CHAPTER 7

7. CONCLUSIONS AND FUTURE WORK

7.1 Introduction

In this thesis, several novel algorithms have been presented under research in the area of semantic video analysis. The completed research work has been found useful in both generating high-quality publications and improving the relevant applications. The potential applications cover a wide range including digital library, video on demand, telemedicine, video database management and multimedia retrieval as well as object and event based intelligent video understanding.

Semantic video analysis is an interesting and challenge research topic, and the main difficulties for automatic and intelligent interpretation of video data can be summarised as follows. Firstly, it is impractical to develop one single algorithm to deal with all videos, especially with arbitrary contents. Even the simple task of shot boundary detection, still suffers from manually introduced artificial editing effects. Therefore, how to make the balance between a general framework and certain specified requirements is always a question in such a context. Secondly, it is hard to implement

intelligent video analysis without domain knowledge. Although some successes have been achieved in extracting low-level features and training various classifiers, ambiguity is unavoidable especially in coping with content-rich video data. In fact, we may also misunderstand the content if relevant background and context information for the video is missing. Therefore, knowledge support seems essential to achieve effective analysis. Thirdly, another problem is content adaptation in semantic video analysis, i.e. how to actively fuse various clues from multiple sources when available. This refers to information extracted from spatial and temporal domain and/or those textual and audio ones. Apparently, additional support from speech and text processing can significantly improve the accuracy and effectiveness of semantic analysis. Although specific information may not be available in the video under processing, the analyser needs to be capable of integrating any meaningful contents in an active way.

Four main research topics relevant to semantic video analysis have been presented in this thesis, which include video segmentation, frame alignment, video summarisation as well as video annotation and retrieval. In addition to the discussions in each chapter, details of the main contributions of this thesis and some suggestions for future investigation are summarised and given in the following sections.

7.2 Main Contributions

The main research objectives of this thesis are semantic video analysis. After a comprehensive survey of existing techniques, four different themes are respectively

discussed, including model-based shot boundary detection, fast and robust frame alignment, activity-driven summarisation of rush videos and highlights-based video annotation and retrieval. The contributions of this thesis in semantic video analysis can be summarised as follows, in accordance with the four main research objectives as presented in Chapter 1.

For Objective 1:

A model based approach has been proposed and successfully applied for shot boundary detection and video segmentation, which can be further highlighted as follows [59, 70].

(i) To extract several novel features as local content indicators in compressed MPEG videos and employ AdaBoost for feature selection and fusion, such that robust shot detection is achieved with a very small set of features;

(ii) To categorise shot cuts into five classes for accurate modelling, plus to propose a three-stage coarse-to-fine process for cut detection including pre-filtering for efficiency and validation using phase correlation on DC images for robustness;

(iii) To model three gradual transitions through analysis of their visual appearances for effective detection of combined shot of cuts, fade out/fade in and dissolve effects;

(iv) To implement the whole system in the compressed domain for efficiency.

Evaluation results for TRECVID test data indicates that the proposed algorithms achieve the best results on cut detection, sixth best on gradual transition detection, and

third best on overall performances [5]. This evaluation also shows that the proposed method is effective and robust on a wide range of video sources, and outperforms many other systems using machine learning approaches like support vector machine (SVM) [5, 59].

For Objective 2:

Subspace phase correlation along with an improved subpixel strategy is proposed for accurate and robust frame alignment and image registration, and the main contributions can be highlighted as follows.

(i) To derive a fast solution using projection-based subspace phase correlation (SPC) to estimate 2-D shifts for frame alignment and image registration;

(ii) To prove that the proposed SPC scheme is insensitive to zero-mean noise than existing conventional approaches using 2-D phase correlation, and it also yields higher peaks than its 2-D counterpart;

(iii) To propose gradient-based SPC and prove it is robust to non-zero-mean noise. To model non-overlapped regions between the images under registration as such non-zero-mean noise and employ gradient-based SPC for effective registration;

(iv) To propose an improved subpixel strategy for more accurate estimation of subpixel shifts.

