daily Events failed to perform well since it is such a broad category and had a very limited number of unique language features. Their approach has yield poor results for the Entertainment and Science & Technology categories due to the insufficient training examples where no unique feature items could be found. When selecting training and testing sets they have used random methods to pick the two different sets. The authors have found out that their method provided the best performance with an accuracy of eighty percent, when eighty percent of the samples were used for training. (7) When applying the natural language parser to pass the

key words, they have kept the first twenty unique keywords to process further. Authors have further found out that the first twenty keywords have the best ratio to improve the accuracy. In order to calculate weights for each combination class and keyword and perform the classification they have used the following formula:

$$w_{ij} = P(c_i | f_j)^2 (\log(m_j) + 1)$$

where,

w_{ij} is the weight value for each pair of category i and feature item j ,

m_j is the the number of stories with feature item j appearing in the training samples, and,

$P(c_i | f_j)$ is the conditional probability of category i given the feature item j , which could be computed as the percentage of stories with the feature item j that are labeled as category i . (7)

After w_{ij} is calculated, the total sum of the weights for all keywords in a news segment, it is then assigned to one of the categories from politics, daily events, sports, entertainment, weather, business and science, or health & technology, depending on which category correlates to the highest sum.

2.2 Audio based video classification

Audio based video classification is another form of categorization method that has gained a significant popularity when it comes to video classification technologies. Compared with pure text based classification, audio based classification methods tend to yield more accurate results. Because of this reason, many video classification methods are based on audio based approaches

rather than text based approaches. Furthermore, compared to solely visual based approaches, audio based approaches seem to incur considerably lower processing cost. Based on the differences between audio and video files, audio based approaches require less computational power than the visual based approaches. If a particular scene separated into an audio based scene and a visual based scene, the audio based scene tends to carry more information than its video counterpart. This is also quite apparent when the file sizes are examined. For video based analysis, at least 10 – 15 seconds worth of visuals are needed to identify some key characteristics. However, in an audio clip, it may be sufficient to use 1 – 2 second clips for characterization. In fact, many researchers have used audio clips ranging around these lengths for their work. In order to examine the necessary components in an audio file, it is necessary to process the signal using a steady sample rate. Some of the commonly used sampling rates are 44.1 kHz and 22050 Hz. After deriving the samples using the sampling rate, those individual samples can be gathered to form frames. This frame-forming process is highly analogous to how visual scenes are processed. Additionally, to further enhance the process it is possible to collect these frames and define frame boundaries so that those frames can be identified using a key frame. After collecting these frames audio files are processed in two different domains: frequency domain and the time domain. (6) In the time domain the amplitude of a signal with regards to the time is analyzed, whereas in the frequency domain, the amplitude with regards to the frequency is considered. This time domain to frequency domain conversion can be done with the Fourier transformation.

When comprehending audio-based features, it is much more beneficial to take the human interpretation about sound into account. For example by using the audio information we can derive three different layers of information: low-level acoustics, midlevel sounds, and high level

sounds. (8) Lui & Wang et al. have worked on a video classification scheme that is solely based on audio feature extracting, with a sampling rate of 22050 Hz for the audio sampling (8). They have focused on audio attributes such as non-silence ratio, standard deviation and dynamic range of the volumes, frequency, noise to voice ratio, standard deviation of the pitch and energy levels, etc. To detect the frames that are silent they have compared the volume and zero crossing rate (ZCR – the times that an audio wave crosses the zero axis) of each frame to preset thresholds. In this work authors have figured out that, if both volume and ZCR are less than the threshold values the frame can be declared as silent. Throughout their research the importance of using ZCR measurement is highlighted. Furthermore ZCR values help to avoid the low energy speech frames from being classified as ‘silent’.

