A Novel Framework for Robust Annotation and Retrieval in Video Sequences

Arasanathan Anjulan and Nishan Canagarajah

Department of Electrical and Electronic Engineering
University of Bristol, Bristol, UK
{A.Anjulan, Nishan.Ganagarajah}@bristal.ac.uk

Abstract. This paper describes a method for automatic video annotation and scene retrieval based on local region descriptors. A novel framework is proposed for combined video segmentation, content extraction and retrieval. A similarity measure, previously proposed by the authors based on local region features, is used for video segmentation. The local regions are tracked throughout a shot and stable features are extracted. The conventional key frame method is replaced with these stable local features to characterise different shots. Compared to previous video annotation approaches, the proposed method is highly robust to camera and object motions and can withstand severe illumination changes and spatial editing. We apply the proposed framework to shot cut detection and scene retrieval applications and demonstrate superior performance compared to existing methods. Furthermore as segmentation and content extraction are performed within the same step, the overall computational complexity of the system is considerably reduced.

1 Introduction

Video annotation is an active field of research in content based video retrieval and summarization. Typically these systems include three steps: video segmentation, feature extraction and indexing. The existing work in video annotation can be divided into two main groups: video segmentation algorithms and content extraction algorithms. Video segmentation algorithms try to divide the video sequences into meaningful subgroups called shots. Over the years, a number of techniques, varying from colour histogram to block based approaches with motion compensation have been proposed for this purpose[1,2,3,4,5]. However an accurate shot cut detection algorithm which works with all kind of video sequences with a single set of parameters is still a challenging problem. Most of the existing content extraction algorithms select one or more key frames as being representative of each shot; feature extraction techniques such as wavelets or Gabor filters are widely used to then extract features from these frames. An efficient key frame selection method, which works with all kinds of videos with little redundancy, is still a difficult problem. Different imaging conditions and camera and object motions make it nearly impossible to represent a shot by a small number of frames without oversampling and thus increasing the complexity and memory requirements of the system. On the other hand, any attempt to

H. Sundaram et al. (Eds.): CIVR 2006, LNCS 4071, pp. 183–192, 2006.

reduce the number of key frames may result in content loss and thus a failure to properly represent the shot. Furthermore, segmentation and content extraction are handled separately in the literature and very little research has been done to perform these two operations within an efficient unified framework. Since each of these techniques use different methods, combining them into a single framework with reasonable computational complexity has been a major problem. In this paper we propose a novel framework for content based indexing and retrieval. This framework allows efficient video segmentation, content extraction and indexing within a single framework.

Early approaches in key frame selection propose to selecting the first frame in each shot as the key frame[6,7]. However one key frame per shot is not always sufficient as there can exist a number of salient changes within a shot due to camera or object motion. Conversely, Ardizzone and Cascia[8] suggest making the number of key frames proportional to the length of the shot. They propose taking a key frame for each second. This approach is likely to oversample the sequence, as the semantic content may not often change that quickly. Zhang et al[9] propose a method to extract key frames based on a similarity measure between adjacent frames. They propose selecting the first frame in a shot as the key frame and compare the following frames with the key frame for content similarity. If a significant change occurs, then that frame is also selected as an additional key frame and this process continues until the end of the shot. The idea behind this method is that any content change between frames suggests significant activity in the shot and should be represented by multiple key frames. Vermaak et al[10] suggest that key frames should be maximally distinct and individually carry the most information. Here the input video is transformed into a sequence of representative feature vectors and this representation is used to define a utility function. A key frame sequence that maximises this function is obtained by a non-iterative dynamic programming procedure.

The initial inspiration of our work is obtained from the work done by Sivic and Zisserman [11,12]. They use local invariant region descriptors to represent key frames. Text retrieval techniques are adapted for fast and efficient retrieval. Local region descriptors are vector quantized into clusters and used as visual ”words” in retrieval applications. The regions obtained in key frames are tracked and any region not lasting at least three frames are rejected. In experiments, they show good performance in scene and object matching. However their system is based on key frames and any failure in key frame extraction will affect their system. As they agree that significant change in imaging conditions may limit the performance of the system because of the limited overlapping regions among key frames. This problem however is overcome in our approach by extracting key features throughout a shot rather than extracting them only from key frames.