Comprehensive results using synthetic data with manual subpixel shifts, real MRI data without or with various levels of Gaussian noise and a video sequence have fully validated the effectiveness and robustness of the proposed techniques. The overall performance outperforms several existing techniques as reported in [142,149-151] in terms of mean-squared-error measurement.

For Objective 3:

A novel algorithm is proposed for the summarisation of rush videos, which were also included in the submission to TRECVID'08 on BBC rush summarization. The main contributions are highlighted as follows.

(i) To model rush videos in a hierarchical manner using the formal language techniques to guide further analysis;

(ii) To extract an activity level from compressed videos for effective detection of shot and sub-shots namely V-unit;

(iii) To model three kinds of junk frames for the effective removal of them;

(iv) To define a new set of similarity measurement between frames and shots and propose adaptive clustering to detect retakes;

(v) To propose content-adaptive summarisation generation on the basis of extracted activity levels.

According to the report announced by TRECVID 2008 [5, 74, 129], the proposed method has succeeded in producing very good results. Measured in terms running time, it achieves a processing speed at 6.07 times of real-time video playing and is the 4th fastest system or the 3rd fastest team. Also it is scored 0.57 in terms of fraction of inclusions found in the summary, the 7th best among 43 groups of results. Besides, its summarised videos are 26% shorter than the target size at 2% of the original data set. According to an overall evaluation parameter PF, the proposed system is ranked the 3rd or the 2nd best. Since the proposed method does not require high-level semantics such as human object detection and audio signal analysis for summarization, which provides a more flexible and general solution for this topic.

For Objective 4:

A new approach is presented for highlights-based automatic video annotation and content-based retrieval, and the contributions are summarised as follows.

(i) To extract several features from compressed videos and propose knowledge-supported modelling for effective shot detection;

(ii) To propose an improved scheme for accurate detection of camera motion patterns via eliminating outlier motion vectors;

(iii) To propose statistical modelling of skin and non-skin pixel colours for effective human object detection;

(iv) To propose an effective solution for highlights based automatic annotation and retrieval of video using semantics.

Among all the content-based retrieval applications, there is a well-acknowledged gap between low level features which are provided by the developers and high level semantics which are desired by unskilled users for effective query [53, 63, 80, 116]. To solve this typical problem, extracting semantics for automatic annotation is highly expected, especially in dealing with content-rich video data. Accordingly, the work here is a good attempt to extract semantics from low-level features for video annotation and retrieval, and encouraging results have achieved to successfully retrieve about 80% of the video highlights [130]. In addition, the proposed statistical modelling technique is found surpassing the one in [112] even implemented in the compressed domain.

7.3 Future Work

Although the work which has been presented in this thesis demonstrates a certain level of success in the relevant fields, there still exists some potential for improvement and further investigation. Accordingly, some suggestions are listed as follows.

(i) For shot boundary detection, much scope remains for improving the effective detection of gradual transitions, although there is very limited scope for improving cut detection. Accurate modelling of various special effects such as wipes could be one

solution to achieve this target [50], and another direction could be finding some unified framework to apply learning-based strategies and identify different shot changes simultaneously [2, 10].

(ii) Regarding frame alignment, one can easily extend the subspace phase correlation to other directions than the x- and y- axis. In addition, another interesting topic is how to borrow the concepts of the Fourier-Mellin transform [145-147] to fast estimate zooming and rotation factors.

(iii) For video summarisation, further areas for improvements include more accurate key frame extraction, more junk frames removal, and more inclusion of interesting contents. These would require some new concepts in algorithm design to accurate measure the frame and shot similarity, especially the similarity of temporal sequences under certain spatial and temporal variations. Audio and speech processing may also provide some useful information for this topic.

(iv) In highlights extraction for annotation and retrieval, further investigations could be undertaken for face detection and recognition from the detected skin candidates to improve highlight extraction and semantic video indexing, retrieval and annotation. In addition, how to recognise more semantic concepts and events from general videos is another unsolved ongoing research hotspot for such applications.