Applied Soft Computing 46 (2016) 1008–1029

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Review Article

Hybrid soft computing approaches to content based video retrieval: A brief review

Hrishikesh Bhaumik[a,*], Siddhartha Bhattacharyya[a], Mausumi Das Nath[a], Susanta Chakraborty[b]

[a] Department of Information Technology, RCC Institute of Information Technology, Canal South Road, Beliaghata, Kolkata 700 015, India
[b] Department of Computer Science & Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711 103, India

ARTICLE INFO

Article history:
Received 7 October 2015
Received in revised form 4 March 2016
Accepted 23 March 2016
Available online 27 April 2016

Keywords:
Content-based video retrieval
Video segmentation
Soft computing
Hybrid soft computing

ABSTRACT

There has been an unrestrained growth of videos on the Internet due to proliferation of multimedia devices. These videos are mostly stored in unstructured repositories which pose enormous challenges for the task of both image and video retrieval. Users aim to retrieve videos of interest having content which is relevant to their need. Traditionally, low-level visual features have been used for content based video retrieval (CBVR). Consequently, a gap existed between these low-level features and the high level semantic content. The semantic differential was partially bridged by proliferation of research on interest point detectors and descriptors, which represented mid-level features of the content. The computational time and human interaction involved in the classical approaches for CBVR are quite cumbersome. In order to increase the accuracy, efficiency and effectiveness of the retrieval process, researchers resorted to soft computing paradigms. The entire retrieval task was automated to a great extent using individual soft computing components. Due to voluminous growth in the size of multimedia databases, augmented by an exponential rise in the number of users, integration of two or more soft computing techniques was desirable for enhanced efficiency and accuracy of the retrieval process. The hybrid approaches serve to enhance the overall performance and robustness of the system with reduced human interference. This article is targeted to focus on the relevant hybrid soft computing techniques which are in practice for content-based image and video retrieval.

Contents

* Corresponding author. Tel.: +91 9830198815.
E-mail addresses: hbhaumik@gmail.com (H. Bhaumik), dr.siddhartha.bhattacharyya@gmail.com (S. Bhattacharyya), mausumi.dasnath@gmail.com (M.D. Nath), susanta.chak@gmail.com (S. Chakraborty).

http://dx.doi.org/10.1016/j.asoc.2016.03.022
1568-4946/

1. Introduction

With the immense popularity of video recorders and reduction in cost of digital storage devices, voluminous amount of videos are uploaded at an incredible rate for online browsing and retrieval purposes. As video encompasses the other three media types, i.e. text, image and audio, combining them into a single data stream for transmission and retrieval has drawn the attention of researchers and application developers to a great extent. Image/video retrieval has been an active research domain since the last four decades.

In comparison to images, videos possess characteristic features. Although the organization of a video is not well defined, the content is more affluent as compared to individual images. Archiving and indexing of images and videos based on semantic content is a challenging task. Subsequent retrieval and browsing of video data manually is a laborious and time intensive task from the users' perspective. During the retrieval process, the primary objective is to present the videos of interest to the user. As such, this broad domain of research is referred to as content-based video retrieval (CBVR). Since a video is a conglomeration of time sequenced images, research in the field of CBVR has been supplemented to a great extent by advances in the field of content-based image retrieval (CBIR). In some approaches, CBVR applications have been augmented with audio cues. As such, approaches in audio indexing, retrieval and classification [1] are important for advances in CBVR.

The term "content-based" refers to features like color, texture, intensity, trajectory of objects or statistical characteristics at the low-level. Mid-level features refer to feature points taken on an image or a series of images as in a video. These feature points are detected by algorithms for feature point detection and description such as SIFT [2], SURF [3], BRISK [4], DAISY [5], GIST [6], ORB [7], etc. The detected feature points in images are matched in order to compute the amount of similarity. Mid-level features are incapable of evaluating the semantic content in a video. This drawback is alleviated using high-level features such as edges, shapes, motion vector, optical flow, event modeling (in videos), timbre, rhythm (in audio), etc. involving different levels of semantics. High-level features are capable of handling semantic queries like "retrieve videos where blue sky and snowy mountains are present". These queries require matching the semantic content in the video database. The major hurdle behind processing high-level queries is the semantic gap that exists between the high-level features and low-level ones.

The process of "retrieval" involves matching features extracted from the frames of the video with the query given by the user. The features are used to create the feature vectors for each video. All videos in the database are represented as a point in an n-dimensional space, where n represents the number of features under consideration. It is pertinent to mention here that taking too many features enhances the computational cost since the dimensionality of the feature vector increases. Over time, researchers have adopted various means for reducing dimensionality. Principal component analysis (PCA) [8] is a very widely used technique for reducing the number of features by mapping the set of original features (P) to another set (Q). The feature set Q contains derived features from the set P. Importantly, the cardinality of the set Q is lesser than P. Researchers have also relied on rough sets for identifying features with high discrimination power and thereby reducing the dimensionality by eliminating less important features. The features taken into consideration form the feature vector for each object (video) in the database. Distance between the feature vectors acts as a measure for similarity between the videos. The commonly used distance measures are Euclidean, Mahalanobis, Hausdorff, Chi-square, etc. The end product of the retrieval process is a set of videos ranked according to relevance. In some systems user relevance feedback is used to tune the system in order to produce more meaningful results. This also helps to model human perception in a better way.

For better understanding, an analogy is drawn between a document and a video sequence (refer Fig. 1). A document consists of paragraphs much like a video being composed of scenes. A paragraph in turn represents a group of inter-related sentences similar to inter-related shots, forming a scene in a video. Further down the hierarchy, the sentences are composed of words, like the shots consist of frames. A frame in a video denotes a single image, whereas a shot is a consecutive sequence of frames recorded by a single camera. A scene represents a collection of semantically related and temporally adjacent shots, portraying and imparting a high-level story. A group of scenes comprises a sequence/story. Frames and

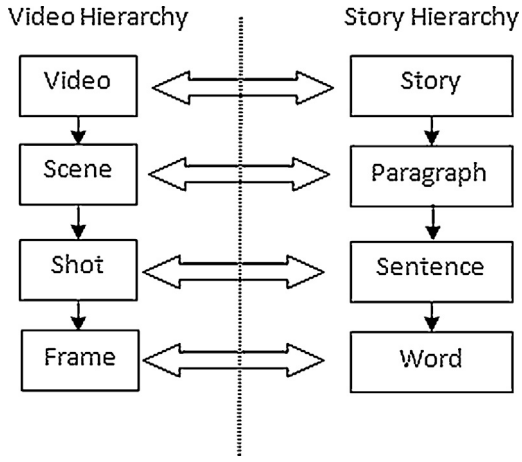

Fig. 1. Analogy between video and story hierarchy.