2.3 Visual based video classification

Many video classification efforts that have been done so far is based on some kind of a visual based approach. This is very intuitive given that anyone would agree the visual element is the most important dimension out of the three dimensions of a video. Therefore in order to gain from these visual clues most researchers have incorporated visual based approaches to their work. Most of these research that use visual features tend to extract features on a per frame or per shot basis. Basically, a video is a collection of images commonly referred to as frames. All of the frames within a single camera action comprise a shot. A scene is one or more shots that form a semantic unit. (6) To clarify this point let's consider an example: a scene in a baseball game. The moment when the pitcher starts to pitch the ball to the moment when someone catches the ball can be considered as one semantic unit, in other words one whole scene. Even though the camera moves from one angle (from pitcher) to another angle where it focuses on the batsman the whole

unit can be considered as a scene. Sometimes authors refer to a scene as a shot. Throughout this paper we will use scene and a shot interchangeably to refer to a same semantic unit within a given video. Many of the prior research use some kind of scene/shot-based approach as the basis of their classification. Instead of analyzing a whole video one frame at a time, it makes logical sense to analyze it scene by scene. A whole scene can be represented as one logical unit of the whole story. For example in an action movie there can be a scene where two people are fighting, or a car chasing scene. We can represent these scenes using one key frame. As mentioned earlier even though a particular scene has multiple frames, in order to get a representative picture only one key frame is being used. When it comes to these scenes, the above-mentioned cinematic principles can be applied as well. The scenes which are extracted from a horror film may contain much more low-lighted scenes compared to scenes from a comedy movie. Also scenes from an action movie may contain lots of fast moving frames compared to scenes from a romantic movie. Nevertheless, these types of scene-based approaches have their own disadvantages. First of all direct identification of scenes in a given video clip can be a tedious task. Unless it's done manually, it may not be intuitive to develop an automatic scene selection algorithm for a video clip. The major reason for this is that scene boundaries can be very hard to identify for some video clips. Sometimes a whole video clip may only contain one logical shot, and at other times it can be multiple shots. The definition of a shot boundary can be different from one person to another based on multiple factors such as their taste, the way they analyze a clip, etc. Anyhow, we cannot undermine the importance of these shot based approaches. After all it is one of the major factors that can contribute to a very good classification scheme.

Girgensohn et al. (9), in their paper describes a video classification method based on the visual features only. They have chosen to classify video frames into six different classes: presentation

graphics, long shots of the projection screen lit, long shots of the presentation screen unlit, long shots of the audience, medium close-ups of human figures on light backgrounds, and medium close-ups of human figures on dark backgrounds. Frames have then been extracted from MPEX videos every 0.5s. Each frame is converted to a 64 x 64 grayscale intensity image. (6) After extracting the frames they have used a various different transform algorithms to transform the vectors and perform the classification.

2.4 Hybrid approaches for the video classification

As emphasized above, a video has a three dimensional nature to it and most of the work that has been done so far incorporates hybrid classifying methods. Utilizing characteristics of video, audio, and text attributes tend to produce much more efficient results than incorporating only one such feature. This is because most of the time in order to get a coherent meaning out of a video, the audio, visual and textual features of a particular video file must be taken into account. Most researchers have utilized at least one more combination of video, text and audio along with their primary choice to overcome unnecessary fluctuation of final results. However the combination of all or a multiple of these features have its own barriers: such as combining all the metrics into one single metric so that end result would behave according to the most accurate feature. Different authors have explored a number of methods: some have chosen to build one single feature vector while others have trained classifiers for each modality and then used another classifier for making the final decision.

Wei et al. (10) have developed a classification method based on face and text processing for different types of TV programs. The features used for this distinct classification were obtained

from the tracking of faces and of super-imposed text in the video stream. This face and text tracking consisted the basis of their classification method and they have put a great deal of dependability to this tracking system. The entire classification method is based on how well this tracking system behaves. They have identified two issues involved in such object tracking methods: the detection of the targets in each frame and the extraction of object trajectories over frame sequences to capture their movements. For the face tracking purpose authors have used a tracking scheme that utilized YUV color coordinates (We used YUV as one of the metrics. It will be described later in this report) for skin-tone region segmentation to adapt to the MPEG-1 and MPEG-2 framework. Authors claim that their system was accurate owing to the utilization of different features and the novel iterative partition process. Interestingly, they haven't applied these face detection techniques to every single frame. It would have been somewhat inefficient to apply high intensity face detection processing to every single frame of every single video in both training and testing sets. To improve the speed of the face tracking, they have taken advantage of the content continuity between consecutive frames by considering the joint detection of faces and trajectory extraction. More importantly, they have utilized one very significant cinematic principle: the variation of faces within a continuous shot is usually small. Hence they have only applied the face detection to the first few frames of each shot. For each detected face, the mean and the standard deviation in color, the height, the width and the center position have been computed. Authors have picked four types of TV programs to do their classification: news, commercials, sitcoms and soap operas. To classify a given video segment into one of these four categories, they have mapped it into the same feature space and evaluated its probability of being each category by the weighted distances to the centers of the news, commercial, sitcom and soap clusters. Another interesting phenomenon these authors have incorporated in their research is the

