2009

VIDEO SCENE DETECTION USING CLOSED CAPTION TEXT

Gregory Smith
Virginia Commonwealth University

School of Engineering, Computer Science Department
Virginia Commonwealth University

This is to certify that the thesis prepared by Gregory L. Smith entitled VIDEO SCENE DETECTION USING CLOSED CAPTION TEXT has been approved by his or her committee as satisfactory completion of the thesis requirement for the degree of Master of Computer Science.

Dr Ju Wang, Department of Computer Science

Dr. James E. Ames, IV, Department of Computer Science

Dr. Yuichi Motai, School of Engineering

[Click here and type your Committee Member's Name and School Name.]

[Click here and type the Department Chair's Name or Representative's Name.]

[Click here and type your School or College Dean's Name]

Dr. F. Douglas Boudinot, Dean of the Graduate School

May 15, 2009

VIDEO SCENE DETECTION USING CLOSED CAPTION TEXT

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Computer Science at Virginia Commonwealth University.

by

GREGORY L. SMITH
BACHELOR OF COMPUTER SCIENCE
RENSSELAER POLYTECHNIC INSTITUTE, 1985

Director: DR. JU WANG
ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE

Virginia Commonwealth University
Richmond, Virginia
May 2009

Acknowledgements

This work would not have been possible without the support and encouragement of my thesis advisor Dr. Ju Wang and instructor Dr. Yuichi Motai. Their advice and supervision during my thesis and coursework were instrumental to the success of this project.

I would like to thank my family who supported me throughout my Master's program. The original inspiration for this work was the video biography I created for my parent's fortieth wedding anniversary. It was the positive reaction to the video by them and my extended family that convinced me that every family should have a video biography. It is to all of them that I dedicate this work.

Table of Contents

List of Tables

Index of Tables

Abstract

VIDEO SCENE DETECTION USING CLOSED CAPTION TEXT

By Gregory L. Smith, B.S. Computer Science

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Computer Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2009

Major Director: Dr. Ju Wang
Associate Professor, School of Engineering, Computer Science Department

Issues in Automatic Video Biography Editing are similar to those in Video Scene Detection and Topic Detection and Tracking (TDT). The techniques of Video Scene Detection and TDT can be applied to interviews to reduce the time necessary to edit a video biography. The system has attacked the problems of extraction of video text, story segmentation, and correlation. This thesis project was divided into three parts: extraction, scene detection, and correlation. The project successfully detected scene breaks in series television episodes and displayed scenes that had similar content.

CHAPTER 1 : Introduction

Communicating family histories has, until relatively recently, been the domain of personal diaries, journals, letters, and even stories told in the oral tradition. In the modern age electronic communication has become the norm. People are using telephones, cell phones, voice mail, email, text messaging, blogs, and "twitters" to communicate. These newer forms of communication are increasingly short and ethereal. Since these media are much less permanent than traditional forms, the natural documentation of family histories is being lost to digital rot.

In recent years, a small cottage industry has emerged to create family video biographies. These videos are created through the use of consumer video and computer equipment. Families can capture their histories in a permanent form that is both lasting and entertaining for future generations.

However, the cost of creating these videos is very high. It can cost as much as $10,000 - much higher than the average family is able to afford. To reduce this cost a new technology is necessary.

Video Biography is arguably the simplest form of film. It involves the videotaping of subjects (either singly or in pairs) and rearranging the videotaped interviews into a

coherent story. As such, it is a very "talking heads" presentation of friends and family members of the subject.

Usually, there is one subject – the life of a person or couple. As such, the subject matter is composed of friends and family telling stories of the person's life and times. If the subject is a couple, the film usually tells the story of how the couple met and the events bringing them to where they are today. The video biography often includes short stories of the subject's ancestors and offspring.

The video biography is told both in first person ("I grew up in upstate New York") and in third person ("He was a funny young man"). When edited, it is usually chronologically organized. The final product can include artifacts (pictures, music, home movies and videos, and letters) to add color to the story.

Editing a video biography involves viewing and reviewing hours of interviews to distill the video down to a 30- to 90-minute final product. It is generally accepted that the ratio of editing time to actual video is about 60 minutes of editing for every minute of video produced [Hampe1997]. At $100/hr for video editing a family video biography can cost $3000-$9000.

To make video biography as affordable as ordinary photographic portraits, the 60:1 ratio must be reduced. The techniques Video Scene Detection and Topic Detection and Tracking (TDT) can be applied to potentially reduce this time down to 1:1.

CHAPTER 2 : The Video Biography Workflow

A video biography is arguably the simplest form of film. It tells the story of one or two people in a chronologically ordered series of scenes. The scenes are typically head and shoulders shots of friends and family members retelling stories of the subjects. The scene shots typically switch from one interviewee to another as they recount their memories of the subject. Artifacts such as videos, letters, and photographs can be overlaid on top of the interviewee's narrative to add context and color to the story. The video biography can last from 30 to 90 minutes.

When creating a video biography, the biographer interviews the subjects and constructs a rough timeline of the story to be told. Artifacts are gathered which enhance the story. The artifacts are labeled with the names and ages of the people involved and the location of the event. In the case of a photograph, the people in the photograph are named and the date and location of the photograph are recorded.

A rough time line of the events of the subjects' life is created with major events marked (such as births, weddings, deaths, and relocations).

