Automatic Shot Boundary Detection and Classification of Indoor and Outdoor Scenes

A. Miene, Th. Hermes, G. Ioannidis, R. Fathi, and O. Herzog

TZI - Center for Computing Technologies
University of Bremen
Universitätsallee 21-23
D-28359 Bremen, Germany
{andrea|hermes|gtis|fathi|herzog}@tzi.de

Abstract. This paper describes our contribution to the TREC 2002 video analysis track. We participated in the shot detection task and in the feature extraction task (features indoors and outdoors).
The shot detection approach is based on histogram differences and uses adaptive thresholds. Multiple detected shot boundaries that follow each other within a short temporal interval are grouped together and classified as a gradual change beginning with the first and ending with the last shot boundary in the interval.
For the feature extraction task we examined whether it is possible to classify indoor and outdoor shots by their color distribution. In order to analyze the color distribution we use first order statistical features. The shots are classified into indoor and outdoor shots using a neural net.

1 Introduction

The Center for Computing Technologies (TZI), University of Bremen, Germany, participated in the video analysis track in the shot detection task and in the feature extraction task (features indoors and outdoors).

The shot detection approach is based on histogram differences. It is divided into two steps - feature extraction and shot boundary detection. Firstly, the histogram differences are calculated for the entire video in real time. Secondly, shot boundaries are detected. The advantage of this approach is the possibility to set adaptive thresholds for the shot boundary detection considering all extracted features of the complete video sequence. The adaptive threshold is set to a percentage of the maximum of all calculated difference values of the video. In the case of gradual changes, often multiple shot boundaries are detected. Therefore multiple detected shot boundaries that follow each other within a short temporal interval are grouped together and a gradual change is detected beginning with the first and ending with the last shot boundary in the interval. The approach is explained in more detail in section 2.

To extract the features indoors and outdoors we use a feed forward neural net as a classificator, trained by a backpropagation learning rule. The input is a feature vector describing the color distribution of an image. The output is the

probability for each feature (indoors and outdoors) to appear in the image. The approach is discussed in section 3.

2 Shot detection

Quite a lot of approaches to shot boundary detection were proposed in the literature. An overview is given in [Lienhart, 1999,Yusoff et al., 1998]. The principle methodology of shot boundary detection is to extract one or more features from every nth frame of a video sequence, to compute the difference of features for consecutive frames, and to compare these differences to a given threshold. Each time the threshold is exceeded a shot boundary is detected. The various approaches differ concerning the used features.

The shot boundary detection system we used for TREC 2002 is based on the approach presented in [Miene et al., 2001]. As mentioned before , the approach can be divided into two main parts. The first part is to extract all needed features from a video. The second part is to detect the shot boundaries based on the previously extracted features.

For the feature extraction part each frame is converted into a grayscale image. Then a histogram H_G is created. Subsequently, the squared differences between each two consecutive frames

$$H_{G_{Diff}}(n, n-1) = \sum_{i=0}^{255} \frac{(H_G(n)(i) - H_G(n-1)(i))^2}{Max(H_G(n)(i), H_G(n-1)(i))} \tag{1}$$

are calculated. $H_G(n)(i)$ denotes a grayscale histogram value at index i of frame n. $Max(H_G(n)(i), H_G(n-1)(i))$ denotes the maximum of both grayscale histogram values $H_{Gray}(n)(i)$ and $H_{Gray}(n-1)(i)$, and is used as a normalization factor.

This leads to a feature difference list in order to detect shot boundaries, which is compared to a threshold. To determine the adaptive threshold, the maximum of all calculated difference values of the actual video is calculated. The adaptive threshold for the actual video is specified as a percental value of the maximum:

$$Th = \frac{Max\{H_{G_{Diff}}(1,0), \ldots, H_{G_{Diff}}(n, n-1)\} \cdot Th_{percentage}}{100} \tag{2}$$

For gradual changes like dissolves or wipes the shot boundary detection often detects more than one boundary per shot. Therefore, all shot boundaries which belong to the same shot have to be merged into one boundary. This step is illustrated in Figure 1. Shot boundaries are merged together if the temporal distance between their occurrences is less than a threshold. The minimal frame number of the merged shot boundaries determines the start, and the maximum frame number determines the end of the gradual change. The exact boundary position is set to the maximum feature difference value within the merged shot boundaries.

Before preparing our results for TREC we tested our shot detection on three videos from the feature developement collection, for which we determined the

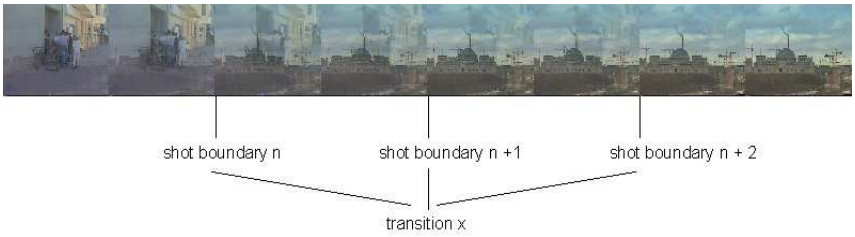

Fig. 1. Merging of multiple detected shot boundaries[Miene et al., 2001].

shot boundaries manually. The results of this experiment are shown in table 2. *File* denotes the file name of the video, *human* the amount shot boundaries determined manually, *auto* the total number of shot boundaries detected by the system, *correct* the amount of correct detected shot boundaries, *false* the amount of false alarms and *missing* the amount of shot boundaries, that were not detected by the system. The last two colums contain the percentage values for precission and recall. For this test we did not distinguish between hard cuts and gradual changes.

file	human	auto	correct	false	missing	recall	precission
00028a.mpg	116	79	79	0	37	68.1	100
00435.mpg	108	54	47	7	61	42.6	85.2
00535.mpg	120	108	100	8	20	83.3	92.6
over all	344	241	226	15	118	65.7	93.8

In the following section we present our approach for the classification of indoor and outdoor scenes.