use of domain knowledge. For example, in news and commercials, optical (gradual) cuts are frequently used by the editor to ensure a smoother and pleasant visual effect, whereas in sitcoms and soap operas most cuts between shots are abrupt. Thus, the percentage of optical cuts among all cuts can be used as an extra feature to update the probability value. When a video segment contains a high portion of optical cuts, a choice of news or commercial is favored over sitcom and soap. (10) Authors have found out that utilizing this kind of extra, yet intuitive domain knowledge information can improve the accuracy of the final result in significant volumes. For their particular study they have been able to improve those results by twenty percent. As far as the outputs are concerned, authors have used twenty-six video segments exclusively for the training set. Out of these twenty-six videos there were five segments of news videos, six commercials, eleven sitcom segments, and four soap opera videos. Their testing set contained thirty-five segments that were different from the original training set. During the final classification process they have been able to correctly label twenty seven out of the these thirty five segments, and eight were misclassified thus achieving a seventy seven percent hit rate. Authors have figured out that most errors originated from news segments labeled as commercials and sitcoms labeled as soaps. One example would be fusion programs such as “Datelines” in NBC containing fewer faces than regular news programs, which essentially seems to misguide the classification trajectories that used face and text detection. (10)

3. Classification Methodologies

In this project, four different classification schemes were carried out in an attempt to find the best approach. The goal was to be achieved by categorizing the data set into four classes: news videos, sports videos, animation videos and music videos. At the beginning, the best approach at hand was to perform a trial and error analysis of the data set. One major hurdle in video classification is the definition of video categories. The video categories chosen for this project is rather unique. As mentioned in the survey paper done by Brezeale and Cook (6), almost all the automatic video classification research that has been done so far does not have a common set of guidelines for different categories. Accordingly, the experiments differed in the number and type of video classes, number of training and testing examples, and length of video clips. Compared to the text document classification there hasn't been that much of research done in the video classification area. Hence, one other difficulty faced was that there weren't any results that could be set up as a target for the chosen categories, animation and music. This was important in order to improve the performance of the classification schemes. In the upcoming sections a high level overview is given about each and every classification scheme followed during the initial phase of this project.

3.1 Rule Based Classifier

The Rule based classifier is one of the simplest classification methods. This approach utilizes a set of pre-determined rules to generate the final result. These rules consist of a collection of "if...then...." clauses.

For example:

$$rule_1 = (RGB\ level > x) \wedge (Sound\ energy\ level < y) \rightarrow Animation$$

$$rule_2 = (YUV\ level < a) \wedge (Frequency\ Bandwidth\ level < b) \rightarrow News$$

$$rule_3 = (Sample\ Rate < e) \wedge (Pitch\ level > f) \wedge (Frequency\ Bandwidth\ level > b) \rightarrow Music$$

The rules for the model are represented in a disjunctive normal form.

$R = (rule_1 \vee rule_2 \vee rule_3)$ where R is the “rule set” and each $rule_i$ ’s are the classification rules or disjuncts.

In this method, the rules get evaluated in a top to down manner. For example, if a certain news video clip has RGB levels and Sound energy levels that are consistent with the first rule, that file would get classified under ‘Animation’. Mutually exclusive rules in a rule set R are defined as any two rules which do not trigger the same record. This property ensures that every record is covered by at most one rule in R. A rule set R has exhaustive coverage if there is a rule for each combination of attribute values. Such rules are known as “Exhaustive Rules”. This property ensures that every record is covered by at least one rule in R. (11)

3.2 Nearest-Neighbor Classifier

Another method explored in this project is the nearest-neighbor classification scheme. The idea behind the nearest neighbor classification method is that it tries to find all the training examples that are relatively similar to the attributes of the test example. Given a video to be classified, if

we can find out which of the training set videos has the highest level of similarity compared to this testing video, then there is a higher chance that this test video may belong to that training set video category. In contrast to the decision tree and rule-based classifiers which attempt to find a global model that fits the entire input space, nearest-neighbor classifiers make their predictions based on local information. However, since nearest-neighbor classifiers only use local information to make their predictions there is a significantly higher chance that these classifiers are susceptible to noise. Therefore, especially when dealing with higher noise data domains, such as video data, extra precautions must be taken while training and testing the data.

