

Rushes summarization by IRIM consortium: redundancy removal and multi-feature fusion

G. Quénot (LIG), J. Benois-Pineau, jenny.benois@labri.fr, B. Mansencal, E. Rossi(LABRI), M. Cord, F. Precioso, D. Gorisse (LIP6-ETIS), P. Lambert(LISTIC), B. Augereau (XLIM-SIC), L. Granjon, D. Pellerin, M. Rombaut (GIPSA), S. Ayache (LIRIS)

ABSTRACT

In this paper, we present the first participation of a consortium of French laboratories, IRIM, to the TRECVID 2008 BBC Rushes Summarization task. Our approach resorts to video skimming. We propose two methods to reduce redundancy, as rushes include several takes of scenes. We also take into account low and mid-level semantic features in an ad-hoc fusion method in order to retain only significant content

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia information systems: video

General Terms

Algorithms, Experimentation.

Keywords

SBD, approximate k-NN clustering, HA clustering, motion features, camera motion, mid-level-features, face detection, audio activity, information fusion.

1. INTRODUCTION

Rushes are raw video footage, which will be edited to get the final version of a movie. This material has one main characteristic: it contains a high level of redundancy. Indeed, due to actors' errors or the director will, there are several takes of the same scene. Furthermore, rushes also contain lots of unscripted parts, unrelated to the storytelling of the movie such as preparation work of the director assistants, director's suggestions, camera setting, clap boards, and undesirable content such as blurred frames, color bars and frames with a uniform color.

As introduced in [1], the aim of the rushes task in the TRECVID 2008 campaign is to find an efficient way to present a preview of rushes showing only the relevant parts of the video, that is undesirable content should be filtered and the resulting summary should contain the relevant events annotated in a Ground Truth.

Furthermore, for the TREC video rushes task a "usability" criterion is defined such as "Summary presents a pleasant tempo/rhythm". There are several approaches for summary building. Starting from a conventional key-framing, up to dynamic summaries with a split screen displaying several shots

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM Multimedia '08, October 27 - November 1, 2008, Vancouver, BC, Canada.

Copyright 2008 ACM 1-58113-000-0/08/1011...\$5.00.

simultaneously [9]. In our approach we resort to the sequential video skimming. Video skimming is a form of video abstraction that tries to compact long videos into short, representative clips, so-called video skims. A number of different skimming approaches have been presented in the literature, combining features derived by analyzing audio, still images and video.

In [10] a rather generic clustering approach with verification of audio consistency was proposed which could not strongly improve the base-line temporal video sub-sampling. Examples of more sophisticated techniques include motion estimation, face detection, text- and speech-recognition. Thus the authors of [11] achieved very good results with regard to redundancy and understandability. This method combines shot boundary detection, junk frame filtering and sub-shot partitioning with removal of redundant repetitions. The summarization of the remaining content is assisted by a camera motion estimator, an object detector/tracker and a speech-recognition tool. Finally, different types of scores are calculated from the obtained information and the decision, if the respective clip is included in the final summary, is taken.

The method we propose aims to reduce the redundancy and to take into account mid-level and semantic features accordingly to the nature of rushes content and annotated ground truth. Indeed the structure of rushes is such that several takes of the same scene are sequentially included into one file. Considering one take as a video shot we will cluster all shots to remove redundancy at a shot level. Then significant events – based on development GT annotation will be detected by fusion of low and mid-level semantic features. The fusion method proposed will retain only significant video segments based on these features thus removing insignificant content. Finally the target summary length will be adjusted by a temporal erosion, simple acceleration or extension of intervals. The paper is organized as follows in section 2 two systems are presented with a different approaches to the initial redundancy removal by clustering, section 3 will present feature extraction tools developed, clustering approaches will be presented in section 4, section 5 describes fusion of heterogeneous features. In section 6 we summarize the results and outline perspectives of this work.

