CHAPTER 1

1. INTRODUCTION

1.1 Background

In the last two decades, the rapid developments in Internet and multimedia techniques have led to the explosive growth of digital media applications. For extracting useful information from massive multimedia data sources, conventional text-based methods rely on adequate and accurate annotation of associated contents. Due to the lack of flexibility and robustness in such annotations, a content-based approach to analysis, indexing and retrieval of media data has attracted much attention. Owing to its flexible nature and huge commercial potential, content-based approaches have been applied in many applications such as digital library, video on demand, telemedicine, etc. [1-4]. In these applications, a fundamental problem is how to extract video semantics and content of interest for effective content representation and delivery. This is highly desired as it could help to automatic the annotation and abstract generation of video content for fast browsing and searching of whole videos. Unfortunately, this problem is still far from completely solved as far as real applications are concerned, although some successful results have been reported.

One basic concern is to benchmark using a huge amount of data, covering a wide-range of content type for consistent and objective evaluations, which is essential to prove the effectiveness of any proposed algorithms. To achieve quantitative evaluations, associated enormous ground truth maps, need to be produced in different levels to satisfy the requirements of specific analysis tasks. Most of these ground truths and quite a few of these evaluations have to be produced manually, which costs a lot of time, labour and money. Fortunately, the TREC (Text REtrieval Conference) Video Retrieval Evaluation (TRECVID) is a widely-acknowledged framework which can be used to help achieve this target [5].

TRECVID provides an independent evaluation of a video "track" in the TREC conference series, an annual worldwide competition event started in 2001 and sponsored by NIST (National Institute of Standards and Technology) in the US with additional support from other U.S. government agencies. The goal of TRECVID is "to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation" [5]. The main tasks of TRECVID are adjusted according to the evaluation progress and the advancement of relevant techniques. In 2007, four main tasks were specified, i.e. shot boundary detection, video summarisation, high-level feature extraction and search. In 2008, two new tasks were added including video copy detection and surveillance event detection whilst shot boundary detection was removed. In 2009, four tasks are included which cover surveillance event

detection, high-level feature extraction, video copy detection and search. As can be seen, all these tasks are focused on semantic video analysis, in which shot boundary detection and video summarisation is respectively considered as nearly completed.

In addition, the work which has completed in this thesis is partly supported by the EU-IST FP6 Funded Research Project "LIVE" ("Live Staging of Media Events", Contract No. IST-4-027312). The two competitions participated in TRECVID are also co-ordinated by the Project. It is worth noting that most of the work in this thesis was submitted to TRECVID, including shot boundary detection in TRECVID 2007 and video summarisation in TRECVID 2008. According to announcements from NIST, very promising results have been achieved in these world-wide competitions which are described in Chapter Three and Chapter Five, respectively.

Although some basic tasks on semantic video analysis have been discussed above, it is necessary to analyse these tasks in detail to motivate the work in this thesis. Accordingly, some existing problems relating to the associated tasks are addressed in the next section.

1.2 Problems

For efficient data organisation and effective processing, accurate and robust shot boundary detection is a fundamental task for further segmentation, summarisation,

annotation and content-based retrieval applications. Due to arbitrary visual appearance, it is difficult to detect shot changes using simple thresholding or applying a single model. As a result, complex modelling is desirable in such a context for improved accuracy and robustness. Herein, both abrupt and gradual transitions will be modelled, where the cut will be further categorised into several sub-classes. All these will be done in accordance with the visual appearances of these transitions and the requirements of TRECVID 2007, a framework used to test the proposed approach.

Problem 1:

To effectively measure the similarity of frame images, fast frame alignment is required, which can be further used to filter motion-caused false alarms from detected shot candidates and register image pairs. Typically, frequency-domain processing using phase correlation has been widely used in such applications. However, it is sensitive to noise and its complexity under fast implementation of 2-D Fourier transform still remains an issue for many applications, where massive amount of data are involved. Consequently, an improved algorithm is required to solve this problem.

