Keyframe Extraction for Video Tagging & Summarization

Damian Borth, Adrian Ulges, Christian Schulze, Thomas M. Breuel
damian.borth@dfki.de
German Research Center for Artificial Intelligence (DFKI)
and University of Kaiserslautern

Abstract: Currently, online video distributed via online video platforms like YouTube experiences more and more popularity. We propose an approach of keyframe extraction based on unsupervised learning for video retrieval and video summarization. Our approach uses shot boundary detection to segment the video into shots and the k-means algorithm to determine cluster representatives for each shot that are used as keyframes. Furthermore we performed an additional clustering on the extracted keyframes to provide a video summarization. To test our methods we used a database of videos downloaded from YouTube where our results show (1) an improvement of retrieval and (2) compact summarization examples.

GI-Topic: KI-BV (artificial intelligence - image understanding)

1 Introduction

Searching video media on the web often means to use the available search interfaces provided by online video portals. YouTube as the market leader provides keyword search based on manually generated meta information and tags. Unfortunately, meta data is limited in its ability of representing the content of a video, and tags are subjective labels that might be misleading in their semantics. This is why content-based video retrieval (CBVR) can improve video search. One key component of CBVR is the extraction of keyframes which then can be analyzed with known image processing algorithm. The resulting features can be indexed and used for further retrieval. As we will show the quality of keyframe extraction influences the overall performance of retrieval systems. Additionally, keyframes can be used to summarize search results and help the user to quickly evaluate the relevance of a video [BMN+03] and for video browsing like in [ROS04], where keyframes are organized along a temporal and visual plane to each other.

2 Our Approach

Video is by its nature a temporally structured media and the segmentation into its basic temporal units (shots) is usually the first step in the processing pipe of content-based retrieval systems. Based on the detected shots there exist several keyframe extraction approaches which then are used as representatives of the shot. A straightforward approach is

to take the first, middle or last frames of a shot as keyframes [O'C91], which might result in information loss within long shots. To overcome this problem Hammoud and Mohr [HM00] proposed unsupervised learning techniques that adapt to the content to the shot. Our approach is a combination of shot boundary detection and an intra-shot clustering of frames to find an adequate number of representative keyframes for the given shot with respect to its visual complexity.

Shot Boundary Detection Among the many different approaches that exist in temporal segmentation [KC01], we use the MPEG-7 Color Layout Descriptor (CLD) as a feature [MOVY01] for each frame and compute differences between consecutive frames. The detection of shot boundaries follows an adaptive threshold technique as stated in [Lie01]. This method works very well for hard cuts but leaks in performance for long dissolves and fades.

Intra-Shot Clustering Modelling every shot as a mixture of Gaussian densities we use the k-means algorithm [McQ67] for keyframe extraction on the CLD feature vectors extracted from every frame of a shot. After clustering we define the nearest frame to the mean of every cluster as a representative for the shot. To estimate the number of clusters and therefore the number of keyframes we use the Bayesian Information Criterion (BIC) [Sch78] for cluster validity measurement. According to the visual complexity of the shot this might lead to multiple keyframes. Figure 1 shows an example with a feature space representation of a complex shot in the center (for visualization purpose reduced to 2d). Visually similar frames are clustered together and their representatives are shown on the left and right side of the image.

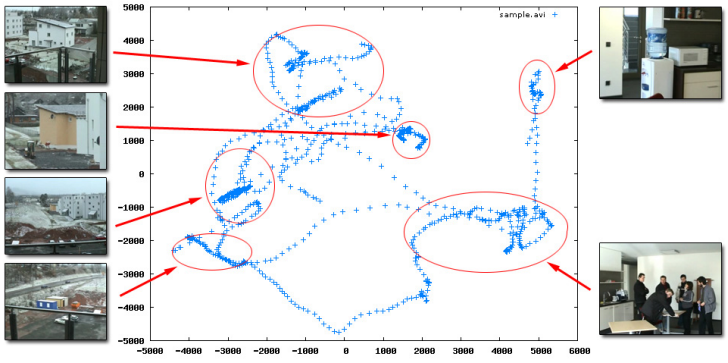

Figure 1: Feature space representation of a complex shot (center). For visualization purpose the feature space was reduced to 2d by PCA. A clustering of the frames leads to keyframes displayed on the left and right side.

Keyframe "Meta-Clustering" A shortcoming of this two step process is the duplication of the same content by multiple keyframes due to the missing inter-shot reasoning. This occurs often when actors have a dialog, or in music video clips where the artist is shown in different shots throughout the video. This tradeoff is addressed in the context of video summarization where we propose a "meta-clustering" on the keyframes extracted in the previous steps. Similar to [BMN+03] we use the idea of grouping keyframe together to create a summary of the video. But in contrast to creating a hierarchical structure defined by groups and scenes, we directly use the already extracted keyframes and cluster them using the k-means algorithm.