This work was realized by French consortium of Research laboratories IRIM comprising LABRI, LIG, LIP6-ETIS, LISTIC, XLIM-SIC and GIPSA –Lab in the framework of GDR-ISIS CNRS. In the following, the contributions of partners are referenced by the names of participating laboratories

2. TWO IRIM SYSTEMS

IRIM consortium submitted two runs with two slightly different algorithms. The Figure 1 presents an overview of IRIM workflow.

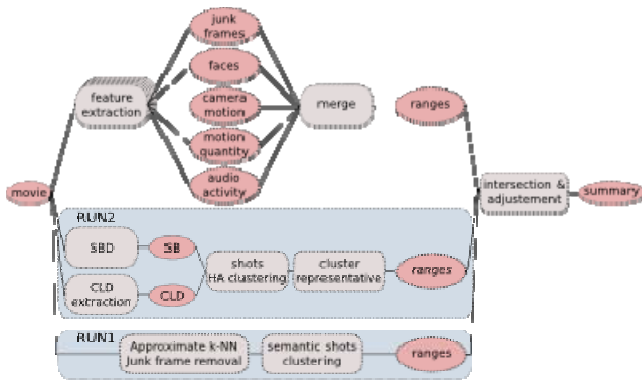


Figure 1: Two IRIM systems

For the two methods, we extract several low and mid-level features from the rushes movie: audio activity, motion activity, junk frames, faces, and camera motion. These features will help to detect important events to keep in the summary. What differs between runs is the redundancy removal algorithm. We use two different methods for clustering into shots. In the first run, we use a method based on approximate k-NN clustering. In the second, it is based on hierarchical clustering. We will first describe the common part of these two runs, such as feature extraction tools.

3. FEATURE EXTRACTION TOOLS

3.1 SBD

One scene take that is a shot is to be logically considered as a basic video unit. This was the assumption in our second run. Thus we need to apply an efficient Shot Boundary Detection (SBD) algorithm. In this work we tried two methods.

The LIG SBD system [5] detects “cut” transitions by direct image comparison after motion compensation and “dissolve” transitions identification by comparing the norms of the first and second temporal derivatives of video frames. It also contains a module for detecting photographic flashes and filtering them out as erroneous cuts and a module for detecting additional cuts via a motion peak detector. The precision versus recall or noise versus silence trade off is controlled by a global parameter that modifies, in a coordinated manner, the system internal thresholds. The system is organized according to a (software) data flow approach.

The ETIS shot boundary extraction method considers cut detection from a supervised classification perspective. Previous cut detectors by classification approaches consider few visual features because of computational limitations. As a consequence of this lack of visual information, these methods need pre-processing and post-processing steps, in order to simplify the detection in case of illumination changes, fast moving objects or camera motion. We are actually combining the cut detection method with our content-based search engine [14] previously developed for image retrieval in order to carry out an interactive content-based video analysis system. The kernel-based SVM classifier can deal with large feature vectors. Hence, we combine a large number of visual features (RGB, HSV and R-G color histograms, Zernike and Fourier-Mellin moments, Horizontal and Vertical projection histograms, Phase correlation) and avoid any pre-processing or post-processing step. We chose the Gaussian kernel for a χ^2 similarity function which has proved its efficiency.

We use a supervised statistical learning approach, requiring a small training set.

Eventually, we used only ETIS SBD system. It gave better recall/precision results on 10 tested files from development set, at the expense of processing time.

3.2 Low –level semantic features

Audio activity – The audio activity gives information about interest of the scene. Indeed, when actors are playing, they speak close to the microphone and the sound level is high, whereas for other parts, such as when staff speaks for instance, sound level is lower. The GIPSA-LAB audio activity detection system performs audio intensity computation using the software Praat [5]. Formally, the intensity I of a sound in air is defined as:

$$I = 10 \log_{10} \left\{ \frac{1}{T \cdot P_0^2} \int x^2(t) \cdot dt \right\}$$

where $x(t)$ is the sound pressure in units of Pa (Pascal), T is the duration of the sound and $P_0 = 2 \cdot 10^{-5}$ Pa is the auditory threshold pressure. In fact, intensity I is computed on effective window length of 0.008s.

From the intensity level I , the audio flow is binary segmented according to its interest using adaptive threshold. The method can be summarized as follows.