Problem 2:

Another interesting problem is to extract content of interests from video and remove redundancies, i.e. video summarisation, to aid efficient content-based representation, analysis, browsing and searching of the raw data. Normally, heuristic rules are applied

in this topic to define meaningful content and remove unwanted junk frames, and these are unlikely to be robust in dealing with a wide range of material, especially in the absence of pre-defined human objects or audio information. Consequently, a general model needs to be developed for effective video summarisation, in particular in dealing with BBC Rush videos under TRECVID 2008 framework.

Problem 3:

To fill the gap between low-level features and high-level semantics, the extraction of highlights from video is a good attempt in such a context. In fact, highlights usually contain human objects under various motion patterns, especially camera zooming events, and they can be detected through a combination of several techniques. Accordingly, these highlights can then be used for automatic annotation and content-based retrieval applications.

Problem 4:

To save a large amount of time decoding video, compressed domain processing is preferable as it avoids the time-consuming inverse discrete cosine transform (IDCT) employed in de-compressing video for analysis. Most of the proposed techniques are implemented in compressed domain, including feature extraction, shot detection, video summarisation, and highlights based video annotation and retrieval, which have proved to be efficient in such a context.

1.3 Research Objectives

In this thesis, the research themes are mainly concerned with semantic video processing and exploring a possible future video annotation and retrieval application, and they were approached as follows:

(1) Algorithm design for effective shot boundary detection from the compressed domain, in accordance with the requirements of TRECVID 2007.

(2) New solution for fast and robust frame alignment to filter motion-caused false alarms in shot detection as well as general image registration applications.

(3) Effective modelling for activity-driven video summarisation without the usage of audio and high-level semantics for general video sceneries, especially the rush videos in TRECVID 2008.

(4) New approach for automatic video annotation and retrieval applications without the manual assistance.

1.4 Research Methodology

Model-based processing is the fundamental technique used in this research, which has been successfully applied to several research problems including shot boundary detection, video summarisation, and content based video annotation and retrieval. Relevant details are presented as follows.

Since there is apparent content changes when shot transition occurs, simple thresholding can be used to examine if such changes are sufficient or above a given threshold. However, this suffers from motion-caused false alarms and fails in dealing with general cases when the level of content changes varies in a wide range. As a result, accurate modelling of these shot transitions are required which includes classifying cuts into sub-categories and modelling several gradual transitions according to their visual appearances. As presented in Chapter 3, such accurate modelling has successfully remove false alarms for effective detection of different shot transitions.

Fast frame alignment is useful in estimating global similarity of images to remove motion-caused false alarms in shot boundary detection, and phase correlation using 2-D Fourier transform is generally employed. However, the complexity and scope still remain issues even under fast implementation, especially for video analysis in which massive amount of data are involved. Besides, 2-D approaches perform less robustly, especially in the presence of noisy data. Therefore, subspace phase correlation is proposed which is insensitive to zero-mean noise and even non-zero-mean noise using its gradient-based extension. Along with an improved subpixel strategy, high accurate and robust results in frame alignment can be achieved and is discussed in Chapter 4.

For effective video summarisation in TRECVID 2008, hierarchical modelling of rush videos in a top-down manner is presented, rather than using heuristic processing. This

model is then utilised for bottom-up based analysis and synthesis in terms of filtering of several junk leaf nodes and activity-driven frame categorisation. In addition, clustering of shots is employed to remove retakes in the video rushes. Since in practice the number of retakes is unknown, an adaptive clustering scheme is proposed on the basis of clustering rules and a hierarchical model. Most importantly, the proposed approach as presented in Chapter 5 does not require high-level semantics such as human objects and audio signal analysis for summarization which provides a more flexible and general solution for this topic.

To fill the gap between low-level features and high-level semantics, automatic extraction of semantics for annotation and retrieval applications is always desired. To achieve this aim, extraction of semantics concepts and events for high-lights based video annotation and retrieval is investigated, where highlights are referred to as human objects under certain camera events, such as close-up events of human objects. Consequently, modelling of skin pixel colours is used for human object detection, and camera motion patterns are determined using the 6-parameter projective camera model. The techniques and the overall system are discussed in detail in Chapter 6.

1.5 Thesis Contributions

The main contributions in this thesis can be summarised as follows in terms of the four research topics including model-based shot boundary detection, fast and robust frame

alignment, summarisation of rush videos, and highlights-based video annotation and retrieval.

