SUMMARIZING RAW VIDEO MATERIAL USING HIDDEN MARKOV MODELS

Werner Bailer, Georg Thallinger

JOANNEUM RESEARCH Forschungsgesellschaft mbH
Institute of Information Systems & Information Management
Steyrergasse 17, 8010 Graz, Austria
{firstName.lastName}@joanneum.at

ABSTRACT

Besides the reduction of redundancy the selection of representative segments is a core problem when summarizing collections of raw video material. We propose a novel approach for the selection of segments to be included in a video summary based on Hidden Markov Models (HMM), which are trained on an annotated subset of the content. The observations of the HMM are relevance judgments of content segments based on different visual features, the hidden states are the selection/non-selection of content segments. The HMM is designed to take all relevant scenes into account. We show that the approach generalizes well when trained on sufficiently diverse content.

1. INTRODUCTION

Our work is motivated by the problem of content management in digital cinema and video post-production. After shooting, a large amount of raw material ("rushes") must be organized and viewed in order to identify the small fraction that will be used in the final content. Video summarization tools can support users in viewing this content and selecting segments of content for use in a production. In previous work we have explored methods for reducing the redundancy of the content, mainly by clustering multiple similar takes of one scene [1]. Once redundant content has been identified a crucial task is to decide which segments from the remaining content shall be included in a summary. The aim of this work is to develop a novel approach for extracting relevant segments to be included in a video summary, which does not rely on knowledge about the video domain or heuristics, but can be trained on ground truth annotations or user interactions. In the context of this paper "segment" denotes a temporally coherent clip of the input video of arbitrary length and not necessarily linked to any properties of the input video (such as shot boundaries). As a video has temporal continuity the selection of each of these segments is not independent. We thus do not treat this problem as a classification problem of segments or subsegments but as a sequence labeling problem.

The rest of this paper is organized as follows: Section 2 discusses machine learning approaches for summarization and content selection, focusing on HMM based methods. In Section 3 we present our Hidden Markov Model (HMM) based approach for segment selection. Section 4 presents evaluation results on the TRECVID [2] 2007 rushes data set and we conclude in Section 5.

The research leading to this paper has been partially supported by the European Commission under the contracts FP6-045032, "SEMEDIA" and FP6-027026, "K-Space".

2. RELATED WORK

Many summarization approaches use some heuristics to determine relevant segments, e.g. segments with activity, interesting camera motion, humans present in the scene etc. Approaches using machine learning techniques apply them to event or concept detection and then include segments containing certain events or concepts in the summary. There are few approaches that use machine learning techniques to build a model of relevant content from annotations or user interactions and apply it to content selection. Some of these approaches are specific to certain domains.

In [3] the authors use HMMs for news story segmentation. The states of the HMM correspond to start/end of a news story, advertisements and other. The observations are based on features from audio and video as well as closed caption data. The authors of [4] use HMMs to classify shots of sports video into 15 classes in order to summarize it. In [5] a HMM based approach for classifying segments of football video into play, break, focus and replay is presented. An approach to event detection (e.g. explosions in action movies) using HMMs is presented in [6]. The work in [7] presents a HMM based approach using compact chromaticity signatures for key frames and scene durations/transition for topic classification (news, commercial, basketball, baseball). The authors of [8] present an approach for classifying segments of video documentaries into commentator, picture and video clip.

A number of papers deal with genre classification using HMMs, e.g. [9]. A hierarchical approach for HMM based semantic analysis of video is presented in [10]. In this work several layers of HMMs for event, segment and genre detection/classification are used.

Other approaches are not limited to a certain domain or genre or perform genre classification, but classify shots into a set of content types. In [11] a HMM is used to classify shots into establishing, dialog and master shots based on the size of the largest face in the shot (master, medium, close up). In [12] a summarization method is presented which is based on classifying video segments into dialog and high/low action content, with transition states in between. The features used are face detection, audio segmentation (silence, speech, music) and zero crossing rate, location change and motion activity. The approach in [13] is based on the concepts of the scene transition graph (STG) and logical story units (LSU). Each hidden state of the HMM corresponds to a cluster in a LSU. The symbols in the observation sequence are the individual shots of the videos. The transition probabilities between the story units are calculated from the data (edges in the STG) and the observation probabilities are given by the relative motion activity. In [14] the video summarization problem is solved using a variant of one-class SVMs called fuzzy one-class SVM that take importance measures derived from

video features into account. Parameters such as the number of segments and the segment length are also considered in the fuzzy membership function.

