Unsupervised recognition and summarization of sport games in sports video

Overview

The problem is formulated as a segmentation task on long television broadcast of mixed Olympic games coverage with a duration of over 4 hours. The Olympic coverage consists of mixed sports, such that no sports specific detectors could be applied; and with visually similar and sequential races of the same sport without interruption where the goal is to isolate each race individually in the final segmentation. This requires a different approach in terms of feature selection, and the choice is made to develop an unsupervised method that is sufficiently robust to allow application in other domains. A summarization component applies labels to the ensuing scene segments by identifying persons present, salient key-words, and the type of sport by means of a classifier trained on Wikipedia articles.

In depth description

We presented a novel, unsupervised approach for segmenting a long video stream. We extracted two multi-modal features to aid us in this task; one based on similar shots. These multi-modal chains served as an input for two methods, Greedy-Clustering, and Density-Clustering. Both were compared against a well-known segmentation approach using Scene Transition Graphs, which they exceeded in terms of F-measure and the coverage metric. While both our methods have comparable F-measures, Density-Clustering is the better option for determining the semantic structure in the video stream, both in terms of identifying boundaries and in the overall number of unique scenes found. Density-Clustering was able to correctly identify 80% of all scenes longer than one minute in the video. Inclusion of an audio feature, in the form of a silence/music/speech detector further improved recall and scene coverage, but at the expense of some oversegmentation. We suspect that our system will do well on broadcasts of an expository nature, where there are clear semantic boundaries between each scene. In our current dataset, we illustrate this by distinguishing between events of different sports, for instance cycling vs. gymnastics, and by distinguishing between events of the same sport but with different semantic content, for instance in swimming, the Men's 100m freestyle 2nd semi-final followed by the Women's 200m freestyle final. Further improvements in scene recognition might be realized by performing speaker diarization, the identification of every speaker, and this would fit well into the chain paradigm. Likewise, the audio environment of the video could be vector quantized, in essence turning background audio into a set of classifiable entities, which would permit the integration of the audio modality in the chain paradigm. Our commentator could be improved by using latent topic models such as Latent Dirichlet Allocation to infer better word to sport associations. We leave these thoughts for examination in future work. Summarization and segmentation are two sides of the same coin. More refined segmentation of a source allows for better analysis and summarization later on. Our commentator is able to describe a semantic event that is occurring and its participants given a good initial segmentation. We emphasize that the core work of segmenting the video stream into semantic events is primarily unsupervised. The parameters for the similarity criterion for each multi-modal chain and the local minima for the Density-Clustering can readily be determined; only the audio component requires additional supervision. With minimal extra knowledge automatically learned from Wikipedia articles, a semantic meaning can be assigned to the resultant scene segments in the form of important keywords and any participating persons. The accuracy of the commentator ranges from 45% to 90% depending on the recognized sport, where video frames with few texts or with text with ambiguous keywords such as "race", "trial" or "lane" were the most difficult to correctly comment. The resultant labels provide a semantic connotation, and can be used for further indexing purposes or other cross-media fusion tasks.

Potential fields of application

This technology can be included as a part of a media search engine as it provides unsupervised indexing and summarization of sport games.

Possibilities for exploitation

We plan to seek collaborative projects with the Belgian industry to further finalize and exploit this technology.

Further Information

Further technical information is available in Poulisse, G.-J., Patsis, Y. & Moens, M.-F. (2012). Unsupervised Scene Detection and Commentator Building Using Multi-modal Chains. *Multimedia Tools and Applications*. DOI 10.1007/s11042-012-1086-0.

Contact person

Prof. Marie-Francine Moens
Department of Computer Science
Celestijnenlaan 200A
B-3001 Heverlee, BELGIUM
sien.moens@cs.kuleuven.be