In our framework, we propose the use of local invariant region features to develop a highly accurate shot cut detection and content extraction method. Stable features are extracted throughout a shot rather than from a small number of key frames. We propose this approach as an alternative to the key frame method. Local regions are tracked throughout a shot with features being extracted from stable tracks. An efficient method is proposed for region tracking to avoid possible repetition of the features. The proposed framework is robust to camera and object motions and can withstand severe illumination changes, spatial editing and noise. The validity of the framework is established first by testing with different kinds of video sequences, and then by demonstrating superior performance compared to existing methods using well known test sequences such as *Run Lola Run* and *Faulty Towers.*

The rest of this paper is organized as follows. The segmentation and content extraction algorithms are described in section 2. In section 3, we explain the experiments carried out to demonstrate the performance of our framework with various video sequences and show superior performance compared to existing methods. We conclude in section 4 with suggestions for future work.

2 Proposed Framework

In our annotation framework we introduce new methods for cut detection and content summarisation. A novel approach, which was previously proposed by the authors[13], is used in cut detection (Local Invariant Region Based cut detection) based on the consistency of the local regions. In Experiments, superior performance is shown compared to existing cut detection methods. To the best of our knowledge, all the existing content extraction and retrieval approaches for video sequences are based on key frames. In this work, however, stable local features, obtained throughout a shot, are used in content extraction and retrieval applications. The detected local regions within a shot are tracked based on the similarity of the region descriptors in adjacent frames. Each new track at any point within a shot is compared to the existing tracks. This enables regions to be tracked through occlusions, thus avoiding repetition of the features. Once a shot cut is detected, the stable tracked regions are summarised based on the length of the run and used as representative features for that shot. Thus in this method, a shot is represented by the stable tracked features throughout the shot rather than the features from one or more key frames. Furthermore both segmentation and content summarisation are performed simultaneously within a single run through the video sequences.

Our segmentation and content extraction algorithms are based on the concept of local invariant region descriptors. A brief explantation and performance evaluation of local region extraction methods can be found in [14]. We choose *Maximally Stable Extremal Regions* (MSER) algorithm by Matas et al. [15] as it performed well with affine and illumination changes. The *Scale Invariant Feature Transform* (SIFT) [16] is used to obtain the region descriptors in our experiments, as SIFT has been proved to be robust against varying imaging conditions [17].

2.1 Video Segmentation (LIRB)

We define a new similarity measure between adjacent frames based on the consistency of local descriptors. For each frame, local region descriptors are calculated independently by using *maximally stable extremal regions* (MSER) and *scale invariant feature transform* (SIFT). The matched descriptors between the adjacent frames are obtained using the greedy algorithm based on a threshold. The consistency measure (CM) between any two adjacent frames is calculated as follows,

$$CM = \frac{N_M}{N_{Max}} \tag{1}$$

where N_M - Number of matches, N_{Max} - Maximum number of regions obtained in any of the frame.

A high consistency value means that most of the selected regions in adjacent frames are matched and a low value means that most of them are dissimilar. If the consistency is less than a threshold value, a shot is declared. As the local region descriptors are highly robust to affine variation caused by motion and illumination changes, the proposed approach can withstand severe camera and object motions. Furthermore local regions are selected across the entire frame which makes the consistency measurement more robust to noise and spatial edits than existing methods. A more detailed explanation of the shot cut algorithm and examples with extreme imaging conditions can be found in [13].

2.2 Content Extraction and Indexing

We extract features from stable local regions throughout the shot, instead of key frames. This is because the key frame method fails when sudden changes occur in camera movement or illumination. Furthermore, features selected from one or more key frames are not robust enough to adequately represent the scenes in a shot.