From the initial interviews of the subjects, a list of potential interviewees is created. The interviewees are contacted and appointments are made to videotape them and record their stories. The biographer may prepare a set of questions for the interviewees beforehand.

During the video interviews, the biographer may use the questions to guide the interviewee through the time line. More often than not, the interviewee remembers events in a non-sequential order. The biographer may allow the interviewee to talk at length

about topics associated with the subjects without any attempt at structuring the interview. This allows free association on the part of the interviewee and can lead to unexpected and useful stories. The biographer may also collect artifacts after the video interview is over.

The fact that stories are not told in time sequential order is one of the reasons that editing a video biography is complicated. It may take several viewings of all the interviews to determine where the commonalities lie.

Once all the video interviews are complete, the process of reviewing the interviews begins. Each interview is watched and broken into scenes. Each scene is given a topic and a slot in the time line. For example "Vic enlists in the Air Force, 1961" may be attached to a particular start and end time in an interview with an interviewee. The segmentation of the stories into scenes is a time consuming process.

Once all the interviews are segmented and tagged, the scenes are reviewed again for how well they tell the story at each point in the time line. The scenes are then collected and "strung together" in an order that will tell the story best. Once all the scenes are selected and ordered, the editing process can begin.

The scenes for each point in the time line need to be combined to tell a coherent story. While it is possible to take a single interviewee and allow them to talk for several minutes about a topic, this is boring for the audience. According to [Hampe1997] thirty seconds is the maximum amount of time any individual should be on-screen at one time.

To facilitate this limitation, the technique of "weaving" can be used to split similar scenes into pieces. Then, the many slices of the different scenes can be combined by interleaving the slices to tell the story in an interesting way.

While modern computer editing suites allow for fun and clever transitions between scenes, these often inject more "energy" into the scene than is appropriate for a family video biography. The scenes and slices can be transitioned by using a simple cross fade between the two scenes.

Once the basic story is edited, artifacts such as home movies and photographs can be inserted over the narration. (Note that [Hampe1997] recommends against an omnipresent narrator in favor of using only the voices of the interviewees to tell the story). Artifacts can also be used to cover edit splices that are unusually unattractive.

Finally, any video titling and post production (such as introductory music) can be added.

CHAPTER 3 : Automatic Video Biography Editing

This paper takes aim at the problem of segmenting the video interviews into scenes with a defined topic. The literature identifies two major areas of research that can be applied to this problem. The first is Story Segementation. Story Segmentation analyzes video content (visual, auditory, and text) and attempts to break content into scenes with the intent to index it for retrieval. The other is Topic Detection and Tracking. TDT is used to analyze video news feeds in real time to identify when a story occurs, and if it is a new story. TDT is most often used in intelligence analysis.

This paper borrows from both research areas to split an interview into stories and determine what the interviewee was discussing. This can save the biographer time in the editing room.

Story Segmentation and Detection

The field of information retrieval includes the automatic review of audio/visual content and indexing it for retrieval. Hauptmann, et al. have processed broadcast news and segmented the content into stories and commercials [Hauptmann1997]. In their work, they used speech recognition, closed captioning, and video cues to separate stories from commercials. Merlino et al. have also created searchable content in their Broadcast News Editor and Broadcast News Navigator [Merlino1997]. [Fan2000] discusses the semantic labeling of content in an information retrieval system for videotaped medical procedures. In it, video images are associated with meanings for retrieval by concept, rather than textual comparison. And in [Alatan2001] series television is analyzed with Hidden

Markov Models to determine if a scene is a dialog, musical, non-dialog or establishing scene.

Topic Detection and Tracking

Allen, et al. describe Topic Detection and Tracking (TDT) and how it is used to segment and classify stories from televised news sources like CNN, etc... [Allen2000] The steps involved in TDT are:

1. Story Segmentation - dividing incoming transcripts into stories
2. First Story Detection - recognizing the emergence of a new topic in the story stream
3. Cluster Detection - grouping of stories based on their topics
4. Tracking - monitoring stories on a specified topic
5. Story Link Detection - determining if two randomly selected stories discuss the same topic

TDT evaluates a stream of text from newswire and speech-to-text systems. The goal of TDT is to break the text down into individual news stories. The news stories are 'tagged' with a topic. If a set of stories are similar enough, they are grouped together. If a news story breaks that has not been seen before (First Story Detection) then an analyst can be alerted to the incoming story.

Automatic Video Biography Editing

As with TDT, Automatic Video Biography Editing (AVBE) has a series of steps. They are:

- Acquisition – acquiring the video and audio interviews
- Conversion – converting the audio into text
- Extraction – extracting the text from the stream
- Segmentation – breaking the interviews into scenes and assigning them a topic
- Correlation – grouping like scenes together
- Weaving – slicing scenes apart and recombining different scenes together to make a coherent whole

Acquisition is the collection of video interviews is normally accomplished by going into the field with a video camera and interviewing subjects. Or, the subjects can travel to a studio and be interviewed off site.

Conversion is the process of converting the spoken words into text that can be interpreted by the computer. It can be accomplished by such means as speech recognition or closed caption translation. There are commercial services that will convert speech to text by using a human transcriber.

