CHAPTER 2

2. LITERATURE REVIEW

2.1 Introduction

Following the introduction to existing problems in semantic video analysis in the previous chapter, a review of relevant literature is presented in this chapter. In accordance with the objectives previously specified, the main contents focus on the following four parts including: i) Video segmentation; ii) Video frame alignment; iii) Video summarisation; iv) Highlights based video annotation and retrieval.

In Section 2.2, existing work on shot boundary detection for video segmentation is discussed. Section 2.3 includes a review of different schemes for video frame alignment, especially frequency-domain methods using phase correlation. In Section 2.4, techniques for the summarisation of both general and rush videos are discussed. Section 2.5 provides a comprehensive review of techniques for video annotation and retrieval, where human detection via skin detection and motion pattern detection are particularly emphasised. Finally, a brief summary is given in Section 2.6 to motivate the work done in this thesis.

2.2 Video Segmentation Techniques

Detection of shot boundary for video segmentation is not a new topic. It was originally introduced decades ago to detect abrupt cuts in videos [1-4]. Since then, many techniques have been developed in both the compressed and uncompressed domains. Generally, techniques in the uncompressed domain can be transferred to the compressed domain, though with lower resolution due to block-based representation of the data, with pixel values in the uncompressed domain substituted by DC components in the compressed domain. In general, compressed domain processing is more efficient yet seems less accurate than the methods from the uncompressed domain, however, some people may feel that the latter help to produce more robust results than that in the pixel domain as it can restrain the sensitivity to camera and object motions [23].

In general, there are at least two steps for shot boundary detection, i.e. extracting features in either the compressed or the uncompressed domain to construct dissimilarity metrics between adjacent frames and then making decisions based on these metrics. If the two frames are sufficiently dissimilar, a shot change is declared. Some recent surveys and/or evaluations can be found in [4, 13-15, 24, 25, 38].

In the uncompressed domain, frame difference is usually measured by using features including (absolute sum of) pixel difference [24, 27], colour histogram [1, 11, 13, 23, 27, 44], texture or edge [11, 26, 27], motion [10, 22, 26, 32, 36, 42], inter-frame correlation in frequency-domain [3, 16], etc. Several experiments have suggested that, in most cases, the performance of edge features is inferior in comparison with other features like histograms [13, 24]. However, edge features have been successfully applied in determining flashlight events to further validate detected shot boundaries [27]. Motion features are usually used to obtain accurate frame difference since both pixel difference and edge difference need good motion compensation for precision and robustness [22, 25, 26, 32]. Inter-frame correlation is another interesting measurement of frame difference, in which phase correlation is used to determine frame similarity in the frequency domain [3, 16]. The approach in [16] takes about 2 seconds to process a frame pair. This means the computation cost is too expensive, which has constrained the practical application of this kind of implementation.

Compressed domain processing is highly desirable as it avoids the expensive inverse discrete cosine transforms (IDCT) used in video decoding. More importantly, it can make good use of many intrinsic pre-computed features in MPEG such as motion vectors and block averages for both accuracy and efficiency. In the compressed domain, the most frequent features used are DC-images [19, 34, 39], DCT coefficients [29], macroblock types [20, 31], motion vectors [29, 31], edges [39, 43], active blocks [30], and bit-rate information [24]. Moreover, some work is also presented in the wavelet compressed domain [37].

It is believed that the work in [18] was the first attempt to use compressed domain features for video analysis. In [29], pair-wise comparison of DCT coefficients (in I-frames) is applied to detect candidate transition regions, followed by a second pass for validation by analyzing motion vectors in parsing MPEG videos for shot boundary detection and key frame extraction. In [12], correlation of frame difference and histogram are employed in detecting dissolve and wipe effects. In [19], DC images are extracted and compared in a pixel-wise order for the detection of cut, fade and flashlight. It is found that pixel differences from DC images perform much better against camera and object motions than that from full images as the former is a smoothed version of the latter. In [20], macroblock type information is employed in determining shot boundaries, flashlight, and captions. In [39], an edge image is extracted from AC coefficients and then a two-stage clustering is utilized to detect cut and gradual transitions by considering histogram and pixel differences of two DC images as well as edge energy and several other frame difference measurements. In [30], changed macroblocks between frames are counted, namely active blocks in determining shot changes. In [43], extracting and tracking of edge objects is introduced for detecting gradual transitions.

