CONTENT-BASED VIDEO INDEXING OF TV BROADCAST NEWS USING HIDDEN MARKOV MODELS

Stefan Eickeler, Stefan Müller

Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg
Duisburg, Germany
e-mail: {eickeler,stm}@fb9-ti.uni-duisburg.de

ABSTRACT

This paper presents a new approach to content-based video indexing using Hidden Markov Models (HMMs). In this approach one feature vector is calculated for each image of the video sequence. These feature vectors are modeled and classified using HMMs. This approach has many advantages compared to other video indexing approaches. The system has automatic learning capabilities. It is trained by presenting manually indexed video sequences. To improve the system we use a video model, that allows the classification of complex video sequences. The presented approach works three times faster than real-time. We tested our system on TV broadcast news. The rate of 97.3 % correctly classified frames shows the efficiency of our system.

1. INTRODUCTION

The increasing amount of digital video in multimedia databases results in a demand for techniques for automatic content-based access to video data. In the last years there have been many different approaches to content-based video indexing. A rough categorization of the approaches yields two main classes. The first class of approaches mainly addresses the problem of reusing video sequences in TV studios. In these approaches [1, 2] the video sequence is segmented into shots and a key-frame is extracted for each shot. The key-frame is stored in a library with a reference to the video tape containing the sequence. To get a vision-based access to the scenes the user presents a query image and the retrieval system searches similar key-frames depending on predefined features (e.g. texture and histogram). These approaches extend the technique of image database retrieval to the retrieval of image sequences.

The purpose of the second class of approaches is information retrieval. The video sequences are not addressed by the content of their images, but by their meaning. A possible query may be: "Find the main articles of last month' broadcast news." The user gets a list of news articles and is able to view the video sequences stored in the database.

Approaches of the second class are presented in [3, 4]. They are based on a two stage scene classification scheme. The first stage is the parsing of the video stream, in order to extract each scene. The second stage is indexing, which assigns the scene to one of several competing classes. The main problem of these methods is that the scene extraction and scene classification are separated steps. If an error in the scene extraction occurs, the correct classification of the extracted scene will be impossible. This is a very important point, because not all scenes are separated by hard-cuts, which are easily detectable, but some are bound by edit effects, which are manifold (especially in TV news) and thus are very hard to detect. Our indexing approach based on HMMs works as follows: Each content class is represented by a HMM. A feature vector is derived for each frame of the video sequence and a probabilistic decoding procedure calculates the sequence of HMMs, that maximizes the probability of having generated this feature vector sequence. This procedure seems to be superior compared to this two stage scheme, mainly because parsing and indexing is performed in one stage. Furthermore HMMs are capable of combining the scene models to a complete video model of the TV news and assigning each frame of the news to a certain state of the news model. The automatic learning capabilities of HMMs can be effectively used to process a very large amount of video data automatically and to analyze the characteristics of video contents in a self-organizing way. The first video indexing systems based on HMMs are [5, 6].

2. CONTENT CLASSES AND VIDEO MODEL OF TV BROADCAST NEWS

To index TV broadcast news it is necessary to define useful content classes of TV news. We defined six main content classes:

- NEWSCASTER: The appearance for this class is very similar among most of the TV stations. On the right hand side of the screen (Fig. 1) the newscaster is reading the news and on the left hand side the "news window" displays the headline and an image related to the news topic.
- BEGIN: Introduction sequence of the newscast
- END: The newscasts ending sequence.
- INTERVIEW: An interview of the newscaster and the interviewed person.
- WEATHER FORECAST: This class contains a weather map and animated or static symbols for the various types of weather conditions.
- REPORT: A single shot of a news report. This class includes all scenes that do not belong to any other defined class.

Four classes are defined for the edit effects:

- CUT: Hard-cut at scene and shot boundaries. This is the mostly used edit effect.
- DISSOLVE: The dissolve appears only within a report.
- WIPE: The wipe is the delimiter for two consecutive news reports.

Figure 1: Frame of the class NEWSCASTER

- WINDOW CHANGE: Change (dissolve or wipe) of the "news window" next to the newscaster. This effect is used as separator between two news topics.

To include a-priori knowledge about the broadcast news into the indexing system, we defined a video model for the TV broadcast news. The video model contains the previously defined content classes and determines the possible transitions between them. In Fig. 2 the video model for the news is presented. All possible transitions are indicated by arcs.

3. FEATURE EXTRACTION

This section describes the calculation of the features used as input for the HMMs. One feature vector is calculated for each frame of the input sequence.