Extraction is the process of extracting the closed caption text from the video and putting it into a database. The captioned text must retain the timestamps back to the original video. This will allow the video to be edited with precision.

In segmentation the computer scans the video text for concepts and breaks the interviews into segments related to the topic the interviewee is discussing. This is the same task as Story Segmentation in TDT.

Correlation, is the process of grouping segments that are similar and organizing them chronologically. This is the same task as Cluster Detection in TDT.

Finally, weaving subdivides the correlated scenes and interlaces them so that multiple interview subjects talk about the same topic.

This paper deals only with extraction, segmentation and correlation.

Since interviews are mainly a head-and-shoulders view of a subject, the visual cues used in traditional TDT are not available for story segmentation. There are no dramatic video transitions, musical cues, camera changes or black-outs to commercials as in network television news programs where TDT is used [Allen2000].

Instead, video biography interviews are very "chatty" with nearly all information conveyed strictly by the spoken word. For this reason we rely solely on the use of text to perform story segmentation and correlation. We assume the video contains closed caption information or a script which gives the text of the video and timestamps of where the spoken text starts and ends.

CHAPTER 4 : Content

Interview content for analysis is not readily available. Creating a set of interviews would be simple enough, but the conversion to text is costly. It is estimated that 8 hours of text would be necessary to create a proper set of training and testing data.

For this reason, series television text was used. The closed caption text for Star Trek (The Original Series, seasons 1, 2, & 3) was extracted from the DVDs and used for the project.

Extraction and Conversion

The first phase of the project addresses extraction and conversion. The process of extracting closed captioned text and converting it into a database was demonstrated using DVDs of "Star Trek - The Original Series." This gave sufficient volume of text and video to demonstrate accurate extraction

To work with the video, the episodes were copied from the DVDs onto the computer's hard drive. The software to accomplish this was the open source program "DVD Decrypter." It stored the video in raw "VOB" format.

Once the video was accessible to the computer, two operations were performed: conversion of the video to MPG-1 form for displaying in a video frame, and extraction of the closed caption text.

NTSC video is displayed as a series of raster lines. There are 525 raster lines in one frame. Frames are presented at 60 per second in two fields of 262.5 lines each. The first frame displays all the odd raster lines and the second frame displays all the even lines.

So, the first frame is called the odd frame and the second frame is called the even frame. [Wiki2008a]

Not all raster lines are visible on an ordinary television set. The first 21 lines are used for the vertical blanking interval. Line 21 of the VBI contains digital data representing closed caption text. The resulting data stream has 960 bits/second and contains alphabetic, special, and control characters. [Wiki2008b]

The closed caption text was extracted from the Star Trek DVDs with the open source program "VobSub." It resulted in a flat file of text (an SRT file) with the following format:

line 1 : line number (monotonically increasing)
line 2 : start time -> end time
line 3 : text
line 4 : text
line 5 : text
line 6 : blank line
<repeats>

While the software did an effective job of extracting the closed caption text, it did not handle special characters or control characters. This resulted in spurious space (' ') characters in the text and occasionally the last few letters of the last line of text would end up on the first line. Special heuristics were applied to filter out the data anomalies and repair the errors.

The text extraction process was tested as part of a graduate level course in pattern recognition. A learning system was developed that would categorize each line of dialog as belonging to one of Kirk, Spock, or McCoy - the three central characters of Star Trek. The steps to extract and train the system were:

- Copy video from DVD to Hard Drive
- Extract Closed Captioning from video to text file
- "Tag" the script with who said what
- Convert the script to comma separated values
- Train the Simple Model
- Test Simple Model
- Gather statistics

We successfully extracted 30 scripts. However the process of tagging the scripts was more time consuming than expected. To tag a file required sitting in front of a television and editing the script as the characters spoke their lines. Often the video would have to be stopped while we caught up with the action. Even with advanced knowledge of the "Star Trek" series it took over two hours per episode to tag the files. Finally, episodes 2, 4, 6, and 7 were tagged.

Database

The resulting script was not easily parsed by the system. So a conversion to Comma Separated Values format was performed by custom software. This code applied heuristics to fix the split line problem mentioned earlier. The columns for the CSV were:

- Episode
- Start Timestamp
- Start milliseconds
- End Timestamp
- End milliseconds
- Speaker
- Dialog text

The CSV files were then converted to a relational database. The tables were:

- EPISODE – a listing of episodes and their associated information
  - *EPISODENO – epsisode number (primary key)*
  - AIRDATE – the date the episode aired
  - TITLE – the title of the episode
  - DVDNO – the disk it was sourced from
  - STARDATE – the stardate as stated in the show
  - CSV? - whether the CSV file was processed
  - MPG? - whether the MPG file was processed

- MPG – a listing of the MPG files that were extracted from the DVDs (episodes can span multiple files)
  - *EPISODENO* - primary key – the episode number*
  - FNAME - filename
  - STARTMILLIS – start position
  - ENDMILLIS – end position

- SCRIPT – the scripts as a single record per line
  - EPISODENO – episode number
  - STARTMILLI – start time
  - ENDMILLI – end time
  - SPEAKER – who was speaking
  - ABOUT – what they were talking about
  - DIALOG – original text
  - DIALOG2 – normalized text after expanding contractions and other heuristics
  - DIALOG3 – conversion of words to concepts

- GAPSCENE – a listing of scenes as determined by sound gaps
  - EPISODENO – the episode number
  - STARTMILLIS – start position
  - ENDMILLIS – end position

- WORDSCENE – scenes as determined by concept matching
  - EPISODENO – episode number
  - TOPIC – concept for the scene
  - STARTMILLIS – start time
  - ENDMILLIS – end time

(See the appendix for the salient SQL queries)

The dialog text was then normalized. In the SCRIPT table, the DIALOG field held one line of original dialog. The DIALOG2 field held the text after it had been corrected for contractions, pronouns, and stemming.