3.3 Naïve Bayes Classifier

The Naïve Bayes Classifier applies the Bayes theorem as the fundamental statistical principle when carrying out the classifications. Bayes theorem utilizes prior knowledge of the classes with new evidence gathered from subsequent input data. The following equation represents the compact form of the Bayes theorem which uses conditional probabilities:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$P(Y|X)$ commonly refers to the posterior probability and $P(Y)$ and $P(X)$ represent the prior probabilities of variables X & Y . The Naïve Bayes Classifier however uses a slightly different version of the formula which includes the conditional independence assumption. (11)

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^d P(X_i|Y)}{P(X)}$$

Furthermore, the Naïve Bayes Classifier carries unique advantages when compared to the other classification methods. One significant advantage would be its robustness in isolating noise points as such points are averaged out when estimating conditional probabilities from data. This is a very important factor when dealing with video data. Also, Naive Bayes classifiers can handle missing values by ignoring that sample during model building and classification. Moreover, the Naïve Bayes Classifier provides a unique advantage when it comes to irrelevant attributes. If X_i is an irrelevant attribute then $P(X_i | Y)$ becomes almost uniformly distributed. In such a case the class conditional probability for X_i has no impact on the overall computation of the posterior probability. (11)

3. 4 Support Vector Machine (SVM)

The Support Vector Machine is a classification technique that has been broadly used in various different practical applications, ranging from handwritten digit recognition to text categorization. (11) One of the main reasons for SVM to have gained such a wide popularity is due to the fact that it works very well with high-dimensional data and therefore avoids the curse of dimensionality problem. Moreover, the SVM has the capability to represent the decision boundary using a subset of the training examples, known as the support vectors. A SVM learning model can be formulated as a convex optimization problem. Also, the SVM can utilize various different algorithms to find the global minimum of the objective function. Other classification methods such as rule-based classifiers and artificial neural networks use a greedy strategy to search the hypothesis space, thus making those solutions only locally optimized. In contrast, the SVM can find a solution that is globally optimized for a given set of data.

Another valuable feature in SVM classifier is its ability to handle dummy variables. SVM can be applied to categorical data by introducing dummy variables for each categorical attribute value present in the data. For example, if the Delay attribute has three values {Analog Delay, Dynamic Delay, Multitap Delay} SVM allows us to use a binary variable to aggregate all three features.

4. Design Process & Experiments

4.1 Classification Categories

For this project, the author initially decided to classify videos into four categories: news videos, sports videos, animation videos and music videos. News video category and sports video category are two of the video categories that have been used by several published research papers. However, author is yet to find any peer reviewed research content that is focused on animation and music video categories. We further classified the music video category in to heavy metal music vs. classical music.

4.2 Video Formats

The Flash video format is used to carry out our experiments. The reasons behind selecting this particular format are multifold. First and foremost, Flash is the predominantly used video format throughout the Internet. Most of the online video portals including YouTube uses Flash format to deliver its content. When a user uploads a video file to the YouTube, it is transcoded in to the flash video format (flv). However, the quality of the video suffers a lot due to this transcoding process. The ubiquitous nature of the Adobe Flash player is one other reason for Flash to be widely deployed in the Internet. By default, videos delivered to the Adobe Flash Player are downloaded progressively into the browser's cache, similar to how images are downloaded when any websites containing graphics are loaded (12).

4.3 Data Set Information

For the classification purposes we've gathered 445 video files which include videos from various news broadcasts, animation videos, music videos and sports videos.

Figure 1 - Data Set Information

Figure 2 - Percentages of Video Categories

The selection of the video data set and its breakdown under each category was random and bound by the time limitations for this project. Since we started with news video classifiers we were able to process 155 news videos for the classification scheme. We were able to process 110 sports videos from different sports, including basketball, baseball and soccer. The hundred animation clips were based on different movie trailers, and cartoon animations. Music videos consisted of different genres of music. More emphasis was put on heavy metal music video clips and classical music clips.