In our approach, the most important features for video indexing are based on the difference image. They specify mainly the motion of the main object in the scene. The difference image $d(x, y, t)$ is derived from the original video sequence by absolute difference of the luminance values $I_Y(x, y, t)$ of adjacent frames. One frame of this difference image sequence contains only the changing parts of the frame compared to the previous frame. It should be noted, that this difference image does not represent the motion in the image directly. However, the size of the gray values $d(x, y, t)$ in one frame t of the difference image sequence can be still considered as some indication of the intensity of the motion in each position (x, y) of the image.

A suitable feature for the characterization of the motion distribution is the center of gravity $\vec{m}(t)^T = [m_x x(t), m_x y(t)]$ according to the equations:

$$m_x(t) = \frac{\sum_{x,y} x \cdot d(x,y,t)}{\sum_{x,y} d(x,y,t)} \quad m_y(t) = \frac{\sum_{x,y} y \cdot d(x,y,t)}{\sum_{x,y} d(x,y,t)} \tag{1}$$

The vector $\vec{m}(t)$ can also be interpreted as "center of motion" of the image.

To increase the ability of the HMMs to model the movements of the center, we include the delta features $\Delta\vec{m}(t)$ of $\vec{m}(t)$ for the "center of motion" into the feature vector. The delta features are defined as:

$$\Delta m_x(t) = m_x(t) - m_x(t-1), \quad \Delta m_y(t) = m_y(t) - m_y(t-1) \tag{2}$$

Another useful feature is the average absolute deviation of the motion in all points of the image from the center of motion $\vec{\sigma}(t)^T = [\sigma_x(t), \sigma_y(t)]$, defined as:

$$\sigma_x(t) = \frac{\sum_{x,y} d(x,y)|(x - m_x(t))|}{\sum_{x,y} d(x,y)}$$
$$\sigma_y(t) = \frac{\sum_{x,y} d(x,y,t)|(y - m_y(t))|}{\sum_{x,y} d(x,y,t)} \tag{3}$$

This feature is very similar to the second translation invariant moment of the distribution, but it is more robust against noise in the image sequence. It can also be considered as "wideness of the movement".

An important feature is the "intensity of motion", expressed as

$$i(t) = \frac{\sum_{x,y} d(x,y,t)}{XY} \tag{4}$$

This feature does not only describe the characteristics of the motion, [7] shows that this feature is also very useful for the detection of hard-cuts.

The video sequence of a report is sometimes influenced by the flashes of photographers. The frames with flashes result in a high value for Eq. (4) and are recognized as hard-cut in the classification process. To avoid this, we use the smaller value of the motion intensity for the frames $(t, t+1)$ and $(t-1, t+2)$.

$$i'(t) = \text{MIN}\left(i(t), \frac{\sum_{x,y} |I(x,y,t-1) - I(x,y,t+2)|}{XY}\right) \tag{5}$$

As shown in [8], the difference image features are very suitable for the recognition of complex movements. The features $\vec{m}(t)$, $\vec{\sigma}(t)$, and $i(t)$, and a feature similar to $\Delta\vec{m}(t)$ are used to recognize 24 different gestures with high accuracy (91.67%). HMMs have been also used in this case for classification. These investigations have been serving as very insightful experiments for us in order to study the possibilities for stochastic modeling of video sequences, and finally led to the basic principle of our approach to video-indexing presented in this paper.

As reported in [7], the intensity of the difference histogram is a useful feature for cut detection. It increases the confidence of cut detection significantly. The intensity of the difference histogram is based on the gray-level (luminance) histogram $h(t)$ and is defined as:

$$hi(t) = \sum_g |h_g(t) - h_g(t+1)| \tag{6}$$

with $h_g(t)$ defined as the bin for the gray-level (luminance) g in frame t. A moving object (news speaker) in front of a uniformly colored background has a small influence on the histogram and results in a low value for the difference histogram feature, while edit effects produce high values.

To reduce the effect of flashes of photographers we use, as in Eq. (5) the smaller value of these features for the frames $(t, t+1)$ and $(t-1, t+2)$.

