HOW AUTOMATIC CONTENT ANALYSIS ENABLES THE ENTERNAIMENT EXPERIENCES OF THE FUTURE

Jan Neumann[1]
Comcast

Abstract

Metadata has always been one of the essential ingredients to help customers find something to watch. It varies from the TV Guide with high-level descriptions to the extended descriptions of recommended items.

These days we have machine learning driven content discovery experiences and clip-based navigational capabilities which rely on the availability of descriptive and semantically meaningful metadata.

Unfortunately, the availability of the metadata is limited due to the high cost of creating such metadata which require significant amount of human supervision. We show how automatic content analysis combines video, audio and text processing with machine learning algorithms to identify relevant moments or temporal segments and their descriptions without or with very limited human interaction.

INTRODUCTION

Nowadays entertainment is consumed in many different ways. Besides the traditional linear consumption of live TV, we regularly enjoy entertainment on our own schedule via DVR recordings or on-demand content services. Vast amount of content is provided to consumers, thus we want the applications we use to consume videos to guide and assist us in a personalized way while we are looking for something to watch that we like and that fits our current mood.

In addition, due to increased consumption of content on mobile devices, we also often do not want to consume whole shows anymore, but only the "best" parts, i.e. the segments that are (or our social networks deems to be) most relevant and interesting to us.

To enable such personalized entertainment experiences, we need to know more about the content shown than just the metadata at the asset level such as titles, credits, keywords. More information includes also semantic segments and moments of a program, who is on screen and when (Figure 1), the theme or topic of the segment, and how relevant a segment is. In addition, if the information is aligned with the timeline of the video, it allows for advanced navigation within and between assets and can be indexed to enable relevant search and recommendations within videos.

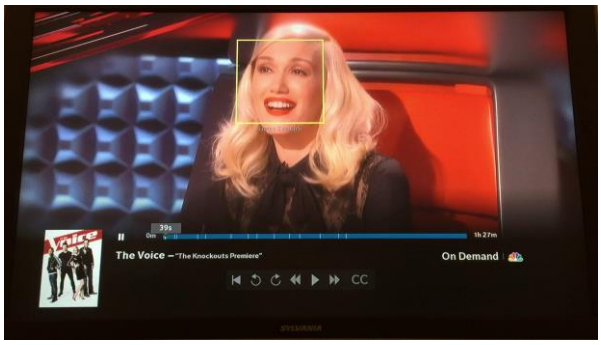

Figure 1 - Face Recognition in TV Shows

Moment and segment based metadata has many applications, such as chaptering of shows for improved navigation, enhanced search within captions and speech from audio so that search interfaces like the Xfinity Voice Remote allow users to search for specific phrases within movies or shows (e.g. "Life is like a box of chocolates").

We can also extract the current game state from the video at regular intervals, and use that information to align the video to an external

[1] The assistance of Hongcheng Wang and Jonghyun Choi during the preparation of this document is greatly appreciated!

meta-data sources such as sports play-by-play information as seen in Figure 2.

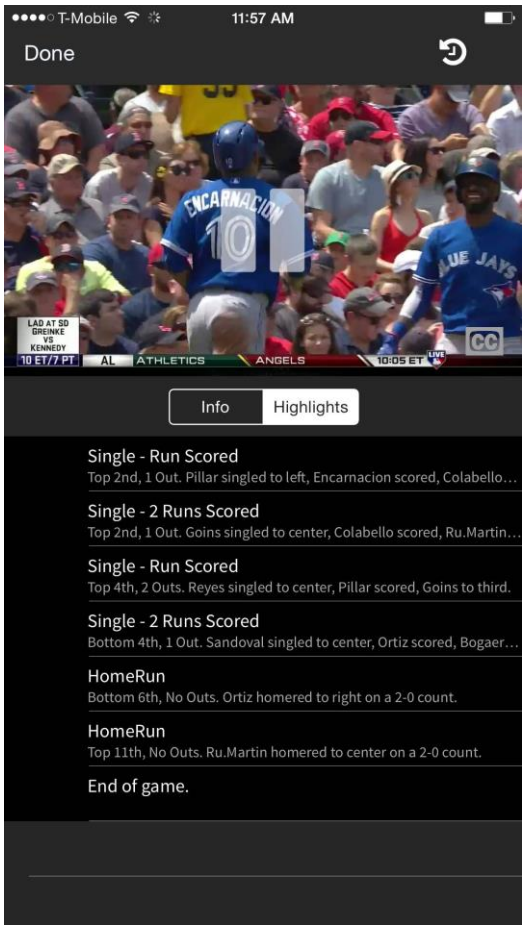

Figure 2 - Mobile App Allowing User to Select Playback of Individual Highlight Plays