The authors of [15] present a method for training a HMM on the user's browsing behavior. A set of ten browsing states is defined and a HMM is trained for each user using unsupervised learning. Video previews (in contrast to summaries they are not assumed to be complete) are generated from this data, i.e. video segments watched while the viewer is in "interesting" browsing states are selected. The selected segments are long enough that complete topical phrases in the audio are preserved. In [16] a video summarization algorithm which selects segments based on *ShotRank* is presented. This is a relevance measure that is gathered by logging the video segments viewed when browsing a video repository.

In [17] a HMM based approach for relevant passage extraction from text documents is presented. It deals with the problem of extracting coherent, variable length text segments of documents which are relevant to a query. The model uses five states, two of them corresponding to a relevant segment (one trained on relevant observation words from the query the other on non-relevant ones that might appear within a relevant passage). The method extracts exactly one relevant passage per document.

3. HMM FOR SEGMENT SELECTION

We formulate the problem of selecting relevant video segments from an input video as follows. Let $F = (f_1, \ldots, f_n)$ be a sequence of feature vectors representing a video, where $n \leq$ the number of samples (equivalent to frames in the visual modality) in the video, i.e. the sequence describes the original video or a temporally subsampled one. F can contain arbitrary features derived from the audiovisual content. In our work F represents relevance and redundancy values derived from different visual features for a sample of the video, but one could apply the same approach to raw feature values (late fusion). For details about the features and how the approach is used in the context of a summarization system see [18].

Our approach is based on Hidden Markov Models (HMM). The feature sequence F is the observation sequence of our HMM. The sequence of hidden states $Q = (q_1, \ldots, q_n)$ represents the selection state of a sample (e.g. redundant, relevant, selected for inclusion in the summary). Each q_i takes a value out of a possible set of state labels (as defined below).

As we use a first-order Discrete HMM (DHMM) we have to map the input feature vectors to discrete scalars. This can be either done by assigning a fixed number of bits to each of the elements of f_i or by vector quantization (VQ). We have tried VQ with different number of bits per sample, but as this improves the result only very marginally at significantly higher computation cost we use the simpler approach.

The same content analysis algorithms are applied to both the training and test set to yield the feature sequences F for each video in the sets. For the training set, the sequence Q of hidden states is created from a ground truth annotation. Without loss of generality we can treat a training set consisting of several videos as one long sequence. The ground truth annotation labels segments as relevant or non-relevant (i.e. we can derive a binary relevance value for each sample) and contains information about scene boundaries (i.e. which parts are raw material for the same scripted scene).

A straight forward approach would thus be to use a model with two possible hidden states. It becomes clear that not all relevant content can be included in the summary (e.g. repetitions need to be discarded). Thus we need to distinguish between two types of relevant segments: those that shall be selected and those that shall not. We cannot simply mark the segments we do not want to select as non-relevant, as we assume that the two classes correspond to different observations and this would spoil the model parameters. We thus define our model as follows:

non-relevant (N, state 1) A sample of the video that does not lie within any of the segments marked as relevant in the ground truth.

relevant (R, state 2) A sample of the video that lies within one of the segments marked as relevant in the ground truth, but shall not be selected for inclusion in the summary.

selected (S, state 3) A sample of the video that lies within one of the segments marked as relevant in the ground truth and shall be selected for inclusion in the summary.

The HMM is ergodic, i.e. changing from any state to any other is possible. It is a priori not clear how to distinguish between relevant and selected segments in the ground truth. We use the longest of the relevant segments of each scene as this implies the least risk of discarding relevant content. We can then train the HMM and estimate the transition and observation probabilities.

A major problem of this 3 state model is that we have very little control over where in the video the segments are selected. The amount and length of relevant segments can be controlled by the annotation of the ground truth, but cannot ensure that the selected segments represent all the scenes in the content set equally well. For example, the set of selected segments may contain several ones from one scene, but none from the other scenes.