With features extracted from the uncompressed or compressed domains, a continuity signal can be constructed by comparing frames in a pair-wise comparison scheme or temporal window filtering [1]. The former is efficient but sensitive to noise, and the latter seems more robust but less efficient. In addition, the question of how to determine a suitable window size appears to be a new problem for temporal filtering.

After constructing the continuity signal, shot changes are determined in several ways, including thresholding [20-21, 28-29, 32], (fuzzy) decision making [9, 11, 40], machine learning [1, 10, 30, 39] and model-based approaches [8-9, 25-26, 33-34], etc. Other methods include those using principle component analysis [35], neural network [44, 48], graph partition model [1] and mutual information [17]. Besides, multi-resolution analysis for gradual transition detection can be found in [1, 15].

Here, model based approaches cover two parts, i.e. modelling of visual appearances of shot changes as production or editing effects [2, 16, 20, 26] and statistical analysis of shot transitions [8, 9, 25, 33, 34, 45]. The former is mainly used to detect a specific shape or pattern to match the proposed model such as detection of linear transitions in [2], monochrome frames for fade-out and fade-in detection in [15], U-shape or downward-parabolic for dissolve detection in [1] and spatially well-separated pattern for wipe detection in [15, 46] etc. The latter usually adopts hypothesis test and probability analysis for shot boundary detection, such as Pearson test in [9], extraction of likelihood ratio and homogeneity test in [34]. In the work reported by Hanjalic [25], Bayesian minimum detection error criterion is used in modelling and analyzing shot changes. Statistical modelling and analysis usually needs some prior knowledge and assumptions such as shot length etc. [7, 10, 25], and they may produce unsatisfactory results if these assumptions cannot be met.

2.3 Frame Alignment Techniques

Frame alignment plays a crucial role in the analysis of multi-dimensional visual data in the digital domain, where at least two images captured under different circumstances, such as from different sensors or at different times, need to be aligned for consistent measurement and processing. This can benefit a wide range of applications, including remote sensing, medical imaging, surveillance, robotic vision, super-resolution for data visualization, image mosaicking, video compression and object recognition [132-140, 144]. For more information on image registration, please refer to some recent surveys in [138-139, 154, 158, 164-168]. The primary purpose of frame alignment techniques in this thesis is for accurate and robust measurement of frame similarity, which can be usefully applied to effective detection of shot boundaries and video segmentation.

Among many existing techniques, phase correlation is a well-known technique for frame alignment or image registration and this has also been successfully used in motion estimation, object recognition and other applications [140-155, 159]. The baseline method utilizes the Fourier shift theorem, according to which shifts in the spatial domain correspond to linear phase changes in the frequency domain. Phase correlation is then further extended to estimate changes of rotation and scale using the

Fourier-Mellin transform and the so-called pseudo-polar Fourier transform [145-147]. Furthermore, phase correlation can also be used for affine motion estimation by multi-resolution analysis [148]. However, estimation of shifts between images with high accuracy remains a fundamental problem, in which potential exists for further research and improvement [149-150, 154].

Although pixel-level registration is adequate for some applications, higher accuracy subpixel registration is generally beneficial to most applications [149, 155, 158]. The need for subpixel registration arises from the simple fact that actual displacements between images are oblivious to the discrete grid employed at the image acquisition stage. Additionally, in other applications such as magnetic resonance imaging (MRI), data are usually sampled of non-integer offsets in the spatial Fourier domain before reconstruction and so subpixel registration by phase correlation is a natural approach in such a context [142]. Other application domains that have historically involved subpixel offsets for registration include down-sampling of images [149-151], spatial interpolation [152], interpolation-free [156] and pyramid-based approaches [157].

Although project-based fast motion estimation has been proposed, such as the work reported in [160-162] and [169-170], these methods usually involve tradeoffs between accuracy and efficiency despite of their good performances [170]. However, in this thesis, the robustness resulting from subspace projection via analysis of derived subspace phase correlation will be investigated and its better reliability will be shown in terms of higher peak attained and robust to noise.