$$hi'(t) = \text{MIN}\left(hi(t), \sum_g |h_g(t-1) - h_g(t+2)|\right) \tag{7}$$

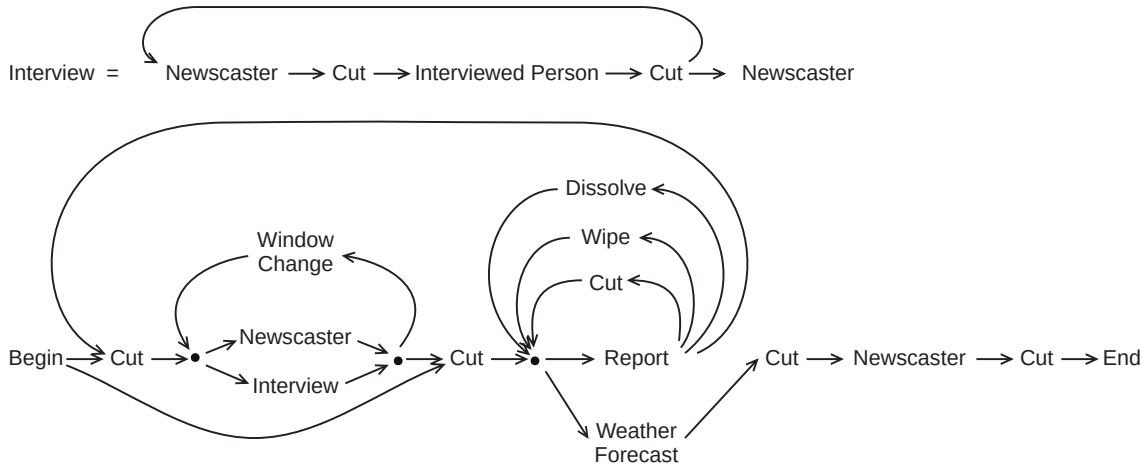

Figure 2: Video model of TV broadcast news

The detection rate of cuts can be further improved by using the equation:

$$hi''(t) = hi(t) - \text{MEDIAN}\,(hi(t-1), hi(t), hi(t+1)) \quad (8)$$

Eq. (8) is based on the median filtering, which is useful to remove impulsive noise from images. Here we use the difference between the filtered value and the original value as impulse detector.

The use of static information results in an improved classification result. We average the Y, U an V color components over the frame t and add them to the feature vector.

$$Y(t) = \frac{\sum\limits_{x,y} I_Y(x,y,t)}{XY} \quad (9)$$

A special feature is designed to detect dissolve edit effects. During a dissolve the value of a pixel should be similar to the interpolated value of the adjacent pixel values $(t-1, t+1)$. This condition is true for static shots, too. Therefore the ratio of the difference pixel to the difference from the interpolated pixel is used.

$$d(t) = \sum_{x,y} \frac{I_Y(x,y,t) - I_Y(x,y,t-1)}{I_Y(x,y,t) - \frac{1}{2}(I_Y(x,y,t-1) + I_Y(x,y,t+1))} \quad (10)$$

To each image of the video are 1280 samples of the audio signal assigned. The logarithmic energy is used to improve the segmentation of the scenes:

$$E(t) = ln \sum_{s} S^2(s) \quad (11)$$

Additional to the energy, a 12 dimensional cepstral vector is calculated for the samples of each frame which enhances the segment classification.

The result of these feature extraction methods is a feature vector consisting of the 25 features.

4. STATISTICAL CLASSIFICATION

Hidden Markov Models are one of the most efficient tools for modeling of time-varying pattern sequences. They have been used very successfully in speech [9] and online handwriting recognition. In our approach we use Hidden Markov Models to build stochastic models of the various content classes of the news. One HMM is trained for each of the previously defined content classes by presenting sample data using the Baum-Welch algorithm.

We use first order left-to-right Hidden Markov Models with 1 to 15 states and continuous output distribution to model the content classes. The video model of the broadcast news (Fig. 2) is converted to a superior HMM, with the HMMs of the content classes as state output. For the classification we evaluate the optimal path resulting from the Viterbi algorithm [9] for assigning the most likely content class to each frame of the video sequence. The video model is used to restrict the optimal path to a sequence of content classes that can only occur as defined in this model. By automatically incorporating this model into the HMM framework and into the decoding procedure, the semantic constraints imposed by the video model can be fully integrated into the recognition procedure. We consider this automatic video model integration as one of the major advantages of our statistical approach to video indexing. It should be also noted that the models for edit effects play a major role in this framework, because they serve as important nodes in the model displayed in Fig. 2. The upper part of Fig. 2 shows also an insightful example for the use of nested structures in our approach: The content class INTERVIEW is composed of several other HMMs. In order to detect an interview, the optimal path has to go through all these classes.

5. EXPERIMENTS AND RESULTS

For experiments we recorded TV news of different TV stations. The TV news were captured from a VHS video tape as YUV 4:2:2 images in a resolution of 192×144 pixels at a rate of 12.5 frames per second and the audio signal with a sampling rate of 16 kHz. To obtain meaningful results we carried out a TV station dependent and a TV station independent indexing experiment of TV broadcast news.