Each line of text was processed to expand contractions. For example, the contraction "I'm" was expanded to "I am".

Because the WordNet database (see below) did not recognize pronouns, certain pronouns were replaced. For example the pronoun "I" was replaced with "self" and "You" was replaced with "ewe". The results of using these alternatives are explained below.

Wordnet stores words in dictionaries of Nouns, Verbs, Adverbs and Adjectives. The words are grouped into Synonym sets (or Synsets). We converted each word of dialog into a concept by looking up the word in the WordNet database and translating it into a token that represents the SynSet that the word was found in.

POS	Unique Synsets	Total Strings	Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

*We used only Noun and Verb POS

The WordNet database holds the root words to thousands of English words. To facilitate looking up these words in the database the Porter Stemmer was used to convert the words to their root. For example "browsing" and "browse" became "brows". The Porter Stemmer is not guaranteed to create actual words. So, the entire corpus was processed through the stemmer. Each stem was then looked up in the WordNet database. If the word was not found it was sent to an exception file. The exception file was then processed by hand and each stem was given a proper root word. The exception file then became the root word translation table.

Each of these translations was applied to the DIALOG field and the result was stored in the DIALOG2 field.

The next normalization converted the DIALOG2 text into concepts. Each word was looked up in the WordNet database. WordNet stores only nouns, verbs, adjectives, and adverbs. The goal here was to reduce the original text into simpler concepts. To this end, only nouns and verbs were processed. Other words including articles and pronouns were thusly discarded.

See the Appendix for examples of text normalization

Segmentation

In the next phase we attempted to divide scenes based on the textual content of the videos. The premise is that scenes are a sequence of dialog lines that have related concepts. When a concept change is detected, this is flagged as a scene boundary.
To demonstrate this concept, three methods were implemented: Simple, Gap, and Concept mode.

In Simple mode, a keyword was looked up in the DIALOG field of the SCRIPT table. That line of dialog (and its associated video) were played back. The start of the playback was computed from the time stamp associated with the dialog.

In Gap mode, the same keyword lookup was performed and this time the closed caption text was scanned backwards until a gap of 2 seconds or more was detected. The gap was detected by finding a line of dialog that ended 2 seconds before the next line started. The same method was used to find the end of the scene.

Gap mode exposed a couple immediate problems. It was not uncommon for a scene to abut another scene without much of a pause between them. This caused the merging of scenes that had no relationship. Also, there are dramatic pauses that can last for more than two seconds. This creates a false break in the scene - especially when Jim Kirk is speaking.

In Concept Mode we attempt to group lines of dialog together based on their similar concepts. The entire corpus was scanned and scored. Each concept word from DIALOG3 in the DIALOG table looked up in the WordNet database. A counter for each SynSet on that line was then incremented. This was continued for each line of dialog in the root word corpus.

With each line of the corpus appropriately scored, we then made another pass over the DIALOG3 corpus to join lines together where the concepts in the lines were similar.

The lines were scored 13 lines at at time. For each grouping, like SynSets were combined. So, for example, if the concept "baby_doctor" appeared in the grouping, the sum of all the "baby_doctor" concepts were combined. The concept with the maximum score became the "concept" of the first line of the grouping. Each line of an episode was scored in this way.

Finally, groupings of lines were combined by detecting a change in the concepts of consecutive lines of the concept corpus. When a change in the concept was detected, a record was written to the "WORDSCENE" table recording the start and end timestamps of the scene and a keyword representing the concept of that scene.

It is interesting to note which scenes were matched with which keywords. For example, scenes in which Mr. Spock is the principle topic of conversation often are translated into the concept "baby_doctor". This seems comical upon first inspection. But upon reflection it makes perfect sense. There was a pediatrician named "Dr. Spock" who revolutionized the way we think of child rearing. Doubtless this name is in the WordNet dictionary and it confuses our "Mr. Spock" with "Dr. Spock" and the system assigns the concept to those scenes.

CHAPTER 5 : Results Analysis

Extraction

Once the scripts were loaded into memory, some simple statistics could be gathered. The frequency of occurrence of words in the text was computed. Of note is the word “captain” which was spoken 140 times in 4 episodes. This is clearly outside the norm for spoken text.

Speaker Recognition

For speaker recognition, the system uses frequency of occurrence to classify text from the scripts as being spoken by Kirk, Spock, or McCoy. The basic assumption is that Kirk will use a certain set of words more frequently than either McCoy or Spock. If we keep track of how many times Kirk uses these words, we can classify new text as spoken by Kirk. If the words in the new text are words that Kirk has used more frequently than Spock or McCoy, then we have some assurance that utterance is Kirk's.