2.4 Video Summarization Techniques

Video summarization, in which original videos are represented by either a still-image based storyboard or a short-clip based dynamic skimming, plays essential roles in efficient content-access, browsing and retrieval of large video databases [52-56]. In principle, the essential strategy is to choose the most meaningful parts of video to form the summary while ignoring the less important parts, which are often referred to as contents of interest (COI). Consequently, how to define suitable COIs is inevitably dependent on both the application domain and the users for whom the video is summarized. Among the existing efforts, due to its attractiveness and wide commercialization, sports video summarization, covering soccer, baseball etc., has been intensively investigated [61, 64-65, 76, 85]. Other typical applications can also be found in news [83, 99], surveillance [57], movies [56, 72, 91, 93], home videos [79], and even stereoscopic sequences [96] as well as videotaped presentations [95]. Some general methodologies that can be applied to multiple application domains are also reported [17, 83]. To address the users' preferences, existing work also covered personalized and user-adaptive summarization techniques [61-62, 79]. Recent literature surveys on relevant techniques have been extensively reported in a number of sources [53, 68, 71, 87].

2.4.1 Previous Work in Video Summarization

In general, video summarization contains four main steps including: i) video segmentation; ii) key frame extraction; iii) similarity-based clustering; iv) summary generation. Segmentation is used to partition original videos into small clips (shots and sub-shots) and then ranking these clips for summarization [57, 63-64]. To measure the similarity of clips, a group of most representative frames are extracted as key frames and many techniques have been proposed for key frame extraction [17, 56-58, 62, 89, 95-96]. Meanwhile, the similarity between frames is measured using simple histogram distance in [52, 81, 83, 90-91, 93] and mutual information etc. in [17]. In addition, the problem of selecting COIs including objects and events can be solved by introducing user-attention models and domain knowledge [54, 57, 92, 98]. In some work, graph theory [52, 97, 101] and dynamic programming [55, 60, 99] are applied for either video segmentation or optimal (suboptimal) clustering for summarization. In generating the summarized video, representative techniques include combination of key frames, video segments, or even a complex layout of these frames whose sizes are determined by their contained information [99].

Typically, video summarization is extracted by only using the information extracted from the video, called "internal summarization" [68]. In contrast, "external summarization" techniques employ additional information for interactive processing. The additional information includes manual annotation of the video such as those in

MPEG-7 descriptors [82-83, 92] and knowledge about the users to achieve personalized summarization [61-62, 79]. For internal summarization, audio information is often utilized together with image features [65-66, 71, 77, 83, 85-86, 91-93], among which camera motion [17, 63, 67] and object motion [52, 58, 64, 95, 98] are frequently employed to model the significance of frames for summarization. Some work using text information overlaid to help with the video summarization is also reported [61].

In addition to these low level features, high level semantics are also extracted for more effective summarization. These semantics provide more accurate descriptions of objects and events at a higher level, in which representative techniques include object detection, tracking and event classification [17, 57, 64, 68, 96]. Since the defined objects and events are solely application dependent, such as human objects under a surveillance environment and normal or abnormal events at an airport etc., it is normally difficult to extend these techniques to the task of general video summarization.

2.4.2 Summarization of Rush Videos in TRECVID

Summarization of rush videos in TRECVID has some significant differences to conventional video summarization due to retakes in the unedited raw video sources [68-69]. These retake clips are from the same shot being captured under various circumstances, such as different camera positions or luminance conditions, changed background and even characters. In addition, between or within these retakes there are junk frames which refer to unwanted and meaningless short clips, such as colour bars, monochrome frames in white or black, etc. As a result, to complete video summarization for TRECKVID'08 rushes, retakes of the same shot need to be clustered and junk clips need to be eliminated.

As for video segmentation, shot boundary detection is usually employed [17, 52]. Since unedited rush videos are dominated by cuts, shot boundary detection becomes relatively easier. Normally, histogram and frame differences are measured and decisions are then made via simple thresholding or complex classifying techniques for shot boundary detection. In thresholding, techniques reported include single threshold, multiple thresholds, or even adaptive thresholding, and classifiers can be SVM (support vector machine), or SOM (self-organizing maps), etc. In addition, features can be extracted from the pixel domain for accuracy or from the compressed domain for efficiency.

To remove retakes, clustering of shots is applied by using KNN (k-nearest neighbours), PCA (principal component analysis), SIFT (scale invariant feature transform), agglomerative clustering, etc [69]. In most of these techniques, the similarity of two shots is measured by a combined similarity of each pair of their associated key frames.

These key frames are representative images for each shot and they can be extracted either by sampling in a shot evenly, or targeted selection via certain criteria. The principle of those criteria are to choose frames of high differences from the two boundary frames in the shot or frames at the midpoint between each pair of high curvature points from cumulative frame differences, etc. Since key frames are only separate points in a temporal clip of shots, such clustering method needs to be further enhanced in order to achieve stronger robustness.