For the station dependent test of the system we recorded ten TV news of the same TV station. The length of the news range from 5 to 15 minutes. To get a maximum amount of training and test data for this experiment, we used the hold-out method. This means that nine news were used for the training of the system and

segment boundary			segment	
insertion	deletion	accuracy	classif.	
10.6%	1.0%	0.3 frms	92.5%	
6.5%	0.0%	0.0 frms	96.9%	
1.0%	6.0%	0.4 frms	95.2%	
2.4%	6.9%	0.2 frms	98.5%	
0.9%	1.9%	0.1 frms	98.2%	TV station
0.0%	0.0%	0.9 frms	97.2%	dependent
1.8%	9.3%	0.2 frms	96.7%	
3.2%	3.2%	0.1 frms	97.0%	
3.4%	2.6%	0.2 frms	99.2%	
0.0%	0.0%	0.0 frms	100.0%	
0.0%	2.2%	0.1 frms	97.9%	
0.0%	2.3%	0.8 frms	95.6%	TV station
0.0%	18.8%	0.9 frms	90.7%	independent
5.1%	9.8%	0.4 frms	90.9%	

Table 1: Recognition Results for the different TV news

content class	correct detections	false detections
Newscaster	98.40%	6.14%
Report	96.40%	1.88%
Begin	100.00%	0.00%
End	100.00%	0.00%
Weather Forecast	80.00%	0.00%
Interview	92.85%	0.00%
Wipe	54.55%	14.29%
Window Change	100.00%	0.00%
Dissolve	91.67%	26.67%

Table 2: Detection- and insertion-rates for TV station dependent test

the remaining sequence was used for the test. This procedure was repeated for every news video. Finally we calculated the measures defined in [10].

The average recognition rate for all broadcasts and all content classes is 97.3 %. Tab. 1 shows the recognition results for the different broadcasts: insertion and deletion rates for segment boundaries, accuracy of the segment boundary in frames and the rate of correct classified content classes. In Tab. 2 the correct detection and false detection rates for each content class are given.

Most of the false detections for the class NEWSCASTER are caused by sequences of an orator at a parliament or a press conference, who has very similar movements compared to the newscaster. The weather forecast has lower recognition rates than the other main content classes. This shows, that the used features are not yet suitable for the recognition of this class and other features have to be included into this system. Texture features should be suitable to improve the modeling of the weather forecast.

For a station independent experiment we recorded four additional news broadcasts of another TV stations. These news are of a similar style. For this experiment the recognition system was trained on the news of the station dependent test and was tested on the additional news. Tab. 1 shows the recognition rates for these news broadcasts. The average segment classification rate is 93.7 %.

The video indexing system works three times faster than real-time. The indexing of a 15 minutes TV news has a computing time of 5 minutes on a Pentium-based PC. We consider this as a significant advantage over other approaches, which are often very time consuming.

6. CONCLUSIONS AND FUTURE WORK

We presented a new approach to content-based video indexing and showed high recognition rates for the task of TV broadcast news indexing. Although we are just at the beginning of our investigation, the indexing results convince us, that HMM-based video indexing is a very promising approach, incorporating the following advantages in comparison to most standard procedures:

- Exploitation of the automatic learning and self-organizing capabilities of HMM.
- Possibility to learn from a large amount of training data.
- Scene classification and scene boundary detection in one process.
- Automatic incorporation of a stochastic video model.
- Real-time capability.

In the future we will apply our indexing approach to more complicated tasks with more content classes such as sports classification.

7. REFERENCES

[1] E. Ardizzone, M. L. Cascia, and D. Molinelli. Motion and Color Based Video Indexing. In *ICPR*, Vienna, Austria, Aug. 1996.

[2] R. W. Picard. A society of models for video and image libraries. Perceptual Computing Section Technical Report 360, MIT Media Lab, Cambridge, MA, 1996.

[3] A. G. Hauptmann, M. J. Witbrock, and M. G. Christel. News-on-demand: An application of informedia technology, Sept. 1995.

[4] S. W. Smoliar and H. Zhang. Content-Based Video Indexing and Retrieval. *IEEE Multimedia*, 1(2):62–72, 1994.

[5] S. Eickeler, A. Kosmala, and G. Rigoll. A New Approach to Content-Based Video Indexing Using Hidden Markov Models. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 149–154, Louvain-la-Neuve, Belgium, June 1997.

[6] J. S. Boreczky and L. D. Wilcox. A Hidden Markov Model Framework for Video Segmentation Using Audio an Image Features. In *Proc. IEEE ICASSP*, Seattle, May 1998.

[7] A. Hampapur, R. Jain, and T. Weymouth. Digital Video Indexing in Multimedia Systems. In *Proc. of the Workshop on Indexing and Reuse in Multimedia Systems*. AAAI, Aug. 1994.

[8] G. Rigoll and A. Kosmala. New improved Feature Extraction Methods for Real-Time High Performance Image Sequence Recognition. In *Proc. IEEE ICASSP*, pages 2901–2904, Munich, Apr. 1997.

[9] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, 77(2):257–285, 1989.

[10] S. Eickeler and G. Rigoll. Measurements for the Evaluation of Video Indexing Systems. Technical report, Faculty of Electrical Engineering - Computer Science, Gerhard-Mercator-University Duisburg, 1998. `http://www.fb9-ti.uni-duisburg.de/report.html`.