A model is kept for each of the three speakers. In training mode, all words are read from a script into memory. In-line training is performed. As each line of dialog is interpreted, an entry for the word is made in the speaker's model. If the word is repeated by that speaker, a count for that word is incremented.

In test mode, a script is read into the computer line by line. For each model, a summation of the words' count is made. Then voting occurs. Whichever speaker received the most “points” determines which class the line of dialog is assigned to.

Accuracy is determined by comparing the expected result (from the CSV script) with the actual result from the Simple Model.

Several tests were performed. The first battery of tests simply trained with a script, then tested with the same script. This shows that the system is at least good at the trivial task of classifying the a priori values on themselves. Then, the system was trained on 3 scripts, and tested on the fourth.

Train	Test	Accuracy
2	2	84%
4	4	82%
6	6	78%
7	7	83%
-	-	-
4,6,7	2	61%
2,6,7	4	55%
2,4,7	6	52%
2,4,6	7	65%

Table 1: Speaker Training Results

As can be seen from the table, the system loses little resolution when testing against the same data it was trained from. A loss of around 18% accuracy is observed.

Also, from Table 1, when three scripts were used to train, the Simple Model scored between 52-65% . This is nearly twice as good as random guessing which scores 33%.

Three different options to Speaker Recognition were analyzed:

- Single Word Occurrence
- Other Speaker Inclusion
- Elimination of "noise" words.

Analysis of the single use of a word, regardless of how many times it was spoken was attempted. The idea here was to normalize the frequency of occurrence and determine if the mere presence of a word could be used to classify a speaker. Single word occurrence did not score as well as frequency of occurrence. This implies that frequent usage of a word by a speaker is significant when classifying.

There were many speakers other than Kirk, Spock, and McCoy in the Star Trek episodes. Allowing a fourth class of "other" could allow a more refined result set. In this case, all other speakers were classified into the "other" model. Attempts to use this method did not work well. The "other" speaker often won out over Kirk, Spock, or McCoy. And no wonder, it was as if everyone on the Starship Enterprise were outshouting the lead characters.

Words like "a", "of", and "but" may have biased or otherwise confused the frequency of occurrence model. The WordNet database of words was used to filter out all but nouns, verbs, adjectives, and adverbs.

Strangely, this method did not perform as well as Simple Model. It seems that even "noise" words are significant indicators of a speaker's speech patterns.

In each run, the options were used in binary fashion. First, none of the 3 options were used. Then each possible binary combination of the 3 options was tried. Table 2 is an example of the "2,4,6" run from Table 1, above. In all runs, Simple Model alone outperformed the alternatives (alone or in combination).

single	*others*	*exclusions*	*result*
FALSE	FALSE	FALSE	(freq. of occur.) 65%
TRUE	FALSE	FALSE	51%
FALSE	TRUE	FALSE	52%
TRUE	TRUE	FALSE	52%
FALSE	FALSE	TRUE	54%
TRUE	FALSE	TRUE	52%
FALSE	TRUE	TRUE	52%
TRUE	TRUE	TRUE	52%

Table 2: Comparison of alternative options from a run of scripts 2,4,6 in Table 1

An attempt was made at using the Hidden Markov Model (HMMPak from Troy McDaniel of Arizona State). The method would read all words from a test script into the model and use the words as a visible layer. A hidden layer with twice as many nodes as the visible layer was created. Batch mode training was performed using K-Means computation. It was thought that words in sequence would yield a strong accuracy model, similar in concept to speech recognition. Again, separate models were created for each of Kirk, Spock, and McCoy.

The results were disappointing. All the classifier results registered near 0%. Different initial values for the hidden layer affected the outcome. A difference in tense of words or word order caused the test data to score zero. A future project will revisit this method.

Segmentation

Using audible gaps in the sound track was a successful mechanism for determining scene boundaries. The scenes had a beginning, middle, and end. However, this method often yielded false starts and false endings as mentioned earlier.

When using the concept mode for scene segmentation we get some interesting results. Longer scenes are cut into smaller scenes because the concept that the scene was tagged which indicated the actual scene content, and not just sound gaps. The table below gives an example of scene breaks computed using the sliding window approach.

Scene Number	concept	Start Milli	End Milli	Scene Duration
1	person	6840	47130	40.29
2	politician	47130	69853	22.72
3	unwrap	69853	234300	164.45
4	undertaking	234300	235301	1
5	interpret	235301	269085	33.78
6	unwrap	269085	271087	2
7	interpret	271087	290657	19.57
8	unwrap	290657	299115	8.46
9	carry_through	299115	325975	26.86
10	shout	325975	327477	1.5
11	carry_through	327477	329479	2
12	shout	329479	407607	78.13
13	travel	407607	456072	48.47
14	shout	456072	558124	102.05
15	engineering	558124	642875	84.75
16	typify	642875	661227	18.35
17	engineering	661227	662728	1.5
18	happen	662728	726075	63.35
19	end	726075	742725	16.65
20	person	742725	768284	25.56
21	utter	768284	796479	28.2
22	carry_through	796479	820636	24.16
23	match	820636	881730	61.09
24	carry_through	881730	914179	32.45
25	match	914179	931830	17.65
26	chemical_element	931830	1038053	106.22
27	carry_through	1038053	1067950	29.9
28	engineering	1067950	1098197	30.25
29	person	1098197	1159591	61.39
30	carry_through	1159591	1271119	111.53
31	induce	1271119	1276408	5.29
32	person	1276408	1339388	62.98
33	evaluate	1339388	1344559	5.17
34	cognition	1344559	1392190	47.63
35	person	1392190	1435400	43.21
36	chemical_element	1435400	1446411	11.01
37	person	1446411	1462477	16.07
38	carry_through	1462477	1475073	12.6
39	travel	1475073	1531046	55.97
40	engineering	1531046	1582731	51.69
41	person	1582731	1596628	13.9
42	engineering	1596628	1818600	221.97
43	unwrap	1818600	1830728	12.13
44	chemical_element	1830728	1834649	3.92
45	person	1834649	1865563	30.91
46	chemical_element	1865563	1871019	5.46
47	person	1871019	1873521	2.5
48	chemical_element	1873521	1917115	43.59
49	kind	1917115	1970535	53.42
50	chemical_element	1970535	1984766	14.23
51	engineering	1984766	1987268	2.5

