学者书屋系列

Video Retrieval Using Spatio-temporal Information

时空视频检索

任 伟◎著

经典

Harbin Engineering University Press
哈尔滨工程大学出版社

Video Retrieval Using Spatio-temporal Information

时空视频检索

任 伟 著

哈尔滨工程大学出版社

内容简介

本书重点挖掘了视频的时空关系，探索了利用机器学习的方法进行视频切割、语义分类。本书分七章，阐明了图像的各种特性，论述了视频的特征，系统介绍了视频的时空逻辑关系、视频的统计分析方法，研究了如何捕捉视频的时空特性，如何利用人工智能神经网络进行视频切割，如何训练计算机“学会”用人类的思维进行视频语义分类、检索。各章节撰写排列体现了从简到繁、由浅入深、从理论到实际、从技术到系统的特点。

本书可以作为高等学校信号与图像处理、计算机科学、机器学习、人工智能、机器视觉等领域的研究生教材和参考书，也可以作为在这些领域从事相关工作的高级科学技术人员的参考书。

图书在版编目(CIP)数据

时空视频检索 = Video Retrieval Using Spatio-temporal Information：英文/任伟著. —哈尔滨：哈尔滨工程大学出版社，2010.4

ISBN 978-7-81133-611-5

Ⅰ.①时… Ⅱ.①任… Ⅲ.①数字图像处理-英文 Ⅳ.TP391.41

中国版本图书馆 CIP 数据核字(2010)第063258号

出版发行 哈尔滨工程大学出版社
社　　址 哈尔滨市南岗区东大直街124号
邮政编码 150001
发行电话 0451-82519328
传　　真 0451-82519699
经　　销 新华书店
印　　刷 黑龙江省地质测绘印制中心印刷厂
开　　本 787mm×960mm 1/16
印　　张 13.75
字　　数 230千字
版　　次 2010年5月第1版
印　　次 2010年5月第1次印刷
定　　价 28.00元
http://press.hrbeu.edu.cn
E-mail:heupress@hrbeu.edu.cn

Preface

The problem of semantic video scene categorisation by using spatio-temporal information is one of the significant open challenges in the field of video retrieval. During the past few years, advances in digital storage technology and computer performance have promoted video as a valuable information resource. Numerous video retrieval techniques have been successfully developed. Most of the techniques for video indexing and retrieval have extended the previous work in the context image based retrieval. In this process, video sequences are treated as collections of still images. Relevant key-frames are first extracted followed by their indexing using existing image processing techniques based on low-level features. For the research in the book the key question is how to encode the spatial and temporal information in video for its efficient retrieval. Novel algorithms are proposed for matching videos and are compared them with state-of-the-art. These algorithms take into account image objects and their spatial relationships, and temporal information within a video which correlates with its semantic class. Also, the algorithms perform hierarchical matching starting with frame, and shot level before overall video level similarity can be computed. The approach, then, is exhaustively tested on the basis of precision and recall measures on a large number of queries and use the area under the average precision recall curve to compare the methods with those in the literature. As a part of this book an international video benchmark Minerva was proposed on which the results have been discussed. The experiments show that the proposed retrieval models are superior in their

performance to the two baseline well-established models used. Also, they are robust to additive noise in data and computationally efficient for testing.

任 伟
2010 年 4 月

Contents

Chapter I Introduction

The main purpose of this chapter is to introduce in brief the research area of video retrieval. A number of approaches have been suggested for effective video retrieval but only a few use spatio-temporal modelling of data. The book discusses in brief the main challenges in this context and introduce a Video Spatio-Temporal Analysis and Retrieval (*VSTAR*) model.

1.1 Motivation

As a result of Very Large Scale Intergration circuit (*VLSI*) technology that is unleashing greater processing power, decreasing cost of storage devices, increasing network bandwidth capacities, and improved compression techniques (Bashir and Khokhar, 2003), digital video is more accessible than ever. Besides professional users, many households today receive digital video information from multiple sources such as cable television, satellite dishes, the world-wide-web, CD/DVD/tapes, etc. In addition, users can create multimedia content using their personal cameras, computers, and 3G mobiles. To help users find and use relevant information effectively, advanced technologies need to be developed for indexing, browsing, filtering, and searching the vast amount of visual content available in video databases. Such techniques are important in various areas of professional and consumer applications such as education, digital libraries, entertainment, content authoring tools, geographical information systems, bio-medical systems, investigation services, surveillance and many others.