The extracted local regions are tracked throughout the shot based on the feature matches between adjacent frames. Some of these regions may disappear in particular frames and then reappear later in the shot. This may happen because of occlusion or failure of the MSER algorithm due to extreme conditions. We call these tracks as discontinuous. A real example of a discontinuous track is given in Fig 1 for a shot taken from the video sequence *Tennis*. Fig 1(a) shows the starting frame and the rescaled frame part to highlight the selected region. Fig 1(b) shows the heighlighted regions in the track. The region in question is tracked from frame 585 to 588 and lost in frame 589 because of the movement of the face away from the camera. However the region reappears in frame 608 and is tracked until frame 613. Although these are two different tracks, they represent the same region, thus giving the same content information. In a content extraction system, these two tracks should be joined and considered as one track. This is achieved as follows. Each new starting track at any point in the shot is compared with all the existing tracks within that shot, to avoid possible repetition of the features due to discontinuous tracks. This also enables tracking of regions through occlusions. For example, consider a frame in the middle of a shot with n tracks, $[t_1.........t_n]$; here the length of the track t_i is m_i frames. Track t_i goes through m_i frames and each point in the track contains a 128 element SIFT descriptor vector. Therefore the i^{th} track can be summarised as, $[\mathbf{d}_1.........\mathbf{d}_{m_i}]$. where $\mathbf{d}$ represents the SIFT descriptor. If a region descriptor, $\mathbf{d}$, obtained in the current frame does not have any matches, then it will be compared with the averaged region descriptor of all existing tracks. For the i^{th} track, the averaged descriptor, $\overline{\mathbf{d}}_i$, will be obtained as follows, $\overline{\mathbf{d}}_i = \frac{1}{m_i}\sum_{i=1}^{m_i}\mathbf{d}_i$. The non matched region vector, $\mathbf{d}$ will be compared with all existing averaged tracks to find the closest averaged track. If the distance between $\mathbf{d}$ and the closest averaged track is less than a threshold then it will be considered as a continuation of that track, otherwise a new track will be formed. In the example shown in Fig 1, the new unmatched region in frame 608 matched with the earlier track from frame 585 to 588 as shown in the figure. Therefore these two tracks are joined together and will be considered as a single track.

Once a shot boundary is detected, the feature vectors in the stable tracks throughout that shot will be averaged and stored. The stable tracks are selected based on the length of the tracks through frames. We select a track if it goes through at least 7 frames. If the total selected tracks is greater than 200 for any shot, first 200 most stable tracks are selected. When a query image is presented to the system, the local region descriptors are obtained for that image and compared with the stored shot features. Based on this comparison, the best matched shots will be selected and presented as the matches. If the best match value is less than a threshold value, then no match will be possible.

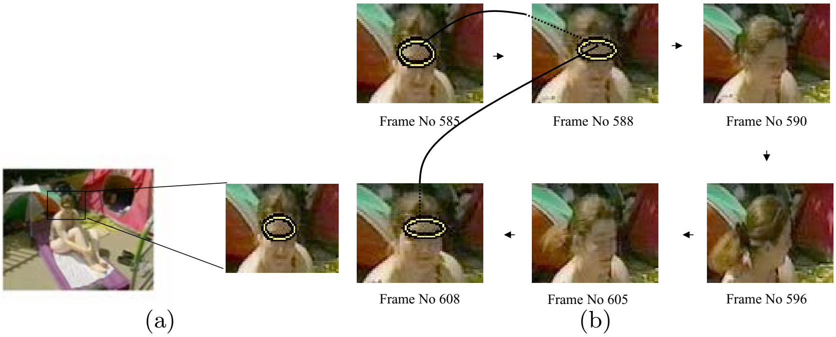


Fig. 1. An example of a discontinuous region track through occlusions (a) Starting frame of the track and the rescaled region for more clear view (b) Rescaled regions in the track. The tracked region is lost in frame 589 because of the movement of the face away from the camera. However, it reappears in frame 608 and joined with the earlier track.

3 Results

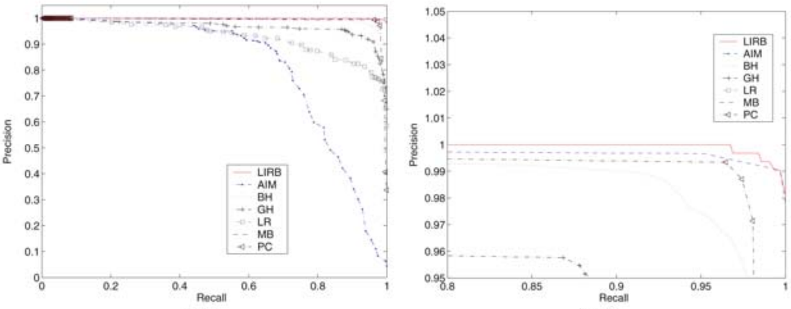
The proposed framework is applied to firstly video segmentation (see 3.1) and secondly scene matching applications (see 3.2).