52	person	1987268	1994275	7.01
53	engineering	1994275	1996277	2
54	carry_through	1996277	2002784	6.51
55	chemical_element	2002784	2003785	1
56	carry_through	2003785	2088069	84.28
57	engineering	2088069	2099080	11.01
58	carry_through	2099080	2154635	55.56
59	chemical_element	2154635	2159140	4.51
60	carry_through	2159140	2177525	18.39
61	arouse	2177525	2195092	17.57
62	carry_through	2195092	2204101	9.01
63	arouse	2204101	2295226	91.13
64	person	2295226	2319383	24.16
65	kind	2319383	2321385	2
66	carry_through	2321385	2338319	16.93
67	engineering	2338319	2445426	107.11
68	person	2445426	2461392	15.97
69	flog	2461392	2539670	78.28
70	unwrap	2539670	2602032	62.36
71	cry	2602032	2606236	4.2
72	utter	2606236	2608322	2.09
73	approve	2608322	2611775	3.45
74	utter	2611775	2641772	30
75	approve	2641772	2661125	19.35
76	flog	2661125	2775939	114.81
77	carry_through	2775939	2835999	60.06
78	supply	2835999	2837501	1.5
79	carry_through	2837501	2851515	14.01
80	supply	2851515	2878175	26.66
81	chemical_element	2878175	2892856	14.68

Table 3: Scene Detection Results

CHAPTER 6 : Conclusions

The system has demonstrated a good start at scene detection using closed captioning. More work needs to be done to hone the classifier so that it recognizes a scene that is very short. Currently many scenes run 2-3 minutes or longer. It seems that a concept in a line of dialog 10 lines away can pollute an earlier one. For example, the word "ship" had a stronger score than the word "death." More research is necessary to determine how to refine the scene segmentation process.

The system performs quite well in this limited domain. The scene selections made by the "Word" mode are quite good compared to the "Gap" mode. Even the simple mode of keyword lookup provides a powerful research tool for those interested in querying a television series for quotes and quips.

The ability to scan the corpus and merge lines to form scenes based on concepts instead of keywords proved most powerful. The concept scenes generally made sense and were more logical than scenes created by Gap Mode.

Challenges

When concept counts were close concept counts were lose, scene detection oscillates between the two concepts making for choppy scene detection.

Some scenes in the episodes can be quite short. When this happens the system can be easily derailed, combining lines of dialog with other, unrelated scenes.

The mapping from words to concepts did not have the desired "smoothing" effect, although it did have the desired effect of discarding a number of "noise" words.

The window size of the grouping algorithm greatly affects scene detection and smoothness of the scenes.

Other Applications

This technology may find use in **Law Enforcement**. The analysis of video confessions is a time consuming operation. Especially where more than one perpertrator is involved. There is actually little difference between comparing two interviews for a family biography and two confessions. The use of concept matching could reduce the time to retrieve confession details and also correlate or detect contradictory facts.

List of References

[Alatan2001] Multi-Modal Dialog Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing. Multimedia Tools and Applications archive. Volume 14 , Issue 2 (June 2001). Pages: 137 - 151 ISSN:1380-7501. A. Aydin Alatan, Ali N. Akansu, Wayne Wolf

[Allan2002] J. Allan, Ed. 2002 *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers.

[Allan2000] First story detection in TDT is hard. Conference on Information and Knowledge Management archive. Proceedings of the ninth international conference on Information and knowledge management. Pages: 374 - 381. 2000. ISBN:1-58113-320-0. James Allan, Victor Lavrenko, Hubert Jin

[Browne2004] Browne, P. and Smeaton, A. F. 2004. Video information retrieval using objects and ostensive relevance feedback. In*Proceedings of the 2004 ACM Symposium on Applied Computing* (Nicosia, Cyprus, March 14 - 17, 2004). SAC '04. ACM, New York, NY, 1084-1090. DOI= http://doi.acm.org/10.1145/967900.968120

[Chen2002] Motion Activity Based Shot Identification and Closed Caption Detection for Video Structuring. Lecture Notes In Computer Science; Vol. 2314 archive. Proceedings of the 5th International Conference on Recent Advances in Visual Information Systems. Pages: 288 - 301. 2002. ISBN:3-540-43358-9 . Duan-Yu Chen, Shu-Jiuan Lin, Suh-Yin Lee

