UNIVERSITY OF CENTRAL FLORIDA AT TRECVID 2003

Yun Zhai, Zeeshan Rasheed, Mubarak Shah

Computer Vision Laboratory
School of Computer Science
University of Central Florida, Orlando, Florida

ABSTRACT

In this paper, we describe the methods for shot-boundary detection, story segmentation, and feature extraction developed at the Computer Vision Laboratory, UCF, for TRECVID 2003 forum. To detect shot boundaries in videos, we use a multi-resolution color histogram intersection technique in a coarse-to-fine fashion. Detected shot transitions are further classified into abrupt and gradual transitions. We have also developed a method to segment video clips into stories. Our method is based on only visual cues and is useful to separate commercials from the news stories. In addition, we have contributed to the feature extraction task and have developed methods to detect two features, namely, Non-Studio Settings and Weather News. The feature detection also relies on the visual information and exploits the structure of the shot transitions and the common annotations in the news videos. We present our results and its evaluation released by NIST.

1. INTRODUCTION

The increasing amount of video data available to us poses challenges to develop tools for video indexing and searching, so that users can efficiently navigate through it. For this purpose, the National Institute of Standards and Technologies (NIST) has conducted TREC video retrieval evaluation (TRECVID) forum. Together with over thirty teams, the Computer Vision Laboratory, UCF, has also contributed to this task. We have participated in three of four tracks: shot-boundary detection, story segmentation, and feature extraction with features 'Non-Studio Settings' and 'Weather News'. Our shot-boundary detection algorithm uses a coarse-to-fine approach based on multi-resolution color histogram intersection. The detection of shots is further extended to the detection of shot transition types including abrupt and gradual transitions. For story segmentation, we cluster shots into stories using only visual cues. Each segment is further classified into either news or non-news story. For feature extraction in videos, 17 features are provided by NIST, including outdoor, people, building, non-studio settings, weather news, sport events, and etc. We have developed algorithms to detect two features: non-studio settings and weather news. It is important to note that our methods rely on information embedded in video tracks only and exploit the shot transition patterns generally observed in news videos.

The rest of the paper is organized as follows: Section 2 addresses the method for shot-boundary detection. Section 3 addresses the method for story segmentation and classification. The detection of 'non-studio settings' and 'weather news' features are discussed in Sections 4.1 and 4.2 respectively. Section 5 concludes our work.

2. SHOT-BOUNDARY DETECTION

We use a coarse-to-fine approach based on color histogram intersection of frames. In the first step, the approximate location of shot boundary is detected at a lower frame rate. This is followed by transition position localization at a higher frame

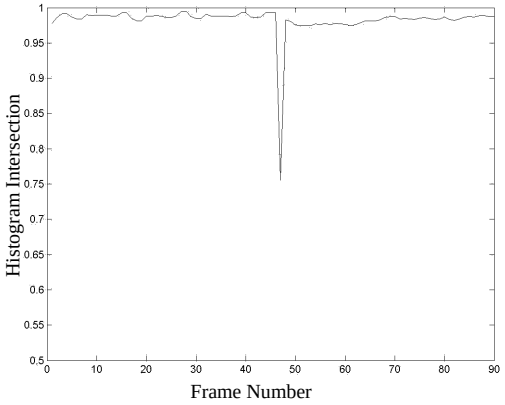

Fig. 1. Histogram intersection plot for a short video containing a shot with an abrupt transition.

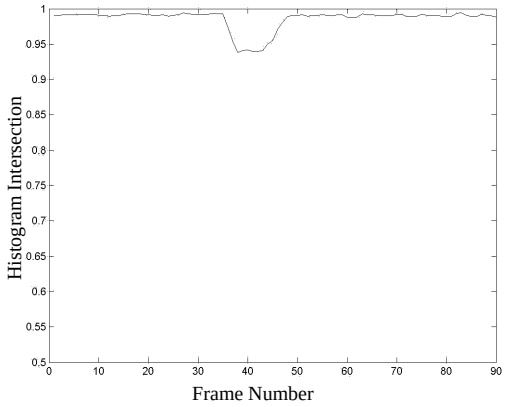

Fig. 2. Histogram intersection plot for a short video containing a shot with a gradual transition.

rate.

