Pattern Recognition 42 (2009) 267 – 282

Contents lists available at ScienceDirect

Pattern Recognition

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

State-of-the-art on spatio-temporal information-based video retrieval

W. Ren[a], S. Singh[b,*], M. Singh[b], Y.S. Zhu[a]

[a]The Key Laboratory of Integrated Microsystems, Shenzhen Graduate School, Peking University, China
[b]Research School of Informatics, University of Loughborough, Loughborough LE11 3TU, UK

ARTICLE INFO

Article history:
Received 30 December 2007
Received in revised form 21 August 2008
Accepted 22 August 2008

Keywords:
Video retrieval
Semantic knowledge
Content-based analysis
Spatio-temporal information

ABSTRACT

Video retrieval is increasingly based on image content. A number of studies on video retrieval have used low-level pixel content related to statistical moments, shape, colour and texture. However, it is well recognised that such information is not enough for uniquely discriminating across different multimedia content. The use of semantic information, especially which derived from spatio-temporal analysis is of great value in multimedia annotation, archiving and retrieval. In this review paper, we detail how the use of spatiotemporal semantic knowledge is changing the way in which modern research the conducted. In this paper we review a number of studies and concepts related to such analysis, and draw important conclusions on where future research is headed.

1. Spatio-temporal information for video retrieval

Content-based video retrieval is a very important area of research and several practical systems have been developed over the last decade with the aim of improving retrieval performance and tested on large-scale databases such as TRECVID http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html. Video classification and retrieval problems can be hierarchically categorised with a taxonomy, an example of which is presented by Roach et al. [1]. A key characteristic of video data is its associated spatial and temporal information that delivers semantically coherent narrative. Temporally consecutive frames have explicit spatial constraints with object inheritance, spatial relationships and motion information from their previous frames. Temporal trajectories of spatial relations among objects are as important as temporal object trajectories to represent object activities and reveal semantic evolution of spatial properties over time. The holy grail of almost all content matching-based video retrieval systems is to improve precision and recall metrics both through the process of improved content representation and use of good quality similarity metrics [2], as well as using a range of relevance feedback architectures and algorithms to allow the system to learn with time what is and is not a good match [3–6].

Unfortunately, temporal and spatial characteristics have not been adequately addressed in most video retrieval systems despite their obvious importance. In such systems, retrieval techniques work on indexing video by treating video sequences as collections of still images, extracting relevant key-frames, and comparing their low-level features. Over the past years, the representation of spatio-temporal data has been extensively discussed. It has inspired the development of mathematical foundations to represent spatio-temporal logic (STL) and reasoning [7], spatio-temporal database models and query languages for the description and manipulation of spatio-temporal objects [8,9], the temporal extension of current spatial data models within GIS [10,11], and a new generation of spatio-temporal video retrieval systems [12]. Spatiotemporal information in video deals with the evolution of spatial objects that change over time. Spatio-temporal modelling in video retrieval is a crucial step for using semantic information on image object relationship to improve the quality of content-based video retrieval. Such information can be used to tag video content and used as the basis for similarity computation between query and database videos. The similarity metrics and matching approaches depend heavily on the representation of spatio-temporal information, e.g., motion feature, spatio-temporal relations, object trajectory, video transition, etc. However, how to effectively model and represent spatio-temporal information is not straightforward. A spatio-temporal model usually first partitions the video into physical meaningful units (shots). This is followed by modelling the spatial relationships among objects in each frame. A final step analyses the temporal evolution of spatial relationships among objects over temporal intervals in each shot as well as in the whole video sequence. More importantly, a spatio-temporal model should suggest a practical solution for effective indexing and comparison. In summary, a spatio-temporal model should provide for:

(a) Representation of the structural elements of video data such as frame, shot, and sequence at different levels of abstraction.

* Corresponding author.
E-mail address: s.singh@lboro.ac.uk (S. Singh).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.08.033

(b) Description of the spatial composition among video objects in each frame including directional and topological relations, and temporal composition among frames within shot and sequences.

Spatial and temporal compositions are two important aspects for the representation of a spatio-temporal model. There are two main approaches for modelling such information:

(a) An integrated approach where objects, their spatial relationships and events are considered as a 3D (three-dimensional) volume with time being the third axis. One can construct a volume of spatio-temporal data in which objects in consecutive images are stacked to form a third temporal dimension. In this approach, the video events can be represented by the analysis of this 3D space based on object trajectories, shape analysis and motion analysis. A sequence of frames $(f_1 f_2, \ldots f_N)$ is represented by a volume in (x,y,t) space, where (x,y) are the discrete spatial coordinates in each frame (f_i), and time (t) is a discrete temporal coordinate that specifies frame number. The key benefit of this representation is that objects' spatial and temporal continuity is explicitly and conjointly provided. Shape and position change of a video object over time (t) is considered in terms of translation, scaling, and rotation of the object. A semantic scene can be delivered as variances of visual appearance from sequence to sequence. This is sequence-to-sequence indexing model. Spatio-temporal information relating to object movement is identified by tracing the trajectories of objects in this 3D (x,y,t) space. The motion trajectories of objects are defined as a physical change in the geographic position of the objects in the video. The trajectories are derived from changing the location of particular points on the objects, or from tracking contours of the objects over time. The former is trajectory slice model, whereas the latter is called trajectory volume model. In this model, time (t) is a critical component. The representation of this model is highly time dependent. Therefore, using different time scales will impact on the representation of this model, and further impact on the final results of indexing and matching. For instance, when we try to match two actions under different time scales by shape comparison, the solution is not straightforward. This complexity is mainly due to the camera motion which induces a global motion in the video in addition to the object's motion during performance of an action. Additionally, it may be due to the action performed at a different speed or the object motion probably observed at different time instants with different temporal extents and under different viewpoint.
The representation of a video sequence as a volume in (x,y,t) space was first pioneered in Buxton and Buxton [13], in which a spatio-temporal gradient scheme is introduced for motion computation and inferring a static scene's depth information. Aldelson and Bergen [14] more explicitly proposed a motion sequence represented as a single pattern in x–y–t 3D space. Since then, the spatio-temporal volume has been predominantly studied in image processing. Bolles et al. [15] first investigated slices of the spatio-temporal volume to recover geometrically static scene structure from motion. Later they exploit spatio-temporal volume for object tracking [16]. Following this idea, other researchers have studied spatio-temporal helix [17], temporal slice analysis [18], oriented energy measurements [19], etc., and applied these concepts to spatio-temporal analysis of video sequences. We give details on this in Section 4.

(b) A separate modelling of spatial relationships (based on spatial logic relations) between object pairs, from temporal modelling based on how these relationships might vary, change in camera position or object movements, position of change in scenes (cut), change in illumination, colour, texture and shape across frames, etc. The information gathered is now fused together either by concatenating spatial and temporal vectors, or through a weighted combination. One option is to keep the information separate once extracted and the SQL type query can be applied—the video that matches the query on the majority of the spatio-temporal features is chosen as the best match. An example is SEMCOG system [20], which represents spatial constraints among objects by using 2D (two-dimensional) string and describes temporal action by using Allen's [21] 13 temporal logic relationships along with distance constraints. Queries use a semantic language—CSQL and VCSQL, which is similar to the standard SQL. These two types of information fusion models can deal with very complicated cases of video retrieval. However, the former is not addressed properly, whereas the latter does not support a comparison by using similarity metrics.

In this paper, we review the state-of-the-art spatial and temporal models with the aim of using these for image and video retrieval. This paper is organised as follows. In the rest of Section 1 we give an overall brief review of spatiotemporal models for video retrieval. Section 2 reviews spatial modelling of video and image data. In Section 3, we discuss research on temporal modelling. Finally, in Section 4, we present a brief review on spatiotemporal information fusion.

2. Spatial information modelling in multimedia retrieval

2.1. Spatial representation

Spatial information can be formulated with the following two methodologies:

- The first approach is to use weak spatial constraints and capture spatial local information to represent low-level texture features. Examples include Gabor wavelets [22], local histograms [23], co-occurrence matrices [24], colour correlograms [25], composite region templates (CRTs) [26], etc.
- The second approach is to represent global qualitative spatial relations that support high-level semantic textual queries. Examples include symbolic projections [27,28], spatial logic [29], θ–R representations [30], etc.

We are more interested in the second type of spatial representation. Spatial qualitative relations between objects are very important for video and image retrieval to support effectively high-level spatial queries. An overview of the major qualitative spatial representation and reasoning techniques is available in Cohn [31]. In the following three sections we discuss three major representation models: (a) 2D strings and its variants (Section 2.1.1); (b) spatial logic (Section 2.1.2) and (c) other models (Section 2.1.3). A number of these models have been inspired by the initial work of Allen [21].

Allen [21] introduced an interval-based temporal logic, which considered objects/events along a 1D (one-dimensional) time axis as a set of temporal intervals based on comparative relations. This differs from point-based approaches, prevalent at that time in the logic and reasoning literature. Allen [21] defined 13 mutually exclusive relations which hold between two intervals: {*before*, *meets*, *overlaps*, *during*, *starts*, *finishes*, and their inverse relations, and *equal*}. Allen's 13 relations can be expressed in terms of at most three order operators ($<$, $>$, $=$). The elegance and simplicity of Allen's temporal interval algebra has inspired several further developments both in temporal and spatial reasoning. It has been formalised as topological relations in 1D spatial domain. It promotes development of symbol projection for spatial image indexing. Lee and Hsu [32,33], for example, represented 13 types of topological relations in 2D-C string, shown in Fig. 1, using the principles of Allen's temporal