Finally, automatic content analysis can be used to classify the activity and theme happening within segments of a show, it can be used to automatically classify content as suitable for various audiences by using adult content and language recognition [Ries], and object and content-based video compression can also impact the coding efficiency of the video pipeline positively.

The availability of such metadata, however, is limited due to the high cost of creating such metadata with human input or supervision. Metadata can be crowd-sourced as on sites such as YouTube or Facebook by analyzing the tags and comments from consumers themselves. This user generated data is usually unconstrained and unstructured though and it requires sophisticated algorithms to make good use of it.

We show how automatic content analysis combines video, audio and text processing with machine learning algorithms to identify relevant moments, segments and their descriptions without or with very limited human interaction. The generated data can then be used to provide consumers with a more interactive and personalized experience. Examples are more accurate program (segment) recommendations, in and between program navigation, and enhanced search capabilities including free-form queries using voice interfaces.

We also describe the details of how we use automatic content analysis techniques to create metadata for navigating from highlight to highlight within DVR recordings of sports broadcasts (Figure 2), organizing news segments according to their topics (Figure 5), and recognizing the persons currently visible in a scene similar to Amazon's X-ray feature, see Figure 1.

WHY DO WE NEED AUTOMATIC CONTENT ANALYSIS?

The availability of metadata is a fundamental building block to help customers interact with content and engage with it at a deeper level.

Unfortunately, the content metadata that is provided by the programmers and third-party metadata providers is often limited, i.e. it only contains high-level tags and descriptions that apply to assets as a whole. This is due to the amount of time it takes to create accurate descriptions and segmentations of content manually.

For live shows where the actual content is not known with sufficient details before the actual broadcast, any metadata would need to be generated on the fly if we are interested in more granular data than is available via the guest list or the genre information.

Companies like Google (YouTube), Amazon, Netflix, Apple, Pandora and Facebook rely heavily on accurate content descriptions to recommend related content so that their customers always have something else to consume while staying in the same eco-system.

Amazon and Netflix have dedicated groups of "movie watchers" that annotate their respective video on-demand catalogs, so that the content discovery algorithms can utilize this information and deliver relevant suggestions, accurate content descriptions and fine grained browse and search capabilities.

Pandora is famous for their collection of 1.3M annotated music tracks which they label as the "music genome" and is the backbone of their recommendation system.

All these companies use automatic content analysis techniques to speed up the generation of their metadata. They use automatic face detection and tracking technology to help their annotators identify the actors within scenes, and are developing techniques to automatically label the scene content in user-generated videos and images [Venugopalan]. Even Pandora who prides themselves on the superior quality of their metadata due to highly trained annotators, uses "machine listening" approaches to scale the annotation process.

As mentioned, another important application of content analysis is temporal metadata that labels contents at the moment or segment level, e.g. ad break detection, when someone appears on screen, when a touchdown is scored, etc.

Navigating through a video to manually annotate it in a frame accurate manner without algorithmic assistance is extremely laborious and time intensive since we can only navigate linearly through a video with standard video editing tools.

Automatic content analysis is very beneficial in this context because it can automatically identify potential annotation locations with associated confidences and present them to the annotator. This transforms the annotation task into a verification and selection task which can be performed much more efficiently by a human.

Finally, automatic tools can also be coupled with crowd-sourcing approaches like Amazon Mechanical Turk to scale up annotation resources on demand.

HOW DOES THE TECHNOLOGY WORK?