3.1 Video Segmentation

The size of the test data for the video segmentation experiments is around 43000 frames with 312 shot positions. The test data contains different kinds of video sequences, varying from movies, TV series, documentaries, sports, wildlife and under water videos. Further details about the test data can be found on[13]. The ground truth shot cut positions were manually defined.

To demonstrate the benefits of our approach, we compare the performance of our algorithm (LIRB) with the following shot detection methods: Pair-wise pixel comparison (PC)[1], Block-based histogram comparison (BH)[3], Likelihood ratio (LR)[1], Average intensity measure[2], Global colour histogram (GCH)[1,3], and Motion based correlation method (MB)[5]. The performance of all the algorithms are compared using well established methods such as Precision-Recall (PR) curves and harmonic mean of recall and precision[18].

Fig 2(a) shows the PR curves obtained for the whole data set. The application of the algorithms to a wide range of media content is important as some algorithms tend to work well with particular type of video and give poor results with other types. For each parameter set, the correctly detected, false and missed number of shots are obtained over the whole data set and the PR curves are plotted. A rescaled version of Fig 2(a) is given in Fig 2(b) to more clearly show the performance near recall value 1. It is clear from these results that our algorithm gives an almost ideal performance and outperforms the rest of the methods. For our approach, the precision value is always greater than 0.98 for any recall value. This is because our algorithm is robust to camera and object movements and can withstand severe illumination changes and spatial editing. Other algorithms fail in such conditions as illustrated by the Fig 2.



(a) (b)

Fig. 2. (a) Performance-Recall curve for the whole data set used in the experiment. (b) A rescaled version of Figure (a) to clearly show the curves near recall value 1.

A good detector should give high values for both precision and recall. A more practical approach to experiment this is to use the harmonic mean (HM) of recall and precision[18], which is defined as,

$$HM = \frac{2P \cdot R}{P + R} \tag{2}$$

This value varies between 0 and 1. A higher value (near 1) means good performance in both recall and precision and a lower value means poor per-

Table 1. Performance comparison based on harmonic mean. Algorithms are in the decreasing order of performance. Correct - correctly detected shots, False - false alarms, Miss - missed shots.

Algorithm	**Correct**	**Miss**	**False**	**Recall**	**precision**	**Harmonic Mean**
LIRB	312	0	6	1	0.9811	0.9905
MB	312	0	7	1	0.9781	0.9889
BH	312	0	66	1	0.8254	0.9043
GH	312	0	169	1	0.6486	0.7869
LR	312	0	221	1	0.5854	0.7385
PC	312	0	613	1	0.3373	0.5044
AIM	312	0	6807	1	0.0438	0.0840

formance for either recall or precision or both. In applications like video annotation, missing a shot cut is more severe than having false alarms. Therefore, for such applications, recall value should be 1. In Table 1, we compared the precision and harmonic mean value for all the algorithms at this condition. As seen in the table, our algorithm outperforms all other methods. Our algorithm gives equally good results for both recall and precision values. In other words, our algorithm detects all the shot cut positions while avoiding most of the false alarms.

3.2 Scene Retrieval

We next evaluate the retrieval performance of our algorithm based on the stored stable local features. Given a query image, related shots taken of the same scene should be retrieved while avoiding other scenes. The shots may be taken under different imaging or lighting conditions, such as different camera angles, zooming positions and illumination changes. Furthermore a shot may cover a large area varying from one place to another and the system should be able to handle these variations. Scenes appearing in movies *Run Lola Run* and *Groundhog Day* are used in scene retrieval experiments which is often used by other researchers. In these movies, the same scenes were filmed a number of times in different imaging conditions, making these ideal video sequences for scene retrieval experiments. The ground truth of the similar scenes are selected manually throughout the whole movies. If a similar place (building or road) appears in different shots, we conclude them as similar shots. Examples of frames from similar shots are given in Fig 3 (a)-(d). Each sub figure contains frames taken from similar shots. As seen in the figure, the frames vary significantly both in terms of the imaging conditions and the areas covered.