Unfortunately, no single approach to video retrieval so far has been shown to be the ideal solution (Aslandogan and Yu, 1999; Naphade and Huang, 2001). Video data is highly diverse and its analysis uses statistical methods whose results lack semantic validity. Hence, a number of video retrieval approaches in the past have investigated

schemes for involving semantic information in video retrieval (Fan et al., 2001; Huang et al., 1999; Liu and Kender, 2000; Liu et al., 1998a; Sudhir et al., 1998; Zhou et al., 2000; Alatan et al., 2001; Lim, 1999; Liu et al., 1998b; Wang et al., 2001; Huang et al., 1998; Weber et al., 2000; Chang et al., 2002; Liu and Hauptmann, 2002; Vasconcelos and Lippman, 1998a; Barnard et al., 2003; Sheikholeslami et al., 1998; Vailaya et al., 1999). Development of efficient semantic features for images and video is an open area of research and much remains to be done. In this book we are motivated to develop a novel spatio-temporal approach that uses semantic information for indexing and retrieving videos. In addition, we are also motivated to provide an international video retrieval benchmark, on the basis of which various studies can be compared. In this overall context, the motivation is to extract semantic features that use spatial relationships between objects in video frames and temporal relationships between frame content, with the main objective of improving video retrieval results.

It is important to note the previous efforts in this area. Extensive research efforts have been made with regard to the retrieval of video and image data based on their visual content such as colour distribution, texture and shape (Aigrain et al., 1996). These approaches are mainly based on still image similarity measurement techniques. Examples include *VisualSEEk* (Smith and Chang, 1996d), *Photobook* (Pentland et al., 1996; 1994), *Blobworld* (Carson et al., 2002), *Virage* video engine (Hampapur et al., 1997), *CueVideo* (Ponceleon et al., 1998) and *VideoQ* (Chang et al., 1998a). The image retrieval techniques allow a user to make queries based on visual image content-properties such as colour, layout, texture, and shape features occurring in the images usually by template matching. Some of these systems also allow the user to make a query by sketching the layout of colour regions or drawing object shape. Feature-based video modelling has been used recently which uses video segmentation and key frame extraction (See Table 2 - 6). After key-frame extraction, these key-frames can be matched using low-level features. Most content-based video retrieval systems have the following limitations.

- *Indexing problem*

The traditional pure feature-based data indexing techniques such as R-tree

(Guttman, 1984), R*-tree (Beckmann et al., 1990), SR-tree (Katayama and Satoh, 1997) and SS-tree (White and Jain, 1996) are unsuitable for video indexing and management because of the curse of data dimensionality, which will adversely impact the methods based on spatial density (Raudys and Jain, 1991; Friedman, 1994).

• *Temporal information utilisation problem*

Current video-retrieval research is based on simply matching key frames across videos, whereas better solutions are needed that match all frames across videos. Because video is a temporal medium, sequencing of individual frames creates new semantics which may not be present in any of the individual shots.

• *Integrated Video Access problem*

a) *Problem with Query-By-Example (QBE)*: *QBE* is widely used in existing video retrieval systems. Query by example approaches are suitable if the user has a similar image or video clip available. However, the query-by-example approach suffers from a critical problem in that an example query video or template video does not fully interpret what the user wants to express. Even though the system can provide plenty of template videos for user selection, there is still a gap between the various requirements of different users. Moreover, when the template key-frame may be taken at a different angle and under a different light condition or has a different scale, query template cannot fully represent actual requirement in reality and the match would not be suitable.

b) *Semantic gap problem*: There is a semantic gap between low-level visual features and high-level visual concepts. Most existing video content retrieval systems ask the user to deliver low-level feature space queries, or choose weight schemes for combining, for instance, wavelet coefficient statistics with Fourier Descriptors, describe cognitive concepts. However, the user may not understand these features. The user does not have confidence in formulating a query. Even though a non-naïve user still finds it difficult to query *JACOB* (Ardizzone et al., 1996c; Ardizzone and Cascia, 1997), because the user is required to specify data-range details that he/she may not know. The naïve user is interested in querying by using high-level semantic keywords rather than using low-level statistical features.

c) *Conceptualizing high-level meaning problem or concept-oriented hierarchical browsing problem*: Hierarchical browsing is another popular way to access video

repositories. However, most existing video retrieval systems such as *VISION* (Gauch et al., 2000), *QBIC* (Niblack et al., 1998), *JACOB* (Cascia and Ardizzone, 1996), *Virage* (Hampapur et al., 1997), *ViBE* (Taskiran et al., 2004) and *CueVideo* (Srinivasan et al., 1999) support hierarchical browsing of a video sequence and do not support concept-oriented hierarchical video database browsing, because of the lack of efficient and standardized mechanisms for video conceptualization (Fan et al., 2004a, Smeulders et al., 2000). The users are interested in hierarchical browsing of the summaries that are presented at different visual concept levels rather than concept-vague visual representations at each level. How to access video databases by conceptualizing high-level concepts need be addressed.

