1 Abstract

Libraries often have large archives of legacy video recordings that lack reliable segmentation and metadata for efficient research, known as the "digital silo" problem. We propose a state-of-the-art multimodal method that combines visual, audio and text cues to automatically segment TV news videos by their natural program boundaries and generate metadata such as topics.

2 Introduction

While there has been extensive research on story-level segmentation of TV news programs, to our knowledsge, program-level segmentation has not been studied so much yet.

Intuitively story-level boundaries must be easier to classify as story, by definition, is a logical group of frames based on specific topics. TV news programs may not necessarily be. The task of dividing TV news recordings by their program boundaries can be challenging even for humans. For instance, CNN had relatively undifferentiated programming back-to-back for hours at a time in the 90s and it can be difficult to locate program boundaries. That said, it may be safe to assume program boundaries are a subset of story boundaries, making story boundaries as decision units for final binary classification for program boundaries.

It may be possible to directly classify program boundaries without finding story boundaries. This avenue is to be explored as the project evolves. Another reason using story boundaries as a program boundary candidate may be a good idea is because the existing caption files have marks that indicate the start of stories and commericals (e.g. SEG_0 type=Story starts). Such marks can be thought of as free labelled samples for training a story boundary classifier.

TBC...

3 Features

In this section, we list features that we believe to have some discriminative power for or correlations with program/story boundaries. Note this is just a tentative laundry list of features to be explored in more detail. Some of the useless features will be pruned in the future. Also notice none of the features will be strong enough to classify program boundaries as a standalone feature. Therefore the key to success would be to develop learning algorithms to fuse the multimodal features properly.

At a high level, some of the features will be used for story segmentation while others may be directly used for program segmentation. There will be three

separate feature extration streams: video, audio, and text. The streams flow independently of each other and only at a final stage will they be combined.

3.1 Video Features

3.1.1 Anchor Shot Detection

We take advantage of the fact that the scenes that include anchors (or their faces) are correlated with story transitions. We sample a frame at regular or noisy intervals and detect human faces. We then run a K-mean clustering method on the detected faces. The dominant cluster centroid and faces assigned to the class of the cluster most likely will be anchor faces. We assume anchor faces most likely appear most often, given a long time horizon (say 1 hour).

Also another visual cue that hints at story transition is, given an anchor shot, to check whether the location of the anchor face (bounding box of the face) is tilted to the left or right, deviating from the center of the frame. When story transitions take place in news programs, it is common that an anchor face is placed to the left with an image of the next story over the anchor's shoulder.

TBC...

Most shows have consistent anchors or presenters in unique sets (again, with some possible re-use within the same channel and no re-use across channels). They might appear as guests on other shows or special coverage, but their appearance on their own show should be consistent in format and repetition.

Shot Connectivity Graph [anchor]→[story1]→[anchor]→
###Black Junk Frames It has been reported in some literature that commercials or news stories are followed by one or two consecutive frames that are completely black/blank. Such black frames do appear in the middle of stories or in other places.

3.1.2 Shot Boundaries

In TRECVID dataset, 94+% of the story boundaries appear in the vicinity of a shot boundary. The same would likely go for program boundaries. Rather than computing this as a binary feature, we can take this as a continuous real-valued score to be used for story boundary detection.

3.1.3 Logo/Title Detection

Most shows have a distinctive set of logo and title visuals. Logo or title candidates can be detected by locating a region whose pixels do not change over many frames.

For more details, https://www.hindawi.com/journals/ijdmb/2012/732514/ details ###Visual Motion Acitivity We can measure how much video is changing with a color pixel difference tracking method, using the percentage of pixels that have changed color between it and the previous frame.

image saliency or tranformed frames in Fourier domain may be useful...

3.1.4 Credit Information

closing credit info screen. copyright/closing credit info

3.2 Audio Features

audio (mp3) can be extracted from original mp4 files using avconv. feeding video into an audio-related feature extractor would be computationally expensive.

3.2.1 Speaker Detection

Changes of a speaker may be correlated with story/program transition. There are a few tools to detect the speaker change, notably based on normalized cross likelihood ratio (NCLR).

3.2.2 Silence Detection

It has been shown long audio pauses are strong indicators of story transitions. This audio feature capitalizes on the pattern that an achorperson pauses before moving on to introduce a new story. One way to estimate the pause duration is to track a maximum time period where volume at timestep t goes below some reference threshold (e.g. fixed/rolling average volume of a given newstream). Such maximum low-volume period is calculated within a regular sampling interval and defined to estimate a pause period.

The long pause period together with some common transition/closing/opening lines may be very useful. Then again, this feature may not be so useful, as opening and closing transitions usually coincide with theme music.

3.2.3 Transition Music

TBC...

Distinctive theme songs (might be the same across broadcasts within a channel [i.e. same music might be used for KABC's 5 p.m. and 6 p.m. newscasts, but that theme will be different than KNBC uses for their newscasts]). There should

be high stability in these theme songs within a given season and network, but possible variation across them.