A frame which contains the scene in question is given as the query. The SIFT features are extracted from the selected MSER regions throughout the frame and compared with the features from all the shots. For normalisation the total number of matched features are divided by the number of features obtained from the query frame. If the normalised value is greater than a threshold, it is

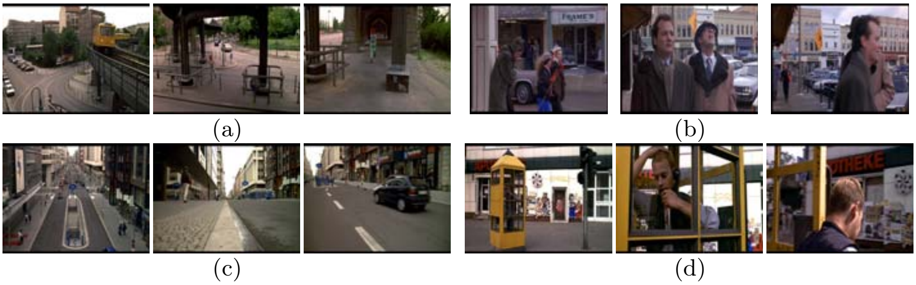


Fig. 3. Examples of frames taken in the same scene. Each of the frame in all the sub figures is taken from different shots taken in the same scene. The frames are varying by imaging conditions and covering different areas.

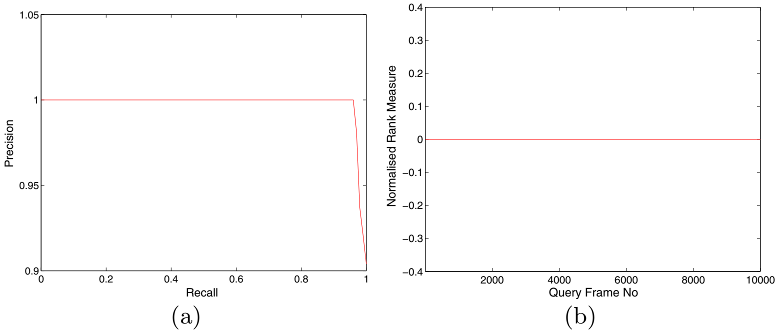


Fig. 4. (a) Average Precision-Recall curve obtained in scene matching applications. 10000 randomly selected frames were used in the experiment. (b) Normalised rank value is plotted for all 10000 randomly selected frames used in precision recall experiment. The rank value is 0 for all the 10000 frames which indicates that all the relevant shots are retrieved as first matches for all the query frames.

presented as one of the matched shots and all the matched shots are ordered in the descending order of normalised matched value.

We use the average PR curve and average normalised rank measure [11] to evaluate the performance of our approach. Fig 4(a) shows the average PR curve. We randomly selected 10000 frames (from movies *Groundhog Day* and *Run Lola Run*) as the query image for the system and the matched shots are obtained. The precision value is calculated as the ratio of the number of correctly retrieved shots to the total number retrieved shots; the recall value is calculated as the ratio of the number of correctly retrieved shots to the number of relevant shots in the database. It is important to note that the scene in some of the selected query frames may appear only in one shot. As seen in the Fig 4(a), our algorithm gives a nearly perfect performance (precision value is more than 0.90 for any recall value). Given an image as query, our algorithm picks up all the shots in the same scene while avoiding any false alarms.

The average normalised rank of the relevant shots can be defined as follows,

$$\widetilde{Rank} = \frac{1}{NN_{rel}}(\sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}(N_{rel}+1)}{2}) \tag{3}$$

where N is the number of total shots, N_{rel} is the number of relevant shots and R_i is the rank of the relevant shot. $\widetilde{Rank}$ is zero if all relevant shots are returned first. The $\widetilde{Rank}$ measure varies between the range 0 and 1 with 0.5 corresponding to random retrieval.

The rank measure is plotted for all 10000 randomly selected frames used in the PR curve experiment, in Fig 4(b). As clearly seen in the figure, the rank value is 0 for all these frames. This indicates that all the relevant shots are retrieved as first matches for all the query frames.

