Aggregated Feature Video Retrieval for MPEG-7 via Clustering

Jiamin Ye
Centre for Digital Video Processing
Dublin City University, Dublin
Ireland.
jiaminye@computing.dcu.ie
http://www.computing.dcu.ie/~jiaminye

MPEG-7 is a generic standard used to encode information about multimedia content and often, different MPEG-7 Descriptor Schemas are instantiated for different representations of a shot such as text annotations and visual features. Our work focuses on two main areas, the first is devising a method for combining text annotations and visual features into one single MPEG-7 description and the second is defining how best to carry out text and non-text queries for retrieval via a combined description.

The dialogue (i.e. ASR transcripts or closed-caption) in videos tells us what has been said but very often not what we can see on the screen. A video retrieval system is encouraged to integrate primitive visual features to find "difficult" relevant shots that can not be found in the dialogue text.

We align the video retrieval process to a text retrieval process based on the TF*IDF vector space model via clustering of low-level visual features (i.e. RGB colour space, colour histogram and edge component histogram). Our assumption is that shots within the same cluster are not only similar visually but also semantically, to certain extent. Our method maps the visual features of each shot onto a term weight vector via clustering. This vector is then combined with the original text features of the shot (i.e. ASR transcripts) to produce the final searchable index.

A typical video retrieval approach is to search different MPEG-7 descriptions separetely and to combine ranked results from the different searches using a sum weighted method. Our approach attempts to integrate the different descriptions into the index and query preparation stages - no combination of ranked results is required.

Our TRECVID2002 and TRECVID2003 experiments show that adding extra meaning to a shot based on the shots from the same cluster is useful when each video in the collection contains a high proportion of similar shots, for example in the documentaries of the TRECVID2002 collection. Adding meaning to a shot based on the shots that are around it might not be an effective method for video retrieval when each video in the collection has low proportion of similar shots such as TV news programmes due to the fact that neighbouring shots typically fall within the same story.