4.4 Software Tools, Development Kits Used

Different propriety and open source version software were employed for this project. All the video processing was done by industry leading Adobe Premiere Pro CS4 software. Premiere Pro version has a rich set of features that can be utilized to analyze and edit any video content. Also for the Flash video processing Adobe premiere comes with its own text-transcribing engine.

For all the audio processing the Adobe Audition software was used. Adobe Audition is among the industry leading software for audio processing and analyzing. Furthermore, Audition has the capability to process multiple audio files at the same time.

For the video classification purpose, the open source Weka data mining software suite was utilized. Weka is based on Java programming language and it has a collection of machine learning algorithms for data mining tasks (13). The users have the capability to incorporate the Weka code base into their own Java code and call the Weka libraries for the classification purposes. Weka is very well known for its rich set of libraries that can be utilized for machine

learning purposes. All the metrics derived through video analysis and audio analysis with the help of the Adobe Premiere and Adobe audition were fed in to Weka for further processing. Weka expects an input document that is in the Weka-specific ARFF file format for further processing. This specialized ARFF file contains the definitions of attributes to relation metrics values.

Furthermore, PHP has been used for front-end development tasks. Front-end comprises of a MySQL database and Apache web server. For digital video playback we used an open source solution called “Flow Player”.

4.5 Experiments

4.5.1 Single Mode Classification Approach - Naïve Bayes Classifier

Throughout this project we had to change our development model accordingly in order to adapt to the unexpected end results. In the very beginning of the project, it was decided to stick to one classification scheme which can be use as an “oracle model” to classify all the genres of videos. Single classification schemes have been used in many video classification approaches carried out so far. For example, the Support Vector Machine based classifiers or the Bayesian classifiers have been utilized in the past to generate such a single classification model. Going along this practice we decided to build our first trial system based on a single classification scheme. Following is the flow chart representation of the workflow:

Figure 3 - Single Mode Classification Approach

For the first set of trials we have used the naïve Bayes classification model to derive the results. From our total set of video data, we had 155 videos manually labeled as “news videos”. For the news classification purpose these data can be referred as positive examples. Rest of the videos in the data set then becomes negative examples. When considering the news classification problem we have focused mainly on the visual features of the video. Video attributes such as frame rate, media duration, video resolution, blur, sharpen levels, Gaussian blur, silence, noise and RGB values were taken into account. Moreover, critical audio attributes such as, the number of beats,

fixed length tempo matching, pitch, audio in point and audio out point have also been included in the feature matrix.

Figure 4 shows a screenshot from adobe premiere window while processing a news video.

Figure 4 - Processing of a News Video

For all the four different categories of news, animation, sports and music videos, we have manually selected forty percent of the videos as the training set to build the Naïve Bayes classification model. The classification training set is used as a pre-determined value set to build mathematical relations between data. The percentage of training set videos to be selected from the original data is entirely up to the user to define. For this specific experiment we assumed forty percent of a training set would be more than sufficient. However, it has been found out that this number plays an important role in the final decision, details of which will be discussed in depth later in this report.

4.5.1.1 Results Derived From the Single Mode Classification Approach- Naïve Bayes Classifier

The derived results from this particular experiment were far from being perfect.

Figure 5 - Results Derived From the Single Mode Classification Approach- Naïve Bayes Classifier

Regarding the empirical data from other related work, these results demonstrated that there is something significantly wrong with our process. At this point, conclusions were drawn about two aspects of our process:

- There exists some unusual elements in the data domain that we have failed to consider
- We have neglected some components in our calculation process

Naïve Bayes classification approach tends to yield good results when the attributes have less correlation among them. However in our data set this might not be necessarily true. For instance, the RGB values tend to correlate significantly with the lighting percentages. The Naïve Bayes classifier performance can be significantly degraded due to these kinds of correlated attributes. Several more runs were executed using the same approach to eliminate any possibilities of miscalculations. But every time we ended up with similar results as above. Therefore at this point it was decided to proceed with another mode of classification.

