Temporal Video Analyzer (TVAN): Efficient Temporal Video Analysis for Robust Video Description and Search

Daniel Rotman, Dror Porat, Yevgeny Burshtein, Udi Barzelay
IBM Research AI
Haifa, Israel

Abstract

With the increasing popularity of video content, automatic video understanding is becoming more and more important for streamlining video content consumption and reuse. In this work, we present TVAN—temporal video analyzer—a system for temporal video analysis aimed at enabling efficient and robust video description and search. Its main components include: temporal video segmentation, compact scene representation for efficient visual recognition, and concise scene description generation. We provide a technical overview of the system, as well as demonstrate its usefulness for the task of video search and navigation.

Introduction

Automated video understanding is one of the biggest challenges in AI today. With the rise in volume of video content, it is important for companies with large-scale video corpuses to automatically extract insights from their video content, thus enabling effective content-based indexing for consumption and reuse. To make the information in videos accessible, it is necessary to create AI technologies which can temporally segment, analyze, index and retrieve desired elements in videos.

To address this need, in recent years video comprehension and navigation tasks have risen in interest in both academic and industrial settings. Tasks such as image-based visual recognition (Russakovsky et al. 2015) and video retrieval systems (Song et al. 2018) aim to aid in the task of video information extraction and browsing. A number of public APIs (Li 2017) can perform a static (image-based) visual analysis of a video, and provide a layer of meta-data which can then be used for video indexing.

In this work we present TVAN, a novel temporal video analysis framework which specifically focuses on the temporal analysis of video. TVAN given an input video of heterogeneous nature, can generate meta-data over multiple levels of temporal granularity—shots and scenes. This metadata can be used by a video navigation interface for browsing the video for audio and visual concepts. Such a system can be leveraged and applied to large video corpuses to extract the information of interest, allowing effective indexing, navigation and retrieval.

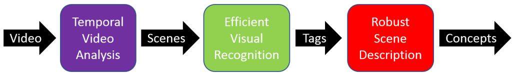


Figure 1: Temporal and visual analysis for videos.

TVAN is composed of three components (see Figure 1):

1. Temporal Video Analysis: We analyze videos and segment them temporally into scenes—a semantic level of division above shots. For videos of heterogeneous nature, this is a critical step to allow understanding of the video with regard to both local and global temporal changes.
2. Efficient Visual Recognition: Prior to performing frame-based visual recognition using state-of-the-art neural networks, we create a compact representation of the scene to eliminate temporal redundancy and minimize computational cost and time.
3. Robust Scene Description: Utilizing the visual tags for the frames constituting the compact representation of the scene, we create a concise and representative description of each scene, resulting in an automatically constructed table-of-contents for the video.

Temporal Video Analysis

The very first phase when analyzing a video is to identify the temporal structure and nature of the video and segment it temporally into homogeneous chapters. This stage is crucial when dealing with heterogeneous videos, as content analysis of a video with multiple semantic chapters will likely result in inaccurate results. We perform this task using our unsupervised learning mutlimodal video scene detection technology (Rotman, Porat, and Ashour 2017).

Video scene detection is the task of temporally dividing a heterogeneous video into its semantic sections, called scenes. Scenes typically relay a specific concept or theme which acts as a component of the story delivered by the video. Formally, scenes are defined as a sequence of semantically related and temporally adjacent shots depicting a high-level concept or story (Rui, Huang, and Mehrotra 1999).

We use our optimal sequential grouping formulation (Rotman et al. 2018) to partition the video using a multimodal fusion of features extracted using state-of-the-art neural networks for visual and audio analysis. This technology, which achieves state-of-the-art performance on various datasets, is the first to be incorporated into such a video search and retrieval system.

Efficient Visual Recognition

Once videos are divided into scenes we proceed to analyze the visual elements throughout the scene. Image-based visual recognition can be used to identify visual elements in a video frame as *tags* coupled with their confidence score.

Collecting these visual tags for videos can be a daunting task, due to the computational cost of the recognition process (and limited resources such as GPUs) and the sheer amount of visual information in videos. Straightforward methods which are implemented in systems today do not take into account the redundancy or temporal change of the visual elements in the video (Hosseini et al. 2017).