3 Experiments & Results

Tagging The keyframe extraction method was used in an experiment for automatic tagging of video material on a database of 2200 videos (total duration 194 hrs.) downloaded from YouTube. In this experiment a video tagger [1] learned concepts like *cat*, *sailing*, *desert* out of a training set of already available tags from YouTube and reached a Mean Average Precision (MAP) of 34.2% on unknown test videos. As features for the experiment we used different combination of color histograms, Tamura features and Visual Words extracted from the given keywords and motion descriptors extracted on shot level. To investigate the impact of the keyframe extraction for a successful tagging we performed the tagging experiment with two additional keyframe extraction methods: `[first]`, where we take the first frame of a shot as representative and `[regular]`, where we regularly sample keyframes within an interval of 7 sec. Based on the chosen combination of used features the results show that our approach increases performance by 2-9% compared to `[first]` and was as performant as `[regular]` but extracting less keyframes.

Summarization We performed "meta-clustering" on several music video clips, material that is characterized by a high amount of cuts and redundant shots. We obtained compact summarizations by grouping duplicate keyframe together. Pure clustering on the entire video without shot distinction was leading to roughly the same amounts of keyframes but seems to miss some content. Figure 2 shows an example of a music video clip where duplicates were successfully clustered together. The first row of the Figure 2 displays the result of our "meta-clustering" based on the previous extracted keyframes, which are shown in the third row. The second row displays the keyframe extracted by a pure clustering based on the entire video. We obtained comparable results for other videos.

[1] described in more detail in [USKB07]

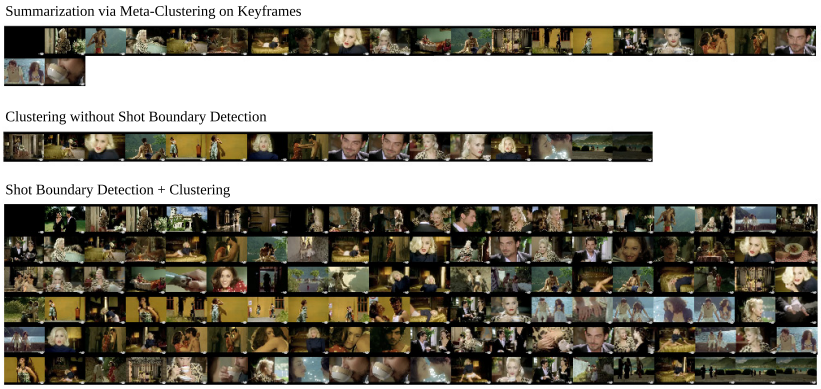

Figure 2: Result of summarization for a music video clip. First row displays "meta-clustering". Second row displays the pure clustering. Third row displays keyframes extracted by our basic method

References

[BMN+03] W. Bailer, H. Mayer, H. Neuschmied, W. Haas, M. Lux, and W. Klieber. Content-based video retrieval and summarization using MPEG-7. *Internet Imaging V. Edited by Santini, Simone; Schettini, Raimondo. Proceedings of the SPIE*, 5304:1–12, 2003.

[HM00] R. Hammoud and R. Mohr. A Probabilistic Framework of Selecting Effective Key-Frames for Video Browsing and Indexing. *Intern. Worksh. on Real-Time Img. Seq. Anal.*, pages 79–88, 2000.

[KC01] I. Koprinska and S. Carrato. Temporal Video Segmentation: A Survey. *Signal Processing Image Communication*, 16:477–500, 2001.

[Lie01] R. Lienhart. Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. *International Journal of Image and Graphics*, 1(3):469–486, 2001.

[McQ67] J.B. McQueen. Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[MOVY01] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, and A. Yamada. Color and Texture Descriptors. *IEEE Trans. on Circuits Syst. for Video Techn.*, 11(6), 2001.

[O'C91] Brain O'Connor. Selecting Key Frames of Moving Image Documents: A Digital Environment for Analysis and Navigation. *Microcomputers for Information Management*, 8(2):119–33, 1991.

[ROS04] M. Rautiainen, T. Ojala, and T. Seppanen. Cluster-temporal browsing of large news video databases. *IEEE Int. Conference on Multimedia and Expo*, 2:751–754, 2004.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[USKB07] A. Ulges, C. Schulze, D. Keysers, and T.M. Breuel. Content-Based Video Tagging for Online Video Portals. In *MUSCLE/Image-CLEF Workshop*, 2007.