4.5.2 Single Mode Classification Approach – Rule Based Classifier

The Rule based classifier has several characteristics that were important with regard to the video data we gathered. In the Rule based classification model, the expressiveness of a rule set is almost equivalent to that of a decision tree because a decision tree can be represented by a set of mutually exclusive and exhaustive rules. We decided to program the Weka software model, so that it will allow us to extract rules and build a classification model based on the Rule based classifier. Also we would be able to study if the Rule-based classifier allows multiple rules to be triggered for a given record to avoid any unnecessary confusion among the correlated data patterns.

We used the sequential covering algorithm to build our rule extraction model. Sequential covering algorithm is a widely used algorithm to extract rules directly from the data. This algorithm tends to yield results in a rather greedy fashion. Since we have four intended

classification categories, the algorithm extracts rules from one category at a time for data sets that contains more than two categories. Following is a step-by-step run down of the implementation of this algorithm: (Original source: (11), amended to incorporate our domain of data)

Let E be the training records and A be the set of attributed-value pairs that belongs to four categories of the data (video, sports, animation & music), $\{(A_j, v_j)\}$

Let Y_0 be an ordered set of classes $\{y_1, y_2, \dots, y_k\}$.

Let $R = \{ \}$ be the initial rule list.

For each class y that belongs to $Y_0 - \{y_k\}$ do

While stopping condition is not met do

$r \leftarrow \text{Call}(\text{Learn_One_Rule}(E, A, y))$.

Remove training records from E that are covered by r .

Add r to the bottom of the rule list: $R \rightarrow R \vee r$.

End while.

End for

Insert the default rule, $\{ \} \Rightarrow y_k$, to the bottom of the rule list R

Below are the summaries of results that we obtained by the implementation of the Rule based classification algorithm:

Figure 6 – Results For the Rule Based Classifier

As clearly observed, the derived results from the Rule-based classification produced results that are slightly better than the previously employed Naïve-Bayes classification method. Especially for the animation video category the Rule-based classification method produced comparatively better results than the rest of the video categories. One reason for this outcome is the difference in color scheme attributes that were incorporated. We used both RGB color schemes as well as YUV color coordinates as the attribute values to define skin-tone region segmentation. RGB uses red, green and blue colors to define the color scheme. These three colors are combined in different portions and percentages to derive another color in this space. The YUV scheme defines the color spectrum in a totally different manner. The Y section of the YUV scheme defines the intensity of the color as opposed to an actual color compared to the RGB scheme. Intensity of the color essentially describes the brightness of the color space. On the other hand, the U and V components, similarly to the RGB scheme, define an actual color. Even though these two color schemes have some similarities between their components, the YUV color scheme can be utilized along with the RGB color scheme to better detect the human skin-tone in a given video. Especially when categorizing animation videos it is a fairly straightforward process to incorporate this kind of human skin tone analysis. A large portion of animation videos tend to contain a high ratio of RGB colors and we can define the brightness of those colors using the YUV color scheme. One reason for the Naïve Bayes Classifier not to pick up this information to its advantage is because, the correlation between the YUV and RGB schemes are comparatively higher when compared with other attributes like pitch level, noise level and frame rate.

Even though the Rule based classifier yielded better results when we examine it in one perspective, it can be easily concluded that this scheme is not capable of producing optimal

results under these conditions. After calculating these results we came to the conclusion that we need to be more careful with regard to our data and its analysis process. After a careful examination we found out two major improvement categories that would help us to significantly improve our results. In the next section we will present those process improvements in detail.

4.5.3 Improvements

4.5.3.1 Improvements to the Current Process

We figured out the way in which we pick our training sets yielded a significant drawback in our final results. Up till this point, we have picked forty percent from each video category as our training set to train the data models. However when we randomly increase the training set percentages we saw a shift in final results. Our final results moved towards the positive direction for each classification method. Zhu & Toklu et al. (7) have used varying percentages of training samples in their research. They have tried to classify videos using Bayes decision methods with training samples that ranged from ten to eighty percent. These authors have selected their training and testing sets randomly. They have also found out that there is a optimal percentage of training samples that can be used to derive the maximum result rate. In their research they have derived that eighty percent of training data would be sufficient to gain the optimum result rate of eighty percent.