A comparable problem is addressed in [17] for text documents, where the authors design a HMM which ensures that exactly one text passage is extracted from each document of a query result set. We adopt this idea in order to better control the segment selection behavior. We include scene boundaries into the model and constrain the selection to zero or one segments per scene. We thus need the estimated scene boundaries in the observation sequence of the test set as additional input. The set of segment boundaries in the observation sequence of the test set is initialized from the detected shot boundaries. Then all boundaries between shots that the retake detection algorithm [1] has assigned to the same scene are removed.

We thus extend the HMM to include a specific state for segment boundaries and we distinguish between relevant and redundant content before and after the selected segment (in order to enforce the selection of zero or one segments). The HMM has then the following states:

non-relevant ($\mathbf{N_{pre}}$, state 1) A sample of the video that does not lie within any of the segments marked as relevant in the ground truth and is before a possible selected segment from this scene.

relevant ($\mathbf{R_{pre}}$, state 2) A sample of the video that lies within one of the segments marked as relevant in the ground truth, but shall not be selected for inclusion in the summary, and is before a possible selected segment from this scene.

selected (S, state 3) A sample of the video that lies within one of the segments marked as relevant in the ground truth and shall be selected for inclusion in the summary.

scene boundary (B, state 4) A sample representing a scene boundary (i.e. the first sample of a scene).

non-relevant ($\mathbf{N_{post}}$, state 5) A sample of the video that does not lie within any of the segments marked as relevant in the ground truth and is after a possible selected segment from this scene.

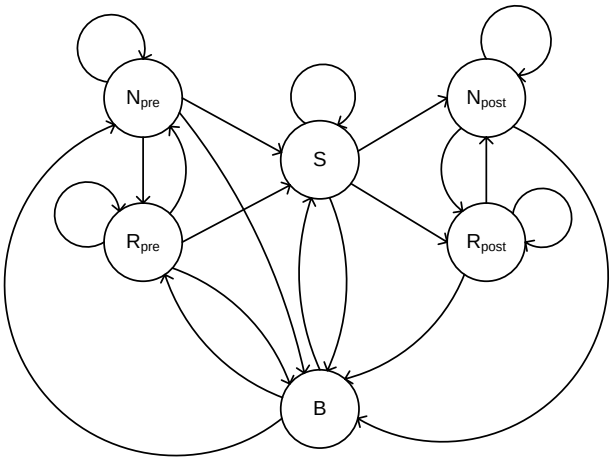

Fig. 1. Topology of the 6 state HMM.

relevant ($\mathbf{R_{post}}$, state 6) A sample of the video that lies within one of the segments marked as relevant in the ground truth, but shall not be selected for inclusion in the summary, and is after a possible selected segment from this scene.

Figure 1 shows a graphical representation of this model. There are four groups of states: states that can be taken before a selected segment in the scene has been encountered ($\mathrm{N_{pre}}$ and $\mathrm{R_{pre}}$), the selected state (S), states that can be taken after a selected segment in the scene has been encountered ($\mathrm{N_{post}}$ and $\mathrm{R_{post}}$) and the scene boundary (B). In general, these groups of states can only be traversed in this order. The exceptions are the direct transition between scene boundary and selected content (as selected content may be directly at the begin or end of a scene) and the "short cuts" from the pre-selection states to the scene boundary (in case that no segment is selected from this scene). Note that this topology is not simply a design decision: Based on the structure of the training data the trained transition matrix of an ergodic HMM will also have zero entries for the transitions missing in this topology.

The proposed method is applied in the content selection step of a video summarization system. In our experiments the feature vectors f_i for the samples consist of relevance values derived from visual features (visual activity, presence of faces) and redundancy values (color bars, other unusable content, redundancy information based on repeated takes). The video sequences are typically subsampled, e.g. by factor 10. The values in the feature vectors are either binary or floating point values in the interval $[0, 1]$. The Viterbi algorithm is applied to estimate the most likely sequence of hidden states for the given observation sequence. From the output of the HMM a list of selected segments is created for the subsequences of samples for which $q_i = \mathrm{S}$. The relevance of such a segment is given as the likelihood of the observation sequence for the segment given the estimated state sequence.