Another way to model video content is to use high-level semantic concepts, using free text/attribute/keywords annotation to represent the high-level concepts of the video content. Examples that describe video database content through manual textual annotation include Little et al. (1993), Weiss et al. (1994), Oomoto and Tanaka (1993), Gauch et al. (2000). However, making annotations is tedious, subjective and time consuming. Using textual annotation or keywords has the following problem.

• *Semantic heterogeneity problem*: Keywords are the most useful for naïve users to specify their semantic query concepts. However, multiple keywords can refer to the same video object class label such as "baby" and "infant". Of course, keywords can also have multiple meanings. There is no standard scene description languages (Szummer and Picard, 1998) which makes it difficult to query video databases.

The recent development of Content-Based Video Retrieval (*CBVR*) systems has made some progress on mapping low-level feature spaces to high-level semantic concepts. However it has only concentrated on some specific knowledge domains such as news video (Bertini et al., 2002), sports video (Kobla et al., 1999b; Li and Sezan, 2002) and face recognition (Martínez, 2000; Raytchev and Murase, 2003). In this book a more generic model of combining spatio-temporal information for video retrieval is proposed. We expect that better results can be achieved if the search is based on the knowledge of objects present within video frames, their spatial relationship and temporal nature of the video.

1.2 Proposed Solution

In the book video content is treated as comprising three levels: basic visual content, basic object content and scene categorisation (see Figure 1 - 1). The object content is regarded as the semantic description of a collection of the basic visual contents. The scene content was further defined as the semantic description of a collection of the object contents. The scene categorization is considered as the global description of videos.

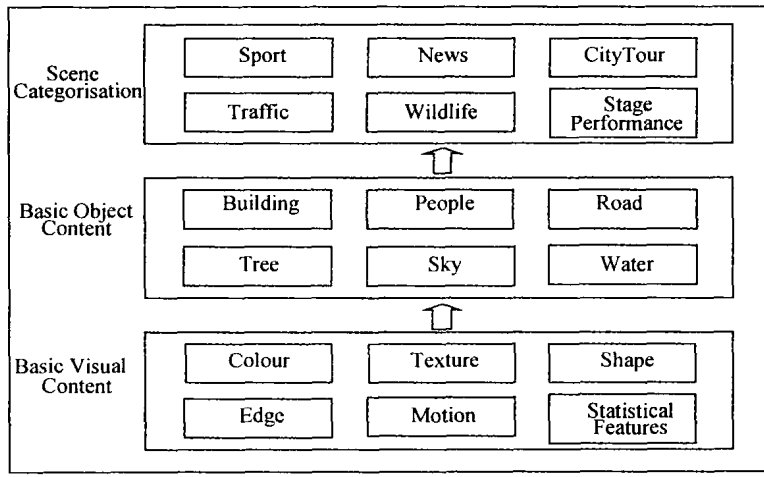

Figure 1 - 1 Definitions of video contents at three levels

Often in queries with video clips, it is desirable to enable high-level semantic queries such as ones involving interesting objects and behaviour (Vilaplana et al., 1998; Vinod and Murase, 1997) or describing a scene by semantic keywords rather than low-level features such as colour, shape, or texture, etc. based on still image comparison. The *MPEG-4* standard (ISO/IEC 14496) provides an ideal data representation scheme for supporting object-based query and retrieval. Video objects can be represented independently of entities in the surroundings or background. The intrinsic properties of video objects such as shape, colour, texture, motion and spatial

coordinates are readily available. In the book, high-level semantic interpretation of video is addressed based on the knowledge of the basic video object content for video retrieval.

In this book two separate schemes for capturing the spatial and temporal relationships are proposed. Spatial data analysis represents the presence and absence of objects and their spatial relationships in terms of distance and context. Furthermore, spatial information also encodes whether there is one dominant object in the scene, whether the object is singular or multiple, and whether the scene is highly cluttered or not. This information can be represented as a string. Similarly, temporal information can be modelled for describing object activity or describing camera activity and effects (camera movement and editing effects such as cut, dissolve and fade). The book has not attempted to analyse object activities (and events) since the aim is to develop retrieval schemes based on video content rather than object behaviour. Instead the book focuses on modelling camera effects that are significant for discriminating between different video genres.