Hence we tried to simulate the same process with our training set data. We varied our training set from thirty five percent to ninety percent. Also we tried to mix and match the training set by dropping randomly selected training set videos back to the testing set. By doing this we were effectively violating the norms of the classification research. The sole purpose of incorporating

training data in to the testing set is to figure out the standard deviation of the results. After several runs, we came to the conclusion that this kind of approach wouldn't add much value to the whole process, because we were essentially shifting the results towards the positive edge by using the already known samples. Going back to the original training set mix and match we found out the perfect equilibrium value for our data set was sixty two percent.

4.5.3. 2 Improvements to the Data Set

One other improvement among many others that we studied and yielded significantly positive results was the normalization of the data set values. Since our data set contains 445 video files from various different classes of categories each and every video file had its own quality. We figured that the actual quality of the video file (whether it's a high resolutions video, low resolution video etc.) would make a huge difference in the attribute values. In order to mitigate this problem we decided to apply a normalization method. The normalization process results in a standardized set of data among all the four different classes of videos. The goal of normalization was to make the entire set of attribute values comparable between two different video categories. For example, the ZCR and pitch values of an audio clip may be significantly different between a high quality music video file and a low quality music video file, even though both are in a common classification field.

4.5.4 Change in Directions

When we figured out that we have to do much further processing with respect to the data set, we decided it would be better to abandon the classification of sports category in the interest of time.

One of the major reasons to choose the sports category was based on the poor results that we already received from the two classification methods: Naïve Bayes and Nearest Neighbor. Variety of the sports videos that we have in our video library may have played a significant role in the accuracy of the final results. Instead of focusing on a one particular sport, we have picked videos from many different sports such as Basketball, Baseball, etc. Difference of the pace of the game, and movements of the players may have caused the result shift. Even though we eliminated the sports category from our project, for the actual classification process we used all those videos as negative test cases.

In order to make up for the lost category we decided to perform more audio processing of music videos. As mentioned earlier, audio processing is much more efficient and a less tedious task compared to video processing. We chose to analyze audio features in both time domain and frequency domain. Adobe audition provides a very rich framework for this kind of analyzing purposes. The following figure depicts frequency domain visuals for an audio file in Adobe Audition:

Figure 7- Frequency Domain Feature Analysis

In addition to the frequency domain feature analysis of the audio, we proceeded further to perform a time-domain feature analysis. Such a time-domain feature analysis was essential to further classify the music videos into two define genres: namely Hard Metal and Classical. We sampled the audio signal at a 22KHz level. We created 1024 samples out of each audio frame.

We paid special attention to various vectors such as noise level, volume, non-silence ratio, pitch etc. As can be observed in the following figure, the time-domain graph has a much more compact representation for a Heavy Metal audio clip due to the high noise, high pitch, and high frequency attributes.

Figure 8 - Time Domain Analysis for a Heavy Metal Video

5 Final Design Model and Results

5.1 Final Design Model

After improving our data sets and training sets using the above-mentioned methods, our results improved significantly. Our target percentage was above ninety percent hit rate for any classification methods. In order to obtain this target, we decided to go ahead with the superior classification method for this video classification data domain: Support Vector Machine (SVM). Many authors have used SVM as their primary classification model. The SVM has the capability to better select margins among the attributes. Since most of our video and audio attributes have rather narrow margins among them we decided to use a proven model like SVM to get better results for our data.

At this point, with three different classification models for all three classes (News, Animation and Music), it was necessary to build another model to aggregate and pick the best results from all three mechanisms. Lin & Hauptmann (14) describes a meta-classification strategy in their research to resolve a similar problem. They have developed a meta classification strategy so that they treat the judgment from each classifier for each class as a feature, and then build another classifier model called a meta-classifier to make the final decision. We too built our final model based on a similar meta-classifier. However we further included a weighted voting methodology, originally described by Zhu & Liou (7) to incorporate results from different classification schemes. Consequently, each scheme has the opportunity to participate in the final decision based on the success rate of each classification class.

Figure 9 - Final Decision Model

5.2 Final Results

Support Vector Machine classification model yielded a very high hit rate for both news and animation segments. However for the music segment, Naïve Bayes Classifier yielded higher results. The reason behind this observation is, because we primarily focused on the audio data attributes compared to the visual attributes for the music video segment. Naïve Bayes in nature are highly robust to isolated noise points because such points are averaged out when estimating the conditional probabilities from the original data.