To rank segmented clips for summarization, detection of some high-level features is employed which includes event detection, video object extraction, object tracking, face detection, and audio analysis [57, 65, 68-69]. Generally, clips including more human objects are considered to have more importance and hence assigned with higher ranks. Certain feature analysis can be used to remove junk frames, as these unwanted small clips are found to be of some fixed pattern in audio-visual appearances. It is worth noting that audio information can be useful in many aspects in applications, such as shot detection, filtering junk frames as well as clustering of retakes.

2.5 Techniques in Video Annotation and Retrieval

In this section, relevant work on skin and human object detection as well as motion pattern detection is reviewed. Among these techniques, knowledge-supported ones are focused for the extraction of semantics and highlights for video indexing and retrieval.

In addition, as most media data is available in compressed formats, compressed-domain processing is employed to avoid the full computational cost of decoding the whole sequence.

Generally, domain knowledge is essential for the extraction of highlights and automatic annotation of the videos, as semantic contents within the videos are normally context-based. For example, highlights in sports videos, like closed captions, slow-motion replays and special zooms [1, 8-10, 29] are quite different from those of political news, as the former is regulated by game-dependent rules but the latter relies on common understanding of real life. However, it is found that almost all these highlights have corresponding camera motion patterns, such as zooming-in, panning or tilting. According to statistical results from [10], such player close-up highlights occupy 39.1%, 25.1%, 39.4%, 48.1% and 57.6% of their predefined shot classes in tennis, soccer, basketball, volleyball and table tennis games, respectively. Therefore, the detection of such motion in videos with human objects seems a straightforward solution for such a context. Related investigations are discussed in detail below.

2.5.1 Motion Pattern Determination

Camera motion is important for at least two reasons: i) it can help to reduce false alarms in shot detection; ii) it can be utilised for event detection and automatic annotation of video content. In most cases, only pan, tilt and zoom factors are considered in such estimations [1, 11, 105]. Usually, there are two ways of estimating the camera motion, and the difference is whether motion estimation between frames is required, i.e. the algorithm makes use of motion vectors from compressed videos or not. Some examples of motion estimation are given below.

Since motion vectors are already stored in compressed videos, compressed-domain processing to estimate camera motion seems an efficient approach. In Zhang et al. [29], sum of difference between motion vectors and change of signs of motion vectors (across the zoom centre) are used in estimating pan, tilt and zooming factors. In Kobla et al. [116], dominant motion in a directional histogram comprising 8 bins is taken to estimate pan and tilt, and then focus of contraction and focus of expansion are used in determining zoom parameter. In [117], arbitrary camera motion is estimated from MPEG videos by removing outlier motion vectors. In [105], Tan et al. introduced rapid estimate of camera motion from P-frames in MPEG videos. However, it lacks accuracy when large moving objects are present. Consequently, it needs to be improved for more robust detection of camera motions.

2.5.2 Skin and Human Object Detection

Employing skin detection to locate human objects in videos is a straightforward approach owing to the fact that human skin has a consistent appearance which is significantly different from many other objects [114]. Some other common methods of detecting human objects include face detection (using Haar-like features, for example) [119-120] and motion and appearance modelling [121-122].

In general, at least three issues need to be considered in skin classification, i.e. colour representation and quantization, skin colour modelling, and classification approaches. In real applications, some post-processing is also required for the detection and recognition of more semantic events including faces, hands or even special skin patches as naked images, etc.

Although many different colour spaces have been introduced in skin detection, such as RGB or normalized RGB [113], HSV (or HSI, HSL, TSL) [104,109,112, 125], YCbCr (or YIQ, YUV, YES) [126], and CIELAB (or CIELUV) [124], etc., they can be simply classified into two categories by examining whether the luminance intensity component is considered. Due to the differences between the training and test data, various results have been reported: Some people argue that ignoring luminance component helps to achieve more robust detection [115,123,128]; however, others still insist that luminance information is essential for accurate modelling of skin colours [111]. Results on skin detection with or without the luminance component are compared in Chapter 6.

Moreover, it becomes widely acknowledged that training from different colour spaces produces comparable results as long as the Y component is included [111], i.e.

invertible conversion between colour spaces can be achieved [102]. Consequently, choosing a suitable colour space merely depends on the intrinsic requirements of efficiency, rather than effectiveness, i.e. the chosen colour space should have its components extracted from images or videos as simply as possible. For instance, YCbCr and RGB spaces are naturally used in compressed and uncompressed images and videos.