- Step 1: the intensity parameter is filtered by a median filter in a window of 100ms. For each video frame, an intensity value is computed as a mean on 40ms window.

- Step 2: the referent level α is computed on a sliding window of 180s. It is defined to insure that for 75% of the window (*percentile*), the energy is upper than the level α which is limited to the interval [57dB, 64dB]. This level calculated each 18s is filtered by a Gaussian filter and interpolated to get one referent level by frame.

- Step 3: the high sound activity is detected if the energy is upper a first threshold $th1 = \alpha + 5dB$.

- Step 4: the high sound activity is segmented from the detected point by defining a window around this point where the energy is upper than the second threshold $th2 = \alpha + 2dB$.

Finally, the video is segmented into high or low sound activity intervals.

Motion activity – The second low-level semantic feature concerns a quantitative evaluation of the motion perceived in image plane. To obtain this measure, the XLIM-SIC uses a three step process all over the entire sequence.

Firstly, we determine the displacement fields between every two consecutive frames. These fields are obtained as solutions of the optical flow equation computed in the native YUV space and using a multi-spectral differential scheme, the flow tensor. This method provides dense displacement fields that are representative of a true color motion.

Secondly, we approximate the displacement fields in finite orthonormal basis of bi-variate polynomials. Each base is a local base as it is obtained by the orthogonalization of the canonic base restricted to the set of points where the norm of the motion vector is higher than a given threshold. The coefficients of the projections of the horizontal and vertical components of a field form constitute a motion signature matrix.

Thirdly, we evaluate the motion activity or the displacement energy as a function of the singular values of the motion signature matrices. The resulting values are then normalized between 0, similar frames, and 1, maximum activity comparable to a black to white cut.

3.3 Mid – level features

Junk frames – “Many color bars, clap boards, all black or all white frames” is the definition given to judge how much “junk” a summary contains. We believe that a frame with a limited number of colors can also be qualified as “junk”. This hypothesis is illustrated in figure 2. All the images displayed exhibit a limited number of color bins in their histograms.



Figure 2: Examples of junk frames

Thus, for junk frame removal, LABRI detects frames with few colors, which according to our assumption will contain uniform color frames and color bars. Our method is based on a thresholding of a color histogram. We apply a threshold on each channel in RGB color space, and classify frames with a limited number of colors as junk frames. When the images are blurry – such as that one on the right in Figure 2, then the histogram is flattened and a thresholding will lead to miss-detection. We thus also apply the same algorithm on downscaled frames (to resolution 8x8) in order to detect diffuse color bars. Indeed these images can be considered as having undergone a low pass filtering. If only “few” significant colors are obtained either on full – resolution histogram or on the histogram of a downscaled frame, then the junk frame is detected. This method gives good results for frames with few colors: we get recall=0.86 and precision=0.88 on 20 tested development videos.

However, this method is not designed to detect clap boards, which should also be considered as junk frames.

In the second system, a supplementary filtering method was applied for detection and removal of junk frames. Here, about one hundred junk frames were selected from development set and near duplicates were detected and removed by approximate k-NN clustering (described in section 4).

Faces – Most of the events in the Ground Truth for development movies are related to human postures or actions. We use skin color detection to acknowledge of human presence. Nevertheless, as we stated in our previous research, the variation of skin color is significant across heterogeneous content sets and it is difficult to get a good training set without the risk of false detection. Thus the ambition of this method consists in training skin color on the video content item to process.

Hence our method makes two detectors co-operate. First we use the face detection algorithm developed by Viola and Jones, extended by Lienhart base, based on Haar-like features, implemented in OpenCV [6]. Then the detected faces are used to train a skin color appearance model, as presented in [7].

OpenCV face detector, trained on frontal face pose, is applied on whole movie, at I-frame resolution. The results are filtered by a geometric filter. The input of this filter is the ratio $r_{i,i+1}$ of

intersection of face bounding boxes in consequent frames I_i and I_{i+1} . The filter length is of 5 frames. On these filtered results, a color based detector is trained using a mixture of Gaussians. Finally, we apply the color-based detector and filter the results with the same median geometric filter.