Figure 10 - Final Results based on SVM and Naïve Bayes methods

We were able to derive more than 85% hit rate for all three classification categories. For News and Animation category the valid classification hit rate was beyond 90%. We were quite confident that we would be able to attain 90% hit rate for music video category if it wasn't for the time restrictions. Following is a graphical summary of all the classification hit rates we were able to produce:

Figure 11 - Number of Correct Hits for News, Animation & Music Videos

6 Future Work and Conclusion

Throughout this project we have explored many ways to fine tune our results. Unfortunately, we didn't have enough time to do a thorough analysis of these different tune-up methodologies. Also we had to abandon the sports category due to time restrictions. These are some of the future-work that must be done to further progress on this particular research project. In terms of a more holistic picture of video classification, one of the main things that need to be improved is the speed of classification. Text document classification has come to a point where it will only take several seconds to classify certain domain of documents. However this is not valid in the video classification workspace. Furthermore, video classification needs a higher amount of processing power in both audio and video components. Therefore it takes up quite a large computation time to carry out the processing. If we can reduce the time boundaries of these processing components and thus reduce the speed of the classification at least by one fold, it would be a great advancement in the field of video classification.

Bibliography

1. **comScore Press Release.** U.S. Online Video Viewing Surges 13 Percent in Record-Setting December. <http://www.comscore.com>. [Online] 2009.
http://www.comscore.com/Press_Events/Press_Releases/2009/2/US_Online_Video_Viewing_Sets_Record.
2. **News@Cisco.** Cisco Visual Networking Index Projects Global IP Traffic to Reach Over Half a Zettabyte(1) in Next Four Years. *newsroom.cisco.com*. [Online] June 16, 2008.
http://newsroom.cisco.com/dlls/2008/prod_061608b.html.
3. **comScore Press Release.** Google Sites Surpasses 10 Billion Video Views in August. *www.comscore.com*. [Online] September 2009.
http://www.comscore.com/Press_Events/Press_Releases/2009/9/Google_Sites_Surpasses_10_Billion_Video_Views_in_August.
4. **Sri Harsha Allamraju, Robert Chun.** *Enhancing Document Clustering through Heuristics and Summary-based Pre-processing*. San Jose State University, Department of Computer Science, San Jose, CA : s.n., 2009.
5. *Information access in context, Knowledge Based Systems.* **J.Budzik, K.J. Hammond, L.Birnbaum.** 2001, Vol. 14, pp. 37-53.
6. *Automatic Video Classification: A Survey of the Literature.* **Darin Brezeale, Diane J. Cook.** s.l. : IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, May 2008, Vol. 38.
7. *Automatic News Video Sementation and Categorization Based on Closed-Captioned Text.* **Weiyu Zhu, Candemir Toklu, Shih-Ping Liou.** s.l. : IEEE International Conference on Multimedia and Expo, 2001.
8. *Audio feature extraction and analysis for scene segmentation and classification.* **Z.Lui, Y. Wang , T.Chen.** s.l. : J. VLSI Signal Process. Syste. Vol. 20, pp. 61-79.
9. *Video Classification using transform coefficients.* **A. Girgensohn, J.Foote.** s.l. : IEEE Intl. Conf. Acoust. Speech Signal Process(ICASP), 1999. Vol. 6, pp. 3045-3048.
10. *TV Program Classification Based on Face and Text Proecessing.* **Gang Wei, Lalitha Agnihotri, Nevenka Dimitrova.** s.l. : IEEE Int. Conf. Multimedia. Expo., 2000. Vol. 3, pp. 1345-1348.
11. **Pang-Ning Tan, Michael Steinbach, Vipin Kumar.** *Introduction to Data Mining*. s.l. : Pearson - Addison Wesley. p. 210. 0-321-32136-7.
12. **Adobe Developer Connection.** Video Learning Guide for Flash: Progressive and streaming video. *www.adobe.com*. [Online] Adobe.
http://www.adobe.com/devnet/flash/learning_guide/video/part02.html.
13. Weka - Data Mining Software In Java. [Online] <http://www.cs.waikato.ac.nz/ml/weka/>.

14. *News Video Classification Using SVM-based Multimodal Classifiers and Combination Strategies*. **Wei-Hao Lin, Alexander Hauptmann**. s.l. : Calnegie Mellon University.