In this report we define automatic content analysis as a combination of computer vision, audio analysis, natural language processing and machine learning algorithms that allow the generation of semantic information for media assets. These analytics capabilities can be applied as a batch process to existing content catalogs (e.g. video on-demand content) and in real-time to linear broadcast streams. Many of the algorithms rely on machine learning techniques such as classification and clustering, and lately deep learning approaches.

Visual Analysis Techniques

Visual analysis is critical to detect and recognize text, objects and logos visible in the scene, segment programs, and classify segments. The standard approach to detect objects in an image and determine the concept category they belong to is usually done in stages.

During the training stage we apply image processing operators such as line, corner or histogram of gradients to create representations or features of examples for each category. Machine learning algorithms then use these representations to train a model that maps from features to categories using either classification or clustering approaches or a combination of the two.

During the detection phase we then process selected frames of the incoming video, apply the trained model to the resulting feature

representations and return the object categories with the highest score according to the model.

Implementations differ in the type of representations chosen, how the data is being stored (raw features or hashed representations), and how the representations are classified, e.g. different classification algorithms such as K-means clustering, support vector machines, neural networks, or random forests. More information about these approaches can be found in classic text books such as [Bishop].

One important aspect of object detection and recognition is to recognize the persons on screen in shows as shown in Figure 1. The first real-time computer vision face recognition system was developed by [Viola & Jones] and a variant of their approach is the basis for the face detection algorithms used in nearly every consumer camera these days. They combined the use of simple features that are easy to use with an efficient learning and classification algorithm that minimizes the number of computational steps necessary for the classification.

Face recognition for content analysis is different from this canonical approach because in contrast to the frontal views that we see in passport portraits we want to identify actors or people on screen who might only be partially visible, are looking away from the camera or are partially in the shadows. This scenario is often described as the "face recognition in the wild" problem [Huang]. At the same time we also have the advantage that for the celebrities that we want to detect in the videos, many images are already available to create accurate models.

Furthermore, since the same actors appear frequently in the same movie or across different episodes of a series, we can improve our results by labeling complete tracks of detected faces and use global constraints such as that the same actor cannot appear in the same scene simultaneously in two different image locations to find a globally optimal labeling solution.

A special variant of object recognition is optical character recognition, OCR in short, where the goal is to read the text that is present on the screen at a given time. This is helpful to establish the context of a show as in news shows, or it can be used to establish timing information in sports broadcasts by analyzing the game clock. Since most text in broadcast videos is artificially inserted as part of the post-processing, we can take advantage of such constraints as increased contrast, static horizontal text on a moving background and canonical size of the fonts. Furthermore, we can also utilize language models and dictionaries to constrain what sequences of characters are likely to appear next to each other (e.g. words, names).

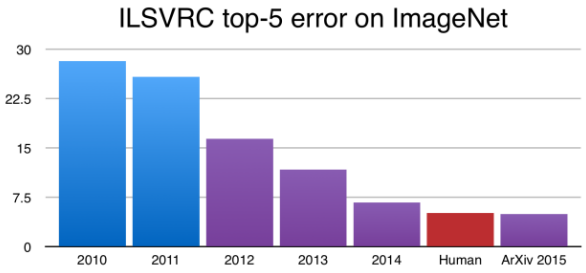

Figure 3 - Improvements in the Accuracy of Automatic Large-scale Object Recognition Algorithms over the Last Years [He]

Advances in deep learning [Bengio] have recently achieved close to human-level accuracy which allows for the automatic creation of metadata with very high accuracy [Krizhevsky].

Figure 3 shows the improvement of large scale object recognition performance for the ImageNet Large Scale Vision Recognition Competition which evaluates how accurately object class recognition algorithms can correctly identify 1 out of 1,000 categories for more than 1.2M images.

The large improvements starting in 2012 are due to the use of deep learning algorithms where the combination of large amounts of data with extensive computing capabilities have moved the field forward at a rapid pace. Recent deep learning based approaches perform face recognition beyond human capabilities [Huang].

Deep learning has not just applications for the classification of static images but can also

be applied to video sequences and is the gold standard for speech recognition. Recently, researchers demonstrated how deep learning models can even learn to "translate" a video to text for automatic content descriptions which can further accelerate the progress in creating metadata for programming [Venugopalan].