We therefore propose an unsupervised learning method for efficient visual recognition for video using max distance cluster tiling (MDCT). This novel formulation represents the frames of the video in a low-level color feature space and assigns them to clusters using a maximal distance constraint. By limiting the maximum distance that frames are clustered together, we can assure that the visual elements in the video are represented while eliminating redundancies.

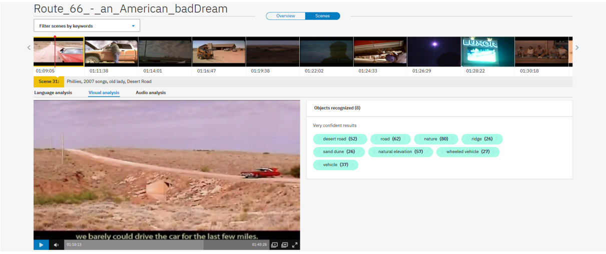
Unlike classical clustering, MDCT is motivated by the fact that two similar frames which would return the same visual tags by frame-based visual recognition will undoubtedly be close together in a low-level color space. Classical clustering can cluster together large scattered groups of points which might have common features but quite probably do not include necessarily the same visual tags.

We develop an algorithm for MDCT called *greedy cluster aggregation* which we derive from the set cover problem (Slavık 1997), where the frames represent the universe and the sets are frames with a smaller distance than the maximum. We can leverage the greedy algorithm for polynomial time approximation of the NP-hard set cover problem which features relatively sound bounds on optimality. In our instance, the frames which cover the most uncovered frames are chosen as cluster centers, and they are added in descending order taking into account a cost function composed of the maximum distance of a frame in each cluster.

Robust Scene Description

Given the visual tags of frames in the scene, we can now proceed to classify the important and representative concepts of the scene. For the frame tags, we use tag confidence, frequency and co-occurrence. We build a novel tag-importance algorithm to identify the most clear, significant and representative tags in the scene.

The confidence of a tag corresponds to the ambiguity in its presence, while the frequency corresponds to its repetitiveness in the scene. The co-occurrence in this instance measures which tags appear frequently together—these tags are likely themes which identify a particular element in the video which is characteristic of the scene. This results in selected tags which are *concepts* that identify the main themes and scene descriptions throughout the video.



Route_66_-_an_American_badDream

Scene 31: Phillies, 2007 songs, old lady, Desert Road

Objects recognized (8)

Very confident results

desert road (52) road (62) nature (80) ridge (26) sand dune (26) natural elevation (57) wheeled vehicle (27) vehicle (37)

we barely could drive the car for the last few miles.

Figure 2: Search engine for navigation in videos.

Search Engine

With these AI technologies composing TVAN for temporal analysis of the video in place, we build a search engine to allow a user to navigate and discover sections of interest and components in a large corpus of videos. By layering additional tiers of semantic information such as speech recognition, natural language understanding and emotion analysis we can give a complete picture of the video and its content. Moreover, the various analytics can be aggregated per scene giving not only the specific identified elements in the video, but also the temporal flow and change of concepts, emotions and keywords throughout the video (see Figure 2).

References

Hosseini, H.; Xiao, B.; Clark, A.; and Poovendran, R. 2017. Attacking automatic video analysis algorithms: A case study of google cloud video intelligence api. In *Proceedings of the 2017 on Multimedia Privacy and Security*, 21–32. ACM.

Li, F.-F. 2017. Announcing google cloud video intelligence api, and more cloud machine learning updates. *cloud. google. com/blog* 8.

Rotman, D.; Porat, D.; Ashour, G.; and Barzelay, U. 2018. Optimally grouped deep features using normalized cost for video scene detection. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 187–195. ACM.

Rotman, D.; Porat, D.; and Ashour, G. 2017. Robust video scene detection using multimodal fusion of optimally grouped features. In *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*, 1–6. IEEE.

Rui, Y.; Huang, T. S.; and Mehrotra, S. 1999. Constructing table-of-content for videos. *Multimedia systems* 7(5):359–368.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Slavık, P. 1997. A tight analysis of the greedy algorithm for set cover. *Journal of Algorithms* 25(2):237–254.

Song, J.; Gao, L.; Liu, L.; Zhu, X.; and Sebe, N. 2018. Quantization-based hashing: a general framework for scalable image and video retrieval. *Pattern Recognition* 75:175–187.