The color-based detection allows us to increase recall, without decreasing precision. However, on 10 tested development movies, with an overlapping of at least 50% between found faces and ground truth, we get recall and precision only around 50%. Nevertheless the results were in average 2-3% better than those supplied by OpenCV detector only.

Camera motion – Camera motions are also parts of the events in the ground truth. Indeed the director often uses camera motions, like zoom or pan for example, to highlight important events in a scene.

For camera motion classification, we use the algorithm described in [8]. First global camera motion is estimated from motion vectors of P-frames of MPEG compressed stream. Then, a likelihood significance test of the camera parameters is used to classify specific camera motions. The algorithm allows for classification of camera motion as pure physical motions: “pan/traveling”, “tilt”, “zoom”, “rotation” or complex motion.

We consider pan/traveling, tilt and zoom as the camera motion of interest. We discard rotation and complex motion. Thus we hope to keep scripted camera motion and remove the majority of unwanted motions.

On 10 development videos, we get recall=0.60 and precision=0.44. The low precision seems due to insufficiency of MPEG 1 macro-block optical flow. Nevertheless, for the sake of computation efficiency, we did not re-estimate motion vectors on decoded frames.

These mid-level features have also been used in the COST292 group submission.

4. REDUNDANCE REMOVAL BY FRAME AND SHOT CLUSTERING

The ETIS summarization approach, used in the first run, is based on the approximate k-NN method, implemented with Locality-Sensitive Hashing (LSH) [13]. Here 64-bin HSV color histograms are used as feature vector. We only consider every 4 frames in the video.

Then, we define a label for the first frame and propagate this label to all the frames among its k-Nearest Neighbors. We repeat this process for all the extracted frames of the video. From this first coarse clustering, we extract video shots using a vote on the label occurrence. We call this process a “semantic shot clustering”. We, then, erase too short shots and remove redundant video sequences from extracted shots in order to keep the longest representative of each semantic shot.

The LaBRI redundancy removal method is based on a hierarchical agglomerative (HA) clustering algorithm as in [12] with a complete-link distance. The stopping criterion in the dendrogram is the increase of coding error. We extract Shot Boundaries with ETIS method described in section 3.1. We then compute a Color Layout MPEG7 descriptor for 4 frames of each shot and use these as feature vector for clustering. Finally, we keep the longest shot as a representative of each cluster, as we believe it will contain most of the scene events.

5. FUSION OF HETEROGENEOUS FEATURES

Fusion of heterogeneous features is a classical approach [3] [4] to improve a content-based multimedia analysis. The fusion strategy proposed by LISTIC is based on an initial stage which is the joint observation of a set of training sequences with all the available features. It comes out the following elements:

- Three data seem essential to the summary construction: the motion activity (for each image it is a 0-1 normalized data denoted *Motion* in the following figure), the sound activity (binary data provided by an adaptive thresholding, denoted *Sound*) and the camera motion (denoted *Cam.*).

- Information on the presence (and number) of faces (denoted *Faces.*) and more generally on the detection of other concepts is not easy to use. Indeed, the confidence is very variable and it is difficult, without a priori information on the sequence content, to know if a summary must or not contain a face, a car, etc.

- The relative duration during which each feature has a significant value is essential for the fusion process: a feature frequently relevant could be fused with a conjunctive operator, whereas a rarely relevant one will be fused with a disjunctive operator.

After junk frame elimination, the fusion is designed in the following way, as reported on fig.3.