The main advantage of deep learning approaches is that they allow us to learn feature representations automatically based on the underlying structure inherent in the data instead of having to create hand-crafted feature extractors based on heuristics which are often suboptimal as shown for both visual and audio features in Figure 4.

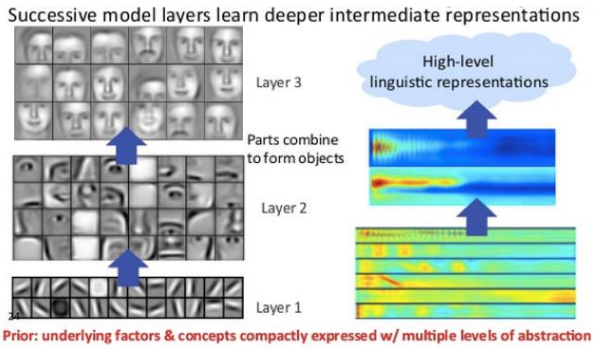

Figure 4 - Multi-layer Representation of Information in Deep Neural Networks [Lee]

Finger printing of content is another application of visual content analysis, for example to identify recurring video segments such as ads or news segments. For efficiency, video finger printing can use a hash representation for each frame, find likely matching candidates using an inverted search index, and then refine the matching using temporal constraints [Kanth].

Another important component of content analysis is the temporal segmentation of videos resulting in chaptering information, as well as the alignment of external meta-data such as sport highlight feeds or movie scripts with the video. We can detect all scene boundaries in a video and group them into larger segments to create chapters that are semantically distinctive.

An essential component of segmentation algorithms is shot-boundary detection which detects the presence of camera view discontinuities, e.g. the boundaries between sequences of contiguously recorded video frames..

We can efficiently find hard cuts between shots by thresholding the difference between the frame signatures of consecutive frames, such as quantized color histograms. We then choose a statistically motivated similarity metric function such as chi-square distance to compute the distances, and choose our thresholds appropriately depending on the content.

More complex transitions such as fades and wipes can also be detected by analyzing the spatial and temporal coherence of the mean and standard deviation of the image brightness values for each frame [Lienhart].

Once we have found shot boundaries, we can use various techniques to group shots into scenes, and scenes into program segments. Helpful cues here are the rate of shot changes, the presence of channel logos, the presence of black frame sequences, etc. These segments can be further classified into programs, commercials, credit roll, etc. using sequential classification approaches such as CRFs [Lafferty] or Recurrent Neural Networks (RNN) [Ng].

Audio, Speech and Closed Caption Analysis

Audio features are also very useful for program segmentation. Typical audio features include the volume, zero-crossing rate, short-time energy ratio, and spectrum flux. These features tend to be complementary to the visual features.

Mining and understanding what is being said in a program is especially helpful to describe what is happening on screen, specifically for talk shows where it is hard to make sense of what is happening on screen based on visual information alone [Schonfeld].

We can detect the topics that are being discussed at the moment on TV by reading all the text visible on the screen and combine it with the information mention in closed caption and speech transcripts. For news shows this can provide us with natural and accurate segment descriptors.

To analyze the text and closed caption we can first apply standard natural language processing techniques [Manning] to detect the parts of speech such as subject, verbs, and objects, before modeling the statistical properties such as co-occurrences between the words and sentences. This can then be used to find representative words and segment the video into coherent clips (Figure 5).

Figure 5 - Chaptering of a News Show

Finally, we could achieve a better and more robust result by a fusing visual, audio, speech and closed caption analysis results due to the complementary nature of different features [Tzoukermann].

EXAMPLE APPLICATIONS

Video distributors are selling a premium video product to its customers. To make this product appealing it is important that they offer an enhanced and more personalized entertainment experience for customers than competing products.

Having access to detailed metadata at the moment and segment level allows them to identify people, locations, activities, music, concepts in scenes, leading to stronger engagement of customers within their ecosystem instead of losing them to second screen devices.

Enhanced and real-time metadata will lead to more interactive and engaging experiences, for example a Trending Topic Guide could show which channel is currently showing a discussion of a topic that a customer is interested in. The viewer could then directly access the relevant programs without having to scan through the channel guide

The guide could also indicate breaking news by noticing that similar topics are being discussed at the same time on multiple channels and further incorporate information from Twitter and other social media sources to enhance the news experience.