Regarding shot boundary detection, the main contributions include i) extracting several novel features from compressed MPEG videos for effective shot change detection; ii) accurate modelling of cuts into sub-categories to enable accurate and robust detection in a three-stage coarse-to-fine process; iii) appearance-based modelling of several gradual transitions; and iv) fast implementation of the whole system in the compressed-domain.

Regarding fast frame alignment, the main contributions include i) deriving of subspace phase correlation (SPC) using 1-D Fourier transform on projected signals; ii) proving that SPC is insensitive to zero-mean noise and gradient-based SPC is also insensitive to non-zero-mean noise; iii) modelling non-overlapped regions between images under registration as non-zero-mean noise and applying SPC for robustness; iv) proposing an improved subpixel strategy for higher accuracy.

Regarding summarisation of rush videos, the main contributions include i) hierarchical modelling of rush videos using formal language techniques; ii) extracting an activity level for cut detection and classifying frames into valid/invalid ones. For valid frames, V-units are extracted as sub-shots containing active frames of high activity levels; iii)

modelling several junk frames for their effective removal; iv) adaptive clustering of retakes; and v) content-adaptive summarisation generation.

Regarding video annotation and retrieval, the main contributions include i) extracting of several features from DC images and motion vectors for shot detection; ii) statistical colour modelling of skin and non-skin pixels for human object detection; and iii) introducing an automatic solution for highlights based video annotation and retrieval.

Please note that most of the work reported in this thesis has been published in peer-reviewed international journals and conferences, and a list of them can be found in Appendix 1.

1.6 Thesis Organisation

In total there are seven chapters in this thesis, and the organisation of the main contents can be summarised as follows.

Chapter Two surveys existing work in terms of shot boundary detection, frame alignment, video summarisation and highlights-based video annotation and retrieval applications. Three stages are introduced for shot boundary detection including feature extraction from compressed/uncompressed videos, continuity signal construction and decision. Regarding frame alignment, frequency domain approaches using phase

correlation are particularly emphasised. For video summarisation, different approaches for the summarisation of general videos and video rushes are discussed. Finally, two aspects of video highlights extraction are discussed which include skin detection for human detection and motion pattern determination.

Chapter Three presents the details of the techniques for the work submitted to TRECVID'07 on shot boundary detections, including abrupt cuts and several types of gradual transitions. According to their visual appearance, cuts and gradual transitions are modelled for effective detection. Cuts are classified into several categories and detected in three stages including pre-filtering, decision-based classification and validation. Three kinds of gradual transitions are modelled for detection which include fade, dissolve and combined cuts, a special case defined in TRECVID.

Chapter Four discusses fast frame alignment, where subspace phase correlation (SPC) along with an improved subpixel solution is proposed. It is proved that SPC is not only effective in estimate frame offsets but also robust to adding-on zero-mean noise and yields higher correlation peaks. In addition, it is further proved that gradient based SPC is robust to non-zero-mean noise, which provides a practical solution in dealing with non-overlapped regions between images under consideration.

Chapter Five provide detailed descriptions of a proposed new algorithm for video summarization, which are also included in the submission to TRECVID'08 on BBC

rush summarization. Firstly, hierarchical modelling is proposed to convert the rush videos from frame-based linear structure to a top-down manner, which is then used to filter junk frames and category frames into active and inactive ones via the extracted activity levels. Secondly, an important concept for denoting continuous active frames, the V-unit, is introduced, which forms the basic element in generating summarised results. Thirdly, adaptive clustering is presented to cluster shots and remove retakes. Finally, experimental results using the TRECVID data are presented and analyzed in detail to obtain objective evaluations.

Chapter Six presents highlights-based video annotation and retrieval, a specific application for semantic video analysis. For video highlights extraction, three separate techniques are discussed including video segmentation, skin pixel classification for human detection, and motion pattern determination. Accordingly, human objects under certain motion patterns are then extracted as video highlights for automatic annotation and content-based indexing and retrieval. As can be seen, the common problem of the gap between low-level features and high-level semantics has been filled in the proposed system. Since all the algorithms are implemented in the compressed domain, the whole system is very efficient.

Finally, Chapter Seven summaries the research works presented in this thesis and also provides some suggestions for future investigation.