Using this feature ...

3.3 Text Features

While a large proportion of the video data come with the caption files, there are still many video files that do not have one. Some have one but the caption file is just incomplete. For the caption-lacking videso, we can consider two approaches: + *Automatic Speech Recognition on audio files* Video captioning through Densecap

ASR may be the way to go for the input compatibility. Both will generate text. ### lexical simliarity & chain strength

http://lxie.nwpu-aslp.org/papers/2012-IEICE-WangXX-A2-SCI-EI-JNL.pdf

3.3.1 Bag Of Words Histogram Distance

Words used for one story may be difficult from those used in another story. We build a vocabulary/dictionary vector out of commonly used words along with the existing caption samples and tally the word occurences within a regular time interval. The result will be a Bag-of-Words histogram. We then calculate the chi-square distance between the two consecutive BoW histograms. If the distance will be real-valued, continuous.

3.3.2 Topical Distance

We utilize the fact that story is a topical group of frames. Two popular methods used for topic modelling are non-negative matrix factorization (NMF) and Latent Dirichlet Allocation (LDA). We can divide a caption (8hour-long in video) into small documents (say 5minutes or some sensible average duration of stories), and run a trained LDA model on the documents which will produce outputs simliar to:

doc1: topic A 50%, topic B 30% ...

we can then calculate a L1/L2 distance between the topic mixtures of each documents. If there's a significant difference, it could strongly indiciate a story transition or even news program transition. (commericals will likely have distinct topic mixtures)

One challenge is to choose the number of topics in advance, which is a hyper parameter.

It seems LDA is going out of fashion and may consider other state-of-the-art methods instead. https://datascience.stackexchange.com/questions/678/what-are-some-standard-ways-of-computing-the-distance-between-documents

3.3.3 Transition Markers

There seem to be a few keywords that appear recurringly in the caption files and they seem to be correlated with program boundaries.

Such keywords (case insensitive) are: caption, commercial, story, SEG. *caption: caption by, captioning by...*
commerical: type=commercial. marks the start of commercial *story: marks the start of story segment* SEG: marks a story transition.

However, this applies only to videos with the captions. Moreover, some caption files are incomplete (like in 1972-01-07) or don't have the keywords where they should be (=a new story starts and there's no SEG type=story.) This skippy behavior applies to all the keywords. Keywords alone are incomplete. e.g. 2006-06-13_0000_US_00000141_V11_MB12_VHS13_H2_JK.txt3

cc-keyword-script is used..

3.3.4 Transition Phrases

There are repeating lines highly correlated with the story/program boundaries such as greetings. We can manually build a collection of reliable transition phrases or find one from the web. names of the stations like CNN, CBS

3.3.5 VCR Index

This is a hacky feature that may be useful for the given dataset, but won't generalize over other news program datasets. The indices of VCRs used for recording are specified in the filenames. The indices may give hint for channel names and recurring structure of programs.

For example, V2 seems to record NBC channels on a regular recording schedule. V2_2006_03_01 may have a simliar structure as the next day V2_2006_03_02 (for daily programs) and a week later V2_2006_03_08 (for weekly programs)

v0: v1: WCBS-TV (cbs) *v2: NBC* v3: ABC *v4:* v5: *v6:* v7: *v8:* v9: *v10:

+The VCR index to TV channel mapping seems preserved throughout the years. Otherwise, using this feature may not be a good idea.

4 Multimodal Fusion

. . . #System Overview input preprocessing using pretrained image classifier (producing visual feature map. . .)

5 Experiments

5.1 Data

https://www.slideshare.net/RJIonline/newsscape-preserving-tv-news

5.2 Inputs

We currently have: *video (mp4)* audio: none. can easily be extracted into mp3 or other audio formats. *text: caption files (srt,txt3,tx4) for videos recorded roughly since 2000. we can use ASR or Video captioning tools to generate captions for the videos that do not have caption files yet.

5.3 Evaluation Metrics

5.4 Solution Model

6 Conclusion

6.1 Challenges

That said, some networks have less distinctive show boundaries. CNN, for example, had relatively undifferentiated programming for hours at a time in the 90s. In more recent years they have a clearer signal of different shows.