4. EVALUATION

For evaluation we use the TRECVID 2007 rushes data set and the annotations provided by NHK [19] which contains lists of relevant segments and an identification to which scene each segment belongs. We have performed two experiments: comparison of the 3 and 6 state model and application of the approach to content from different productions.

HMM	Precision	Recall	Scene p.	Scene r.	Fraction
3 state	0.3641	0.1371	0.5429	0.3220	0.1596
6 state	0.3410	0.2528	0.5405	0.4851	0.2815

Table 1. Comparison of the 3 state and 6 state HMMs. The table lists sample-based precision and recall, scene-based precision and scene recall and the fraction of the duration of input content that has been selected.

4.1. Comparison of 3 and 6 state model

For the comparison we have performed leave-one-out cross validation on the complete TRECVID 2007 test data set (about 20 hours). We use the following measures: precision (correctly selected samples over all samples reported as selected), recall (correctly selected samples over all selected samples in the ground truth), scene precision (scenes for which segments have been correctly reported over number of reported segments), scene recall (scenes for which segments have been correctly reported over number of scenes containing selected segments in the ground truth) and the fraction of content that has been selected. For the scene precision and recall a scene is assumed to be covered if one segment of this scene has been selected.

Table 1 shows the results of this experiment. The 6 state model yields a slightly lower precision value, but significantly higher recall values (sample-based recall nearly doubles, scene recall increases by 50%). Besides the better distribution of content over the scenes this is caused by the higher fraction of content being included in the summary. Note that despite this increase the fraction of selected content is still below the fraction of content annotated as relevant in the ground truth (0.3715).

4.2. Generalization

One important question is how well the proposed approach generalizes when applied to a different content set. We can again use the TRECVID 2007 rushes data set for this experiment, as it is not a homogeneous data set but composed of videos from different TV productions, including genres as different as e.g. a children's program with hand puppets, a hospital series, a mystery series and a documentary about ancient Greece. We selected the five productions ("The House of Eliott" (HE), "Casualty" (Ca), "Jonathan Creek" (JC), "Ancient Greece" (AG) and "Between the Lines" (BL)) for which at least four videos are available in order to have training sets of at least two hours. We performed experiments by training the 6 state model on the content set of each of the productions and applying the model to each of the four other productions and by training on all videos of the data set except one production and testing on the videos of the productions not used in training.

Table 2 summarizes the results of these experiments. It can be seen that the results depend mainly on the training set and not on the pairwise relation of the content sets (e.g. training on JC and testing on Ca yields quite bad results, while training on Ca and testing on JC performs much better). The performance of the model is influenced by the diversity and inhomogeneity of the training data. The videos from some of the productions are more homogeneous (e.g. containing mainly indoor scenes) so that good results are reached for other productions with the same type of content. This assumption is supported by the fact that the model trained on all but one productions

Test / Training	HE prec.	HE rec.	Ca prec.	Ca rec.	JC prec.	JC rec.	AG prec.	AG rec.	BL prec.	BL rec.
HE	n/a	n/a	0.2224	0.3338	0.3008	0.3156	0.2883	0.3694	0.4292	0.4436
Ca	0.1811	0.0266	n/a	n/a	0.3968	0.1863	0.1202	0.0400	0.3889	0.1328
JC	0.1346	0.0002	0.0862	0.0001	n/a	n/a	0.1776	0.0002	0.1765	0.0002
AG	0.3203	0.1636	0.2327	0.2701	0.2221	0.1359	n/a	n/a	0.4098	0.2885
BL	0.3950	0.0395	0.0529	0.0122	0.1164	0.0144	0.3058	0.1118	n/a	n/a
all except test	0.3485	0.1342	0.2355	0.2814	0.2578	0.1740	0.2226	0.0035	0.3679	0.2836

Table 2. Generalization of the approach: the 6 state model has been trained on the production stated in the first column, sample-based precision and recall of applying this model to the other productions are listed. The last line shows the results for training on all videos of the content set except for those being part of the production used for testing.

yields generally good results, which are comparable to those of the leave-one-out cross validation presented in Section 4.1.