[Colace2005] A Probabilistic Framework for TV-News Stories Detection and Classification, Colace, F. Foggia, P. Percannella, G. DIIIE, Univ. di Salerno, Fisciano; Multimedia and Expo, 2005. ICME 2005. pgs1350-1353

[Fan2004] Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing. Jianping Fan; Hangzai Luo; Elmagarmid, A.K. Image Processing, IEEE Transactions on Volume 13, Issue 7, July 2004 Page(s):974 – 992 Digital Object Identifier 10.1109/TIP.2004.827232

[Fleischman2007] Michael Fleischman and Deb Roy. (2007). Situated Models of Meaning for Sports Video Retrieval. HLT/ACL 2007, Rochester, NY. pdf (293K)

[Gauch1999] Real time video scene detection and classification. Information Processing and Management: an International Journal archive. Volume 35 , Issue 3 (May 1999) Pages: 381 - 400. ISSN:0306-4573. John M. Gauch, Susan Gauch, Sylvia Bouix, Xiaolan Zhu

[Hampe1997]Making documentary films and reality videos: a practical guide to planning, filming, and editing documentaries of real events. By Barry Hampe Edition: revised Published by Macmillan, 1997 ISBN 0805044515, 9780805044515 342 pages

[Han2002] An integrated baseball digest system using maxium entropy method. International Multimedia Conference archive. Proceedings of the tenth ACM international conference on Multimedia Pages: 347 - 350. 2002. Mei Han, Wei Hua, Wei Xu, Yihong Gong

[Hangalic2005] Adaptive extraction of highlights from a sport video based on excitement modeling. Hanjalic, A. Multimedia, IEEE Transactions on Volume 7, Issue 6, Dec. 2005 Page(s): 1114 - 1122

[Hauptmann1998] Hauptmann, A. G. and Witbrock, M. J. 1998. Story Segmentation and Detection of Commercials in Broadcast News Video. In *Proceedings of the Advances in Digital Libraries Conference* (April 22 - 24, 1998). ADL. IEEE Computer Society, Washington, DC, 168.

[Kumaran2004] Text classification and named entities for new event detection. Annual ACM Conference on Research and Development in Information Retrieval archive. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Pages: 297 - 304. 2004. 1-58113-881-4. Giridhar Kumaran, James Allan

[Lyu2005] A comprehensive method for multilingual video text detection, localization, and extraction. Lyu, M.R.; Jiqiang Song; Min Cai. Circuits and Systems for Video Technology, IEEE Transactions on. Volume 15, Issue 2, Feb. 2005 Page(s): 243 - 255

[Merlino1997] Broadcast news navigation using story segmentation. International Multimedia Conference archive. Proceedings of the fifth ACM international conference on Multimedia. Pages: 381 - 391.1997 ISBN:0-89791-991-2. Andrew Merlino, Daryl Morey, Mark Maybury

[Namkamura1997] Semantic analysis for video contents extraction—spotting by association in news video. International Multimedia Conference archive. Proceedings of the fifth ACM international conference on Multimedia. Pages: 393 - 401. 1997. ISBN:0-89791-991-2. Yuichi Nakamura, Takeo Kanade

[Shahraray1995] Automated authoring of hypermedia documents of video programs . International Multimedia Conference archive. Proceedings of the third ACM international conference on Multimedia. Pages: 401 - 409. 1995. ISBN:0-89791-751-0 Behzad Shahraray, David C. Gibbon

[Stokes2001] Combining Symantic and Syntactic Document Classifiers to Improve First Story Detection, Nicola Stokes, Joe Carthy, University of Dublin

[Stokes2001] First story detection using a composite document representation Human Language Technology Conference archive. Proceedings of the first international conference on Human language technology research. Pages: 1 - 8. 2001 Nicola Stokes, Joe Carthy

[Wiki2008a]Wikipedia.org "NTSC" http://en.wikipedia.org/wiki/Closed_captioning

[Wiki2008b]Wikipedia.org "Closed Captioning" http://en.wikipedia.org/wiki/NTSC

APPENDIX I : SQL QUERIES

Select all lines of dialog from an episode

```
SELECT * FROM SCRIPT WHERE EPISODENO = ? ORDER BY
STARTMILLI
```

Select all episode numbers

```
SELECT DISTINCT EPISODENO FROM SCRIPT
```

Search for all lines of dialog with a certain word in it

```
SELECT EPISODE.EPISODENO, EPISODE.AIRDATE, EPISODE.TITLE,
MPG.FNAME, MPG.STARTMILLIS, MPG.ENDMILLIS,
SCRIPT.STARTMILLI, SCRIPT.ENDMILLI, SCRIPT.DIALOG
FROM EPISODE, MPG, SCRIPT
WHERE
     EPISODE.EPISODENO = MPG.EPISODENO AND
     EPISODE.EPISODENO = SCRIPT.EPISODENO AND
     SCRIPT.DIALOG LIKE '%${KEYWORD}%' AND
     MPG.STARTMILLIS <= SCRIPT.STARTMILLI AND
     SCRIPT.ENDMILLI<= MPG.ENDMILLIS
ORDER BY EPISODE.EPISODE, SCRIPT.STARTMILLI
```

