Assessing Digital Video Data Similarity

Waleed E. Farag
Indiana University of Pennsylvania, USA

INTRODUCTION

Multimedia applications are rapidly spread at an ever-increasing rate, introducing a number of challenging problems at the hands of the research community. The most significant and influential problem among them is the effective access to stored data. In spite of the popularity of keyword-based search technique in alphanumeric databases, it is inadequate for use with multimedia data due to their unstructured nature. On the other hand, a number of video content and context-based access techniques have been developed (Deb, 2005). The basic idea of content-based retrieval is to access multimedia data by their contents, for example, using one of the visual content features. While context-based techniques try to improve the retrieval performance by using associated contextual information, other than those derived from the media content (Hori & Aizawa, 2003).

Most of the proposed video indexing and retrieval prototypes have two major phases, the database population and the retrieval phase. In the former one, the video stream is partitioned into its constituent shots in a process known as shot boundary detection (Farag & Abdel-Wahab, 2001, 2002b). This step is followed by a process of selecting representative frames to summarize video shots (Farag & Abdel-Wahab, 2002a). Then, a number of low-level features (color, texture, object motion, etc.) are extracted in order to use them as indices to shots. The database population phase is performed as an off-line activity and it outputs a set of metadata with each element representing one of the clips in the video archive. In the retrieval phase, a query is presented to the system that in turns performs similarity matching operations and returns similar data back to the user.

The basic objective of an automated video retrieval system (described above) is to provide the user with easy-to-use and effective mechanisms to access the required information. For that reason, the success of a content-based video access system is mainly measured by the effectiveness of its retrieval phase. The general query model adopted by almost all multimedia retrieval systems is the QBE (query by example; Marchionini, 2006). In this model, the user submits a query in the form of an image or a video clip (in case of a video retrieval system) and asks the system to retrieve similar data. QBE is considered to be a promising technique since it provides the user with an intuitive way of query presentation. In addition, the form of expressing a query condition is close to that of the data to be evaluated.

Upon the reception of the submitted query, the retrieval stage analyzes it to extract a set of features then performs the task of similarity matching. In the latter task, the query-extracted features are compared with the features stored into the metadata; then matches are sorted and displayed back to the user based on how close a hit is to the input query. A central issue here is the assessment of video data similarity. Appropriately answering the following questions has a crucial impact on the effectiveness and applicability of the retrieval system. How are the similarity matching operations performed and based on what criteria? Do the employed similarity matching models reflect the human perception of multimedia similarity? The main focus of this article is to shed the light on possible answers to the above questions.

BACKGROUND

An important lesson that has been learned through the last two decades from the increasing popularity of the Internet can be stated as follows "[T]he usefulness of vast repositories of digital information is limited by the effectiveness of the access methods" (Brunelli, Mich, & Modena, 1999). The same lesson applies to video archives; thus, many researchers start to be aware of the significance of providing effective tools for accessing video databases. Moreover, some of them are proposing various techniques to improve the efficiency, effectiveness, and robustness of the retrieval system. In the following, a quick review to these techniques is introduced with emphasis on various approaches for evaluating video data similarity.

One important aspect of multimedia retrieval systems is the browsing capability and in this context some researchers proposed the integration between the human and the computer to improve the performance of the retrieval stage. Truong and Venkatesh (2007) presented a comprehensive review and classification of video abstraction techniques introduced by various researchers in the field. That work reviewed different methodologies that use still images (key frames) and moving pictures (video skims) to abstract video data and provide fast overviews of the video content. A prototype retrieval system that supports 3D images, videos, and music retrieval is presented in Kosugi et al. (2001). In that system each type of queries has its own processing module; for instance, image retrieval is processed using a component called ImageCompass.

Due to the importance of accurately measuring multimedia data similarity, a number of researchers have proposed various approaches to perform this task. In the context of image retrieval systems, some researchers considered local geometric constraint into account and calculated the similarity between two images using the number of corresponding points (Lew, 2001). In Oria, Ozsu, Lin, and Iglinski (2001) image are represented using a combination of color distribution (histogram) and salient objects (region of interest). Similarity between images are evaluated using a weighted Euclidean distance function, while complex query formulation was allowed using a modified version of SQL denoted as MOQL (Multimedia Object Query Language). Other researchers formulated the similarity between images as a graph-matching problem and used a graph-matching algorithm to calculate such similarity (Lew, 2001). Berretti, Bimbo, and Pala (2000) proposed a system that uses perceptual distance to measure shape feature similarity of images while providing efficient index structure. Hörster, Lienhart, and Slaney (2007) employed several measures to assess the similarity between a query image and a stored image in the database. These methods include calculating the cosine similarity, using the L1 distance, the symmetrized Jensen-Shannon divergence, while the last method is adopted from language-based information retrieval.

With respect to video retrieval, one technique was proposed in Cheung and Zakhor (2003) where a video stream is viewed as a sequence of frames and in order to represent these frames, in the feature space, high dimensional feature vectors were used. The percentage of similar clusters of frames common between two video streams is used as a criterion for measuring video similarity. A set of key frames denoted video signature is selected to represent each video sequence; then distances are computed between two video signatures. Li, Zheng, and Prabhakaran (2007) highlighted that the Euclidean distance is not suitable for recognizing motions in multi-attributes data streams. Therefore, they proposed a technique for similarity measure that is based upon singular-value decomposition (SVD) and motion direction identification.

Another technique was proposed in Liu, Zhuang, and Pan (1999) to dynamically distinguish whether two shots are similar or not based on the current situation of shot similarity. One other video retrieval approach introduced by Wang, Hoffman, Cook, and Li (2006) used the L1 distance measure to calculate the distance between feature vectors. Each of the used feature vectors is a combined one in which visual and audio features are joined to form a single feature vector. In Lian, Tan, and Chan (2003), a clustering algorithm was proposed to improve the performance of the retrieval stage in particular while dealing with large video databases. The introduced algorithm achieved high recall and precision while providing fast retrieval. This work used the QBE paradigm and adopted a distance measure that first aligns video clips before measuring their similarity.

A powerful concept to improve searching multimedia databases is called relevance feedback (Zhou & Huang, 2003). In this technique, the user associates a score to each of the returned hits, and these scores are used to direct the following search phase and improve its results. In Guan and Qui (2007), the authors proposed an optimization technique in order to identify objects of interest to the user while dealing with several relevance feedback images. Current issues in real-time video object tracking systems have been identified in Oerlemans, Rijsdam, and Lew (2007). That article presented a technique that uses interactive relevance feedback so as to address these issues with real-time video object tracking applications.

EVALUATING VIDEO SIMILARITY USING A HUMAN-BASED MODEL

From the above survey of the current approaches, we can observe that an important issue has been overlooked by most of the above techniques. This was stated in Santini and Jain (1999) by the following quote: "If our