The temporal information can be extracted in the form of a feature string. Spatial and temporal features are then fused into a single feature string that can be matched to other such feature strings. Hence, the key elements of the book can be summarised as follows:

a) Object relationships in video frames are defined by the proximity and directionality in 3D.

b) To avoid the problems of image segmentation and low level image processing (which is not the focus of this study), a semi-automatic method of object indexing is used. Key-frames are automatically found and segmented, and the objects of key-frames are manually indexed (e. g., car, person...). Object tracking based on colour and shape features is used to transfer the labels to regions in between two key frames. With this procedure, we assume that one day the whole process can be automated, if image processing operators were to improve significantly.

c) The method uses a library of predefined objects that are likely to be found in the video. There is a class that labels all other regions as "Unknown".

d) For a test query video Q, finding the best matching video from the retrieval

database is equivalent to choosing the best sample by minimising the feature point distance between Q and retrieved samples. Different similarity metrics can be used for this purpose.

e) The effectiveness of the approach can be tested based on standard measures of precision and recall.

The above points of observation only provide a brief list of the salient features of the book. In this book a *VSTAR* model is proposed. This model is aimed at encapsulating semantic information within a video at both semantic and temporal levels to improve the quality of video retrieval. Chapters 4 and 6 detail the model fully and justify the novelty of this work. The key novel contributions presented are summarised as follows.

a) A 3D qualitative spatial model

A 3D spatial model to cover 169 directional and topological relations between two objects in Chapter 4 is described. The model enables us to represent the semantic information about object relationship in images using a string based feature set.

b) Two spatio-temporal matching algorithms to retrieve videos

Two novel video retrieval models *VSTAR* and $VSTAR_{MCFS}$ (afterward called *VSTAR* models for short) are proposed. The algorithms for video matching using the hierarchical scheme of matching first frames, then shots and videos are completely novel (see Chapter 4). The matching schemes use measures of object similarity, temporal similarity and spatial similarity. All of these measures are mathematically defined and developed specially for this research work.

c) Definition of metrics (*TVI* & *SCS*) for measuring *Video Temporal Order*

The video temporal order is a very important factor to measure video similarity. In Chapter 4, two metrics of video temporal order are introduced to measure and compute temporal similarity after frame matching between shots.

d) The other features are presented in the book, such as:

Automatic video transition detection using a machine learning approach

As discussed in Chapter 2, several studies have attempted to improve the performance of video transition prediction. Automatic video transition is not a trivial task. In Chapter 4, the mainly challenges of video transition prediction are also represented. The book proposed an approach that performs feature fusion based on

statistical and motion feature differences between consecutive frames to predict the transition type. A number of methods adopted for analysis from external sources are tailor-made to work on video data. Some novel features for video segmentation are introduced to increase discriminating power. A neural network was trained and tested, and used to automatically detect video transitions. This predication can be done in real-time, does not require any predefined threshold, and attains high accuracy (more than 90%).

The book defines some features for video segmentation including *b-coefficient* (block-coefficient), *c-coefficient* (cell-coefficient), difference of maximum luminance level and difference of percentage maximum luminance level. Some noval approaches to detect video transitions are also introduced such as applying mutual information to detect dissolve, using camera motion information to increase discriminating power. These are discussed in Section 4.3.1.2. A number of object features are presented in the spatial string that are not traditionally used such as object size, single or multiple, etc. Furthermore, the concept of semantic correlation are also defined between objects which can be represented as a correlation matrix and used in a frame matching process.

As a part of this book, extensive data collection is performed and the Minerva benchmark is introduced for video retrieval. Full details of this benchmark are available from www. paaonline. net/benchmarks/minerva. The benchmark is available for the international research community to use and test their algorithms.

1.3 Structure of Book

This book is laid out in four parts. Part I (chapters 1, 2 and 3) details the state-of-the-art research on video retrieval and spatial, temporal information extraction. Chapter 2 presents a literature review on video retrieval approaches and chapter 3 discusses research on spatial and temporal models for image/video. Part II (chapter 4 and chapters 5) describes the video retrieval algorithms and two baseline models. In Chapter 4 a model for spatial and temporal analysis of video objects is proposed and a spatial-temporal indexing scheme for video retrieval is presented. In Chapter 5, two

baseline video retrieval models are introduced. Part Ⅲ (chapter 6) deals with a machine-learning framework for supervised video retrieval. It also describes the experimental setup and model optimisation, evaluate the proposed algorithm, and compare the performance of the proposed algorithm with that of baseline models. Part Ⅳ (Chapter 7) presents the final conclusions. The book also includes six appendices. Appendix A compares compressed and uncompressed video. Appendix B describes an approach of semi-automatic video annotation. Appendix C represents an object-pair correlation matrix. Appendix D introduces key-frames extraction. Finally, Appendix E describes commonly used audio features.