Select all topic information on an episode

```
SELECT EPISODE, TOPIC, STARTMILLIS, ENDMILLIS FROM WORDSCENE
WHERE EPISODE = ?
```

APPENDIX II : NORMALIZTION OF TEXT

oh, captain...
oh captain
oh:American_state
captain:commissioned_military_officer,head,lead

got a minute?
got a minute
got:
a:metric_linear_unit minute:time_unit,unit_of_time

a minute.
a minute
a:metric_linear_unit
minute:time_unit,unit_of_time

it"s spock.
it is spock
it:engineering,engineering_science,applied_science,technologyis:
spock:baby_doctor,pediatrician,pediatrist,paediatrician

you noticed anything strange about him?
ewe notice anything strange about him
ewe:African
notice:announcement,promulgation,sight
anything:thing
strange:
about: him:

no, nothing in particular. why?
no nothing in particular why
no:negative nothing:relative_quantity
in:linear_unit
particular:fact why:reason,ground

well, it"s nothing i can pinpoint
well it is nothing self can pinpoint

well:excavation,surface,come_up,rise_up,rise
it:engineering,engineering_science,applied_science,technology
is:
nothing:relative_quantity
self:consciousness
can:container,preserve,keep
pinpoint:moment,minute,second,instant,locate,turn_up

without an examination
without an examination
without:
an:associate_degree,associate
examination:investigation,investigating

APPENDIX III : EXAMPLE SCENE DETECTION

oh	American_state
captain	commissioned_military_officer
got	
a	metric_linear_unit
minute	time_unit
it	engineering
is	
spock	baby_doctor
ewe	African
notice	announcement
anything	thing
strange	
about	
him	
no	negative
nothing	relative_quantity
in	linear_unit
particular	fact
why	reason
well	excavation
it	engineering
is	
nothing	relative_quantity
self	consciousness
can	container
pinpoint	moment
without	
an	associate_degree
examination	investigation
but	
he	chemical_element
is	
become	change_state
increasingly	
restive	
if	
he	chemical_element
were	
not	
a	metric_linear_unit
vulcan	Roman_deity
self	consciousness
would	
almost	
say	opportunity
nervous	
and	
for	
another	
thing	situation
he	chemical_element
is	
avoid	prevent

food	substance
self	consiousness
check	draft
he	chemical_element
ha	angular_distance
not	
eaten	
at	chemical_element
all	
in	linear_unit
three	digit
day	time_unit
well	excavation
that	
just	
soun ds	
like	kind
mr	title
spock	baby_doctor

APPENDIX IV : EXAMPLE SCENE DETECTION – PART 2

chemical_element	5
consciousness	3
excavation	2
metric_linear_unit	2
time_unit	2
engineering	2
baby_doctor	2
linear_unit	2
relative_quantity	2
commissioned_military_officer	1
angular_distance	1
American_state	1
kind	1
fact	1
container	1
digit	1
draft	1
moment	1
substance	1
thing	1
negative	1
opportunity	1
prevent	1
investigation	1
announcement	1
associate_degree	1
change_state	1
African	1
title	1
reason	1
Roman_deity	1
situation	1

The first line is scored as 'chemical_element' because the words 'he' and 'at' are used frequently

Note that pronoun trouble has colored our scoring. 'He' is not intended as 'helium' but rather as the subject 'Spock'.

APPENDIX V : EXAMPLE SCENE DETECTION – PART 3

Scene	Concept	Start	End	Duration
0	excavation	11044	15331	4.29
1	engineering	15331	16332	1.00
2	excavation	16332	25392	9.06
3	consciousness	25892	33683	7.79
4	American_state	33683	45278	11.60
5	consciousness	45278	101718	56.44
6	African	101718	107474	5.76
7	person	107474	110844	3.37
8	African	112228	114230	2.00
9	linear_unit	114731	115515	0.78
10	linear_unit	131831	135335	3.50
11	linear_unit	140924	156439	15.52
12	linear_unit	199732	202068	2.34
13	African	202068	206072	4.00
14	consciousness	206573	207574	1.00
15	African	207574	214581	7.01
16	consciousness	214581	218084	3.50
17	African	218084	220587	2.50
18	consciousness	220587	229095	8.51
19	consciousness	233099	247914	14.82
20	consciousness	251784	254621	2.84
21	commissioned_military_officer	265598	280363	14.77
22	commissioned_military_officer	286653	297163	10.51
23	commissioned_military_officer	303086	310810	7.72
24	commissioned_military_officer	315565	316566	1.00
25	commissioned_military_officer	324073	347180	23.11
26	chemical_element	347180	367700	20.52
27	dramatist	367700	373706	6.01
28	consciousness	373706	378211	4.51
29	chemical_element	378211	384384	6.17
30	title	384384	416332	31.95

VITA

Gregory Lee Smith was born in Limestone Maine, on 4 December, 1962, the son of Victor L. Smith and Johanna L. Smith. After completing his work at Lake George High School (Lake George, NY), he went on to Rensselaer Polytechnic Institute where he studied Computer Science and received his Bachelor of Science in December 1985. For the next 24 years he pursued a career in Software Engineering, building software systems for such corporations as General Electric Aerospace, TV Guide, and Genworth Financial. He moved to Richmond, VA in September 2000, and in September 2007 entered The Graduate School at Virginia Commonwealth University in Richmond, VA.