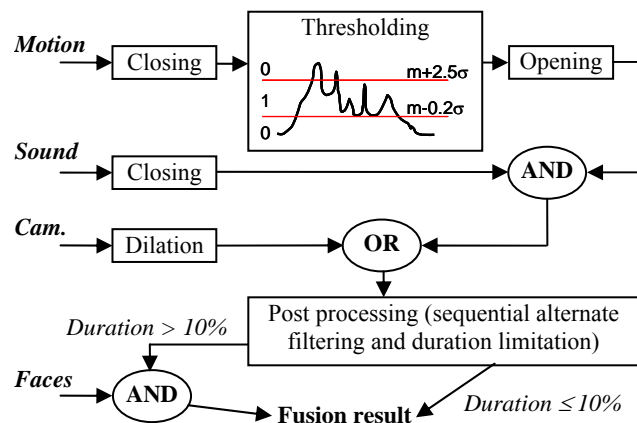


Figure 3: Fusion of heterogeneous features

In this fusion scheme, different mathematical morphology filters are used to smooth the information (elimination of too short information intervals, merging of close information intervals, etc.). These filters are adjusted in an empirical way.

The motion activity is thresholded in an adaptive way by keeping only the data ranging between a low threshold and a high threshold. Information lower than the low threshold corresponds to static scene and is not interesting for the summary, and information higher than the high threshold corresponds to undesirable motions or to cuts.

With these binary data, the fusion consists in a logical AND operator between motion and sound activity data, followed by a logical OR operator with camera motion data. To reduce too long summaries, faces data is introduced using a logical AND operator.

Once fusion of features has been done, it happens that total duration of kept ranges is different than the targeted 2% of the original video. If these ranges are far below 2%, there are chances that we missed some events. In this case, we decided to slightly extend each range, by adding frames at the beginning and at the end, taking care of not adding junk frames. On the contrary, if ranges from fusion are superior to 2%, we erode each range. The erosion consists in removing some frames at the beginning and end of each range, without reducing its duration under 1s (i.e., 25 frames as the frame rate is of 25fps). If after this first step the summary duration is still too long, we accelerate ranges till it is necessary, reducing the frame rate by the factor of 2.

6. RESULTS AND PERSPECTIVES

As presented in [1], eight key criteria are used in the NIST TRECVID summary evaluations:

- DU – duration of the summary (secs)
- XD – difference between target and actual summary size (secs)
- TT – total time spent judging the inclusions (secs)
- VT – total video play time (versus pause) judging the inclusions
- IN – fraction of inclusions found in the summary
- JU – Summary contained lots of junk
- RE – Summary contained lots of duplicate video
- TE – Summary had a pleasant tempo/rhythm

For the DU and XD criteria, our results are not good. For DU, our rank is 41st for run1 and 43rd for run2. Indeed, we were most of the time slightly superior to the target 2%. There are two reasons for this. First, our fusion step produced, most of the times, ranges that globally bypassed the target 2%. Thus we had to reduce their duration. However, we then aimed to have a number of frames strictly inferior to 2% of original number of frames. But we did not take into account that re-encoding would add some frames.

For TT, run1 is ranked 23rd, run2 is ranked 35th. For VT, run1 is 10th and run2 is 29th. The better results for judging run1 can be explained by other remaining criteria, in particular RE and TE.

For IN, run2 is 23rd, run1 is 27th. Both runs are close but for once, run2 is slightly better than run1. This can be explained by the fact that actually clustering in run2 kept far more things than run1 (see RE criterion).

For JU, run1 is 15th and run2 is 22nd. In the first run, the filtering of junk frames was also performed by k-NN-based method using junk frames from development set. Thus stronger filtering resulted in the better removal of junk units.

For RE, run1 is 16th and run2 is 34th. This criterion shows that our hierarchical clustering for the second run was not efficient. For a given movie, it certainly kept several repetitions of the same scene in the same cluster. Indeed we notice that some rushes content items in the development set contained continuous takes for the same scene. As the shot with the longest duration was chosen as cluster representative, we had a higher probability to keep repetitions.

For TE, run1 is 18th and run2 is 27th. It is certainly correlated to repetitions.

Considering system effort, the IRIM runs are rather effort consuming: 15267.8 for IRIM 1 and 41556.6 for IRIM2 with 41st and 43rd places respectively. This is due to the completeness of the feature space proposed by ETIS which was used both in the first run for semantic shots extraction and in the second run for the SBD. Furthermore, the use of Gaussian kernels with highly dimensional feature vectors is time consuming. Nevertheless it gave the best results on development set and thus it was retained for our systems. Furthermore, the program used in the submission was developed in Matlab without optimization. So far, a few simple algorithmic modifications have allowed a speed-up factor of 10, and we are confident that a factor of 100 will be obtained after the program is translated to C, in a release we are planning for the near future.