The automatic creation of relevant short content from long programming would also be an appealing application, for example one could automatically create collections of the "best" SNL or Jimmy Fallon's late night show sketches from last night, the highlights for my favorite sports team or the news segments that match my interest. Chapter annotations allow for easy in-asset navigation of sports highlights and news segments that is more similar to how customers navigate clip based video platforms as seen in Figure 5.

To find these chapter markers, we first use visual analytics to temporally segment the videos into shots and scenes, then we extract topics and summary sentences from the closed caption co-occurrence statistics to group the segments into clusters, and finally use an exemplar based clustering approach to relate clips across broadcasts.

Sports highlights can be automatically aligned with a video by using OCR to automatically detect read the current game clock in the video. Once the game clock is analyzed, we can align the detected clock times with an external game highlight feed to create a temporal annotation of each highlight moment.

More detailed descriptions of the media assets lead to improved search and personalized recommendations that goes beyond simple keywords but instead allow one

to search via semantic concepts which is important as voice interaction with the TV becomes the norm rather than the exception, e.g. Xfinity X1 Voice Remote, Apple TV, Amazon Fire TV.

Finally, automatic content analysis could also create visual descriptions as an assistive technology and thereby noticeably increase the coverage of these solutions.

SUMMARY

In this article we described how the availability of detailed content metadata enables a range of new and exciting user experiences. Due to the time and labor intensive nature of creating such metadata manually it is essential that human annotation is augmented with automatic content analysis techniques such as described in this paper. Only then can the metadata creation process scale to handle large number of assets and near real-time constraints.

Finally, we gave an overview about common approaches to analyze the visual, audio and language content of an asset and how these approaches are combined in the industry today to enable the entertainment experiences of the future.

REFERENCES

[Bengio] Y. Bengio. Learning deep architectures for AI, Foundations and Trends in Machine Learning, 2(1):1-127, 2009.

[Bishop] Christopher Bishop. Pattern Recognition and Machine Learning, Springer, 2007

[He] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, ICCV 2015

[Huang] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49, 2007.

[Kanth] K. V. Ravi Kanth, Divyakant Agrawal, Amr El Abbadi and Ambuj Singh. Dimensionality reduction for similarity searching in dynamic databases. SIGMOD 1998

[Krizhevsky] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012

[Lafferty] John Lafferty, Andrew McCallum and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, ICML 2001

[Lee] Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Y. Ng. Convolutional Deep Belief Networks for scalable Unsupervised Learning of Hierarchical Representations, ICML 2009

[Lienhart] Rainer Lienhart. Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. International Journal of Image and Graphics (IJIG), Vol. 1, No. 3, pp. 469-486, 2001.

[Manning] Christopher Manning and Hinrich Schuetze, Foundations of Statistical Natural Language Processing, May 1999, MIT Press

[Ng] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici, Beyond Short Snippets: Deep Networks for Video Classification, CVPR 2015

[Ries] Christian X. Ries, Rainer Lienhart. A Survey on Visual Adult Image Recognition. Multimedia Tools and Applications, Springer, 2012

[Schonfeld] Dan Schonfeld, Caifeng Shan, Dacheng Tao, Liang Wang. Video Search and Mining, Springer, 2010

[Tzoukermann] Evelyne Tzoukermann, Geetu Ambwani, Amit Bagga, Leslie Chipman, Tony Davis, Ryan Farrell, David Houghton, Oliver Jojic, Jan Neumann, Robert Rubinoff, Bageshree Shevade, and Hongzhong Zhou. Semantic Multimedia Extraction using Audio and Video. Multimedia Information Extraction. Wiley-IEEE Computer Society Press, 2012.

[Venugopalan] Subhashini Venugopalan, M, Rohrbach, Jeff Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko, Sequence to Sequence – Video to Text, Proceedings of the 2015 International Conference on Computer Vision (ICCV), 2015.

[Viola] Paul Viola and Michael Jones. Robust Real-time Object Detection, IJCV 2001