6.2 Further Work

- shot boundaries may not be useful (just like phonetics was not useful in speech recognition)=
- smallest decision unit: segment based on shot boundaries or histogram difference between two frames, entropy, video saliency
- medium decision unit: segment based on stories and learn to combine them to a program. . . anchor shot. . .
- largest decision: unit program

three separate input streams * audio-> feature map * text-> feature map * video-> feature map #Related datasets transfer learning possibilities

NIST http://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html #Related Work

6.3 not important

http://www.cs.cmu.edu/~mehrbod/SSeg07.pdf

https://pdfs.semanticscholar.org/5c21/6db7892fa3f515d816f84893bfab1137f0b2.pdf existence of blank frames

http://www.cs.cmu.edu/~mychen/publication/duygulu_ICME04.pdf existence of blank frames

https://books.google.co.kr/books?id=nCnSy5XXdygC&pg=PA361&lpg=PA361&dq=boundary+segmentation+

http://mmlab.ie.cuhk.edu.hk/archive/2002/CSVT02_Video.pdf http://www.cstr.ed.ac.uk/downloads/publicatio http://www.bcs.org/upload/pdf/ewic_im99_paper3.pdf

6.4 important

best performing CRF and nice summary of features used http://lxie.nwpu-aslp.org/papers/2012-IEICE-WangXX-A2-SCI-EI-JNL.pdf

useful handcrafted features are listed: http://www1.cs.columbia.edu/~smaskey/candidacy/cand_papers/merlino_

https://pdfs.semanticscholar.org/41ed/c1f04cef2af8aa112642a0d3fdc36a395dda.pdf the appearance of an anchor person, audio pitch jump and significant audio pauses

Hsu et al. used a maximum entropy objective to select the most informative mid-level audio and video features and demonstrated an optimal feature fusion method. http://www.ee.columbia.edu/ln/dvmm/publications/03/icme2003pr.pdf

nice summary of related work http://crcv.ucf.edu/papers/civr2005_zhai.pdf

pretty recent and relevant http://cs229.stanford.edu/proj2012/DaneshiYu-BroadcastNews%20StoryBoundaryDetectionUsingVisual,AudioAndTextFeatures.pdf

nice https://www.hindawi.com/journals/ijdmb/2012/732514/

automatic speaker diarization (segmenting audio based on speaker identities) http://www.quaero.org/media/files/bibliographie/bredin_segmentation_of_tv_icassp2012.pdf

a lot of relevant papers http://mklab.iti.gr/publications

object-level detection using caffe https://arxiv.org/abs/1504.06201

Deeplearning https://arxiv.org/pdf/1601.07754.pdf

CNN https://arxiv.org/pdf/1705.08214.pdf http://imagelab.ing.unimore.it/imagelab/pubblicazioni/2015ACMM_

6.5 not sure

3d cnn http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Ng_Beyond_Short_Snippets_20

SBD using CNN https://arxiv.org/abs/1705.08214

topic-based segmentation and indexation of audio transcripts http://www.inesc-id.pt/pt/indicadores/Ficheiros/1146.pdf

SVM, boosting http://www.ee.columbia.edu/ln/dvmm/publications/04/hsu04generative.pdf

In later works, Hsu et al. investigated alternative discriminative models, i.e. Support Vector Machine (SVM), and showed a performance improvement when combining maximum entropy with SVM. http://www.ee.columbia.edu/~lyndon/pubs/spie2004-seg.pdf

Gao et al. [7] combined syntactic and semantic methods for segmentation using an unsupervised learning method. http://mmlab.ie.cuhk.edu.hk/archive/2002/CSVT02_Video.pdf

used CNN to generate tags http://medialab.sjtu.edu.cn/publications/2015/2015_BMSB_Wenjing.pdf

used ANN to do SBD https://github.com/MaxReimann/Shot-Boundary-Detection/blob/master/paper/SBD-Approach-Paper.pdf

maximum figure-of-meritlearning approach http://www.mirlab.org/conference_papers/International_Conference

novel shot boundary detection http://www.cai.sk/ojs/index.php/cai/article/viewFile/185/156

definition of scenee http://videoanalysis.org/Prof._Dr._Rainer_Lienhart/Publications_files/MTAP2001.pdf

scene segmentation metrics http://mklab.iti.gr/files/csvt12_preprint.pdf

6.6 reference implementations

6.6.1 semantic image segmentation

https://github.com/gberta/HFL_code

6.6.2 shot and scene detection

http://mklab.iti.gr/project/video-shot-segm https://github.com/Breakthrough/PySceneDetect
*http://johmathe.name/shotdetect.html

6.6.3 shot and key frame generation

https://github.com/yahoo/hecate https://github.com/andrefaraujo/videosearch

6.7 Todos

http://mklab.iti.gr/files/csvt11_preprint.pdf

7 Todos

index of cutfiles and whether it's complete or not get a decent definition of a program. commercials may exist within the same program. for e.g. news -> commercial -> short weather news -> commercial should this count? (e.g. 2006-06-13_0000_US_00000433_V5_MB13_VHS14_H1_MS.txt3 ~3h15m) * file index (cutpoints, caption avaiable) *shortage of labelled samples: build a decent working classifier (that achieves 90+%) and let it generate samples. The supervising samples are considered labelled but extremely noisy so a noise-robust DNN model can learn from it.* noise (consistent, e.g 2006-06-13_v11) *just take this a single-frame image classification task? (but a stack of frames as input)* massive model -> how to reduce its size ... and apply a simpler model