As for colour quantization, various quantization levels have been suggested, such as 32, 64, 128 and 256 [111, 113]. Higher levels mean that more storage space is required hence lower efficiency will be achieved in the detection process. However, there is no well-accepted scheme in such a context. Therefore, the performance under different levels needs to be compared, especially on test data under varying illumination.

To model skin (and non-skin) colours, two main approaches are generally utilized, i.e. parametric and nonparametric ones. The former usually model skin colours as Gaussian or mixture of Gaussian distributions, and the number of components in the mixed model varies from 2 to 16 [124]. Other parametric models include elliptic boundary models, etc [126]. Parameters in the models are usually obtained by the EM (Expectation Maximum) approach [125]. Non-parametric approaches include histogram-based models [111, 113] and neural networks, etc. [111]. In addition, there are also some imprecise models using fixed ranges of thresholds such as the work in

[128] and [109], although the latter also contains a further step to adapt with image content. It is found that histogram-based approaches and neural network based ones usually generate the best results and outperform parametric approaches [111].

With colour models of skin and non-skin provided, skin pixels are usually determined by using Bayesian decision rules of maximum a posteriori, minimum cost and even maximum likelihood strategies [113]. The latter has only skin colour model and is similar to those using a look-up table for decisions whilst the first two also have a model for non-skin colours and thus the likelihood ratio of the pixel's colour in skin and non-skin models are obtained for decision. Other classification approaches include those using linear or elliptic decision boundaries [104, 124, 128]. Nevertheless, one (or more) threshold(s) is (are) then required for such a decision, and unsuitable threshold(s) may lead to quite poor performance.

Furthermore, existing approaches work mainly on uncompressed images and videos, which make them less efficient owing to the fact that most such media is provided in the compressed format and thus an expensive decompression is required before detection. Instead, the work in this thesis is based on MPEG videos, in which skin pixels are detected directly from the compressed domain, avoiding time-consuming inverse DCT transforms and potential applications are fast detection and indexing of human objects in videos. Consequently, it provides an efficient and fast implementation. Comparing with previous work reported in [123] and [108], an optimal threshold of likelihood ratio between skin and non-skin pixels is derived which skip the iterative processing in [108]. Furthermore, even without a dynamic model as introduced in [112], results from sequences under varying illumination still seem very promising.

2.6 Summary

This chapter started with discussions of the shot boundary detection techniques for video segmentation, which includes feature extraction from the uncompressed and compressed domains, continuity signal construction and decision/classification. Although some basic principles such as apparent changes of image contents can be found for shot change detection, these cannot be applied to all cases for video segmentation. Moreover, there are still quite a few exceptions that do not satisfy these constraints, especially when there are overlapping regions in the images or if the changed contents occupy only small portions of the image. As a result, model-based approaches are presented in Chapter 3 for the effective detection of shot boundaries.

Next, relevant techniques for frame alignment or image registration are discussed. The basic motivation for frame alignment is to remove motion-induced false alarms in shot boundary detection as such false detections satisfy a high overall similarity but large inter-frame difference. Frequency domain processing using phase correlation is emphasised for efficiency in such a context. Typically, 2-D Fourier transform is utilized by existing phase correlation approaches in estimating shifts between images. However, it suffers from less robustness and high complexity even under fast implementation when huge amount of data is involved, such as video analysis. Therefore, a subspace extension to 2-D phase correlation is proposed in Chapter 4 for improved efficiency and robustness.

Regarding video summarisation, two main parts are covered namely techniques for general videos and specific ones as for rush videos in TRECVID. The former is discussed in four steps including video segmentation, key frame extraction, similarity-based clustering and summary generation. The concepts of "internal summarisation" and "external summarisation" are also introduced in generating the summarised videos. For rush videos, a survey of relevant approaches was carried out focusing on junk frame removal and shot clustering. Rather than using heuristic processing, hierarchical modelling of rush videos in a top-down manner is presented in Chapter 5 followed by adaptive clustering and content-driven summarisation for effectiveness.

Finally, applications of semantic video analysis are discussed including video annotation and content-based retrieval. In general, these require a certain degree of manual assistance to fill the gap between low-level features and high-level semantics.

Relevant techniques are surveyed in terms of video segmentation, human detection via skin detection and motion pattern classification. In this thesis, however, a new approach is proposed and presented in Chapter 6 to avoid massive labours and ambiguity caused by manual annotation for highlights based video annotation and retrieval.