Hence in this work we proposed two systems which both used low-level and mid-level semantic features and information fusion. The initial redundancy removal at shot and frame level were different. We think that the proposed framework will lead to creation of more faithful summaries, with better inclusion of GT events, if more semantic features can be added, such as detection of humans in close-up shots, incorporation of more rich audio features. For the redundancy removal, approximate k-NN clustering seems more promising. Furthermore, filtering of content with specific concept detection (such as clap boards or staff in unscripted parts), taking into account the target duration of summaries directly at the information fusion stage will allow us to better reduce the redundancy of the summaries.

7. ACKNOWLEDGMENTS

This work was supported by French research consortium GDR CNRS ISIS.

8. REFERENCES

- [1] Over, P., Smeaton, A. F., and Awad, G. 2008. The TRECVID 2008 BBC rushes summarization evaluation. TVS'08: Proceedings of the International Workshop on TRECVID Video Summarization.
- [2] Ayache, S., Quénot, G. and Gensel, J. 2006 CLIPS-LSR Experiments at TRECVID 2006. TRECVID'2006 Workshop. Gaithersburg, MD, USA.
- [3] Li, Y. and Kuo, C.C.J., 2003. Video Content Analysis using Multimodal Information. Kluwer Academic Publishers.
- [4] Kleban, J., Sarkar, A., Moxley, E., Mangiat, S., Joshi, S. and Kuo, T. 2007. Feature fusion and redundancy pruning for rush video summarization. Proceedings of the international workshop on TRECVID video summarization, Augsburg, Germany, 2007, 84-88.
- [5] Praat, <http://www.fon.hum.uva.nl/praat/>
- [6] OpenCV, <http://opencvlibrary.sourceforge.net>
- [7] Don, A. and Carminati, L. 2005. Detection of Visual Dialog Scenes in Video Content Based on Structural and Semantic Features. International Workshop on Content-based Multimedia Indexing (CBMI).
- [8] Kraemer, P., Benois-Pineau, J. and Gracia Pla, M. 2006. Indexing Camera Motion Integrating Knowledge of Quality of the Encoded Video. Proceedings of 1st International Conference on Semantic and Digital Media Technologies (SAMT).
- [9] E. Dumont, B. Meriardo, "Split-screen Dynamically Accelerated Video Summaries", in Proc of 15th international ACM conference on multimedia, September 24-29, 2007, Augsburg, Germany
- [10] A. Hauptmann, M. Christel, W.-H. Lin, B. Maher, J. Yang, R. Baron and G. Xiang, "Clever Clustering vs. Simple Speed-Up for Summarizing BBC Rushes". Proc. TRECVID BBC Rushes Summarization Workshop at ACM Multimedia 2007, Augsburg, Germany, September, 2007
- [11] F. Wang and C.-W. Ngo, "Rushes Video Summarization by Object and Event Understanding", Proc. TRECVID BBC Rushes Summarization Workshop at ACM Multimedia 2007, Augsburg, Germany, September, 2007
- [12] S. Benini, A. Bianchetti, R. Leonardi and P. Migliorati, "Extraction of Significant Video Summaries by Dendrogram Analysis," in Proceedings of International Conference on Image Processing ICIP'06, Atlanta, GA, USA, October 8-11, 2006
- [13] A. Andoni, M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, "Locality-Sensitive Hashing Scheme Based on p-Stable Distributions", in the book "Nearest Neighbor Methods in Learning and Vision: Theory and Practice", T. Darrell and P. Indyk and G. Shakhnarovich (eds.), MIT Press, 2006
- [14] G. Camara-Chavez, M. Cord, S. Philipp-Foliguet, F. Precioso and A. de Albuquerque Araujo, "Robust Scene Cut Detection by Supervised Learning", in Proceedings of EUSIPCO 2006, Firenze, Italy, 2006