3 Classification of indoor and outdoor scenes

For the feature extraction task we have examined whether it is possible to classify indoor and outdoor shots by their color distribution. In order to analyze the color distribution, first order statistical features are used, which are extracted from the histograms of the three color channels (RGB) and the grey level histogram. The features calculated from each histogram are average, variance, and amount of peaks, normalized to an interval $[0.0, \ldots, 1.0]$. Therefore we calculate 12 statistical features alltogether. In order to classify the shots into indoor and outdoor shots, a feed forward neural net with backpropagation learning was trained. For this task we used the SNNS (Stuttgart Neural Network Simulator) [SNNSv4.2, 2002].

At the input layer the 12 statistical features are presented, that were obtained from the histograms. The output layer consists of two neurons that take on

values between 0.0 and 1.0 measuring the probability for the features indoors or outdoors in the shot. Two hidden layers with 20 neurons each are initialized with random weights. In order to train the neural net, some videos from the feature development collection were chosen. The shots are classified manually to generate 323 training data sets, 178 for indoors, and 145 for outdoors. Figure 2 shows the trained neural net.

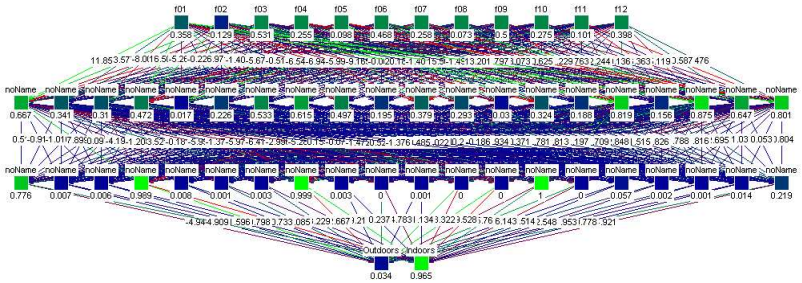

Fig. 2. Trained neural net.

In order to classify the shots from the feature extraction test collection, a set of n key frames is extracted from each shot. Every kth frame of a shot is used as a key frame, but in order to be more independent of inaccuracies during the shot detection and of gradual changes (e.g., wipes, fades, or dissolves) a number of frames around the shot boundaries is skipped (see Figure 3). In order to classify a shot, the set of n key frames is presented to the neural net.

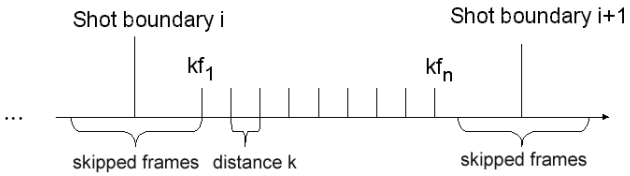

Fig. 3. Extraction of key frames.

For each of the two output neurons a list is obtained containing n values, one for each key frame. The median is calculated for each list to obtain the final indoors or outdoors probabilities for the shot. In order to measure the accuracy of the classification result, the difference between the median values of the indoors and the outdoors neuron is calculated. If the difference exceeds a threshold the shot is classified to contain the feature with the higher probability. The difference is also used for the ranking.

4 Future Work

For the next months and also for our next participation in TREC 2003 we will concentrate on the improvement of our shot detection approach concerning the detection of gradual changes. We have also to examine how to obtain better results for the extraction of the indoors and outdoors features. As mentioned before the neural net was trained with indoor and outdoor example frames from the feature development collection. The results from the TREC evaluation of our results shows that there are serious problems with the classification of the material from the feature test collection. At the moment we are working on the analysis of these problems. One major problem appears to be, that we trained the net only with examples of indoor and outdoor scenes. Therefore the results for scenes containing neither indoor nor outdoor scenes are undefined. Therefore, especially artificial scenes lead to a wrong classification.

In addition we are looking forward to develop further modules for our feature extraction system to be able to extract other features like text or human faces.

References

[Lienhart, 1999] Lienhart, R. (1999). Comparison of Automatic Shot Boundary Detection Algorithms. In *Proc. SPIE Vol. 3656 Storage and Retrieval for Image and Video Databases VII*, pages 290–301, San Jose, CA, USA.

[Miene et al., 2001] Miene, A., Dammeyer, A., Hermes, T., and Herzog, O. (2001). Advanced and adapted shot boundary detection. In Fellner, D. W., Fuhr, N., and Witten, I., editors, *Proc. of ECDL WS Generalized Documents*, pages 39–43.

[SNNSv4.2, 2002] SNNSv4.2 (2002). *SNNS Stuttgart Neuronal Network Simulator User Manual, Version 4.2*. University of Stuttgart and University of Tübingen. http://www-ra.informatik.uni-tuebingen.de/downloads/SNNS/SNNSv4.2.Manual.pdf.

[Yusoff et al., 1998] Yusoff, Y., Christmas, W., and Kittler, J. (1998). A study on automatic shot change detection. *Lecture Notes in Computer Science*, 1425.