5. CONCLUSION

We have presented a HMM based approach for selecting segments to be included in summaries of raw video content. It has been shown that the 6 state model taking scenes into account provides better recall and better representation of the different scenes, but produces longer summaries. While we can statistically control the expected length and number of selected segments via the annotations of the training data we cannot ensure a certain minimum or maximum duration of the selected segments for a particular video. Also it is difficult to eliminate certain types of unwanted content (e.g. color bars) if they appear as short segments between relevant content. We have also shown that the approach generalizes well across different content sets if the training set is sufficiently diverse. An advantage of the DHMM approach is that online training, e.g. from user interactions in a browsing tool, is possible.

6. REFERENCES

[1] Werner Bailer, Felix Lee, and Georg Thallinger, "Detecting and clustering multiple takes of one scene," in *MMM*, Kyoto, JP, Jan. 2008, pp. 80–89.

[2] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006, pp. 321–330.

[3] Stanley Boykin and Andrew Merlino, "Machine learning of event segmentation for news on demand," *Commun. ACM*, vol. 43, no. 2, pp. 35–41, 2000.

[4] Chung-Lin Huang and Chih-Yu Chang, "Video summarization using hidden Markov model," in *Proceedings of International Conference on Information Technology: Coding and Computing*, Apr. 2001, pp. 473–477.

[5] Reede Ren and Joemon M. Jose, "Football video segmentation based on video production strategy," in *ECIR*, 2005, pp. 433–446.

[6] Milind R. Naphade and Thomas S. Huang, "Discovering recurrent events in video using unsupervised methods," in *Proceedings of IEEE International Conference on Image Processing*, 2002, vol. 2, pp. 1522–4880.

[7] C. Lu, M.S. Drew, and J. Au, "An automatic video classification system based on a combination of HMM and video summarization," *International Journal of Smart Engineering System Design*, vol. 5, pp. 33–45(13), Jan. 2003.

[8] Tiecheng Liu and John R. Kender, "A Hidden Markov Model Approach to the Structure of Documentaries," in *Proc. of the IEEE Workshop on Content-based Access of Image and Video Libraries*, Washington, DC, USA, 2000, p. 111.

[9] Darin Brezeale and Diane J. Cook, "Learning video preferences from video content," in *Proc. of 8th International Workshop on Multimedia Data Mining*, 2007, pp. 1–9.

[10] Gu Xu, Yu-Fei Ma, Hong-Jiang Zhang, and Shi-Qiang Yang, "An HMM-based framework for video semantic analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 11, pp. 1422–1433, 2005.

[11] W. Wolf, "Hidden Markov Model Parsing of Video Programs," in *Proc. of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Washington, DC, USA, 1997, p. 2609.

[12] Yagiz Yasaroglu and A. Aydin Alatan, "Summarizing video: Content, features, and HMM topologies," in *Proc. of 8th International Workshop on Visual Content Processing and Representation*, Madrid, ES, Sept. 2003, pp. 101–110.

[13] Sergio Benini, Pierangelo Migliorati, and Riccardo Leonardi, "Hidden Markov Models for Video Skim Generation," in *Eighth International Workshop on Image Analysis for Multimedia Interactive Services*, Jun. 2007.

[14] YoungSik Choi and KiJoo Kim, "Video summarization using fuzzy one-class support vector machine," in *Computational Science and Its Applications*, 2004, pp. 49–56.

[15] Tanveer Syeda-Mahmood and Dulce Ponceleon, "Learning video browsing behavior and its application in the generation of video previews," in *Proc. of the 9th ACM International Conference on Multimedia*, 2001, pp. 119–128.

[16] Bin Yu, Wei-Ying Ma, Klara Nahrstedt, and Hong-Jiang Zhang, "Video summarization based on user log enhanced link analysis," in *Proc. of the 11th ACM International Conference on Multimedia*. 2003, pp. 382–391, ACM.

[17] Jing Jiang and Chengxiang Zhai, "Extraction of coherent relevant passages using hidden Markov models," *ACM Trans. Inf. Syst.*, vol. 24, no. 3, pp. 295–319, 2006.

[18] Werner Bailer, "A comparison of distance measures for clustering video sequences," in *1st Workshop on Automated Information Extraction in Media Production*, Turin, IT, Sept. 2008.

[19] NHK Science & Technical Research Laboratories, "Test modules for TRECVID activity. Use case scenario. Ver.1.2.0E," Apr. 2008.