4 Conclusions and Future Work

A novel framework for video annotation based on local region descriptors is proposed. A new similarity measure is developed and the advantages are demonstrated with accurate shot cut detection and scene matching. The proposed method is robust to camera and object motions and can withstand severe illumination changes and spatial editing. The performance is evaluated with different kinds of video sequences and compared with existing methods. The results demonstrate that our method provides significantly improved performance, especially when there are severe object motion, illumination and spatial editing, compared to existing methods. The local regions are tracked throughout a shot and the stable regions are used to form shot representation. The above shot representation gives better results compared to the conventional key frame method and excellent performance is shown in scene matching applications. Future work will consider identifying individual objects in video sequences based on local region descriptors and the current framework provides the foundation for this extension.

References

1. H. J. Zhang and A. Kankanhalli and S. W. Smoliar, *Automatic partitioning of full-motion video*,Multimedia Syetems, vol 1, pp. 10-28, 1993.
2. A. Hampapur and R. Jain and T. Weymouth, *Digital video segmentation*, In Proc. ACM Multimedia, pp. 357-364, 1994.
3. A. Nagasaka and Y. Tanaka, *Automatic video indexing and full-video search for object appearences*, In Proc. Visual database Systems, pp. 113-127, 1992.
4. Y. Yusoff and W. Christmas and J. Kittler, *Video Shot Cut Detection Using Adaptive Thresholding*, In Proc. British Machine Vision Conference, pp. 362-381, 2000.
5. S. V. Porter and M. Mirmehdi and B. T. Thomas, *Video Cut Detection using Frequency Domain Correlation*, In Proc. International Conference on Pattern Recognition, pp. 413-416, 2000.

6. Herng-Yow Chen and Ja-Ling Wu, *A multi-layer video browsing system*, IEEE Trans on Consumer Electronics, vol 44, pp. 842-850, 1995.
7. B. Gunsel and A. M. Tekalp, *Content-based video abstraction*, In Proc. International Conference on Image Processing, pp. 128-132, 1998.
8. E. Ardizzone and M. L. Cascia, *Video indexing using optical flow field*, In Proc. International Conference on Image Processing, pp. 831-834, 1996.
9. H. J. Zhang and Zhong and S. W. Smoliar, *An integrated system for content-based video retrieval and browsing*, Pattern Recognition, vol 30 pp. 643-658, 1997.
10. J. Vermaak and P. Peraz and M. Gangnet and A. Blake, *Rapid summarisation and browsing of video sequences*, In Proc. British Machine Vision Conference, pp. 424-433, 2002.
11. J. Sivic and A. Zisserman, *Video Google: A text retrieval Approach to object matching in videos*, In Proc. International Conference on Computer Vision, 2003.
12. J. Sivic and F. Schaffalitzky and A. Zisserman, *Efficient Object Retrieval from Videos*, In Proc. EUSIPCO, 2004.
13. Arasanathan Anjulan and Nishan Canagarajah, *Invariant Region Descriptors for Robust Shot Segmentation*, Accepted for the Proc. of IS&T/SPIE, 18th Annual Symposium on Electronic Imaging, California, USA, January 2006.
14. K. Mikolajczyk and T. Tuytelaars and C. Schmid and A. Zisserman and J. Matas and F. Schaffalitzky and T. Kadir and L. Van Gool, *A comparison of affine region detectors*, Technical report, University of Oxford, 2004.
15. J. Matas and O. Chum and M. Urban and T. Pajdla, *Robust wide baseline stereo from maximally stable extremal regions*, In Proc. British Machine Vision Conference, pp. 384-393, 2002.
16. D. G. Lowe, *Distinctive image features from scale-invariant key points*, Int. Journal of Computer Vision, vol 60, pp. 91-110, 2004.
17. K. Mikolajczy and C. Schmid, *A performance evaluation of local descriptors*, In Proc. International Conference on Computer Vision and Pattern Recognition, pp. 257-263, 2003.
18. C. J. Van Rijsbergen, *Information Retrieval*, Butterworths, 1979.
19. H. Mullerand and S. Marchand-Maillet and T. Punt, *The truth about corel-evaluation in image retrieval*, In Proc. International Conference on Image and Video Retrieval, pp. 38-49, 2002.