2.1. Shot Boundary Initialization

A shot is defined as a sequence of frames taken by a single camera with no major changes in the visual content. During a shot transition, the visual similarity of consecutive frames changes. This can be detected by observing the color histograms of the frames. We use a 24-bin histogram in RGB color space, allocating 8 bins for each channel. Let $D(i)$ represents the histogram intersection between frames f^{i-1} and f^i, then:

$$D(i) = \sum_{allbin\ b} min(H_{i-1}(b), H_i(b)) \qquad (1)$$

where H_{i-1} and H_i are histograms of frames f^{i-1} and f^i respectively. A shot boundary at f^i is found if:

$$\begin{aligned} D(i-1) - D(i) &> T_{color} \\ D(i+1) - D(i) &> T_{color}, \end{aligned} \qquad (2)$$

where, T_{color} is a threshold that captures the significant difference between the color statistics of two frames.

For abrupt transitions (Fig. 1), where the difference in color distribution is large enough, the above frame-to-frame histogram intersection performs well. However, for the gradual transitions (See Fig. 2), the color difference between consecutive frames is not significant, the above technique does not reliably capture the shot transitions. To deal with this problem, we temporally sub-sample the original video sequences (every fifth frame in our experiments). The histogram intersection is then applied to the sub-sampled sequences thus amplifying the frame-to-frame visual dissimilarities. (Fig. 3). This initial estimate of shot boundary, which may not be accurate, is refined in the next step.

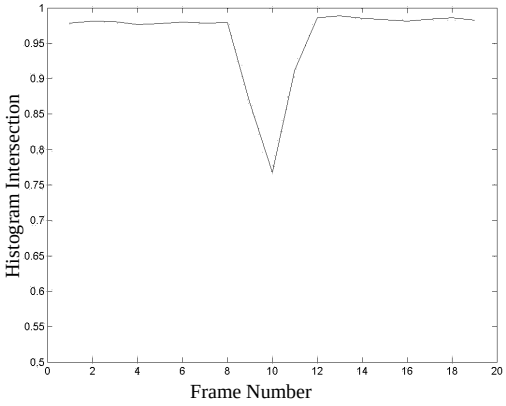

Fig. 3. Histogram intersection plot for a sub-sampled (with Sampling Rate 5fps) video (shown in Fig. 2) containing a shot with a Gradual Transition.

Once the approximate location of a shot transition is obtained, we localize it at the highest sampling rate. This is achieved by detecting the frame with the minimum frame similarity in a window, that is, the local minimum of the color histogram intersection plot. Let P be the initial estimate of shot transition and a be the search range, the lo-

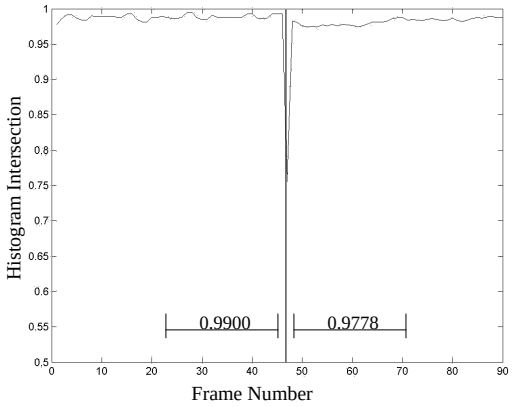

Fig. 4. Histogram Intersection plot for a short video containing a shot with an abrupt transition. The numbers shown on the left and right of the initial boundary are the averages of the histogram intersection over the search windows. The windows size is 30 frames.

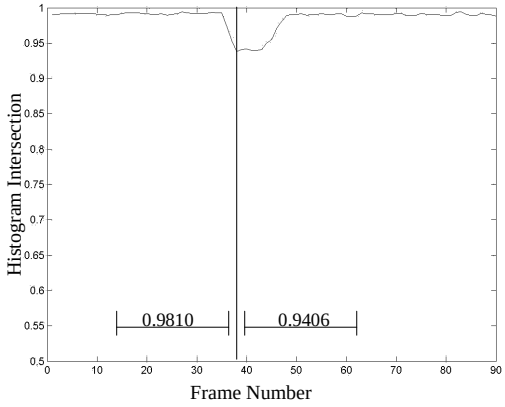

Fig. 5. Histogram Intersection plot for a short video containing a shot with a Gradual Transition. The numbers shown on the left and right of the initial boundary are the averages of the histogram intersection over the search windows. The windows size is 30 frames.

calized transition boundary at frame M is calculated as:

$$M = arg\,min(\{D(P-a), ..., D(P+a)\}). \quad (3)$$

2.2. Illumination Artifact Removal

The detection of the shot transitions is followed by the removal of outliers. We have observed that in the news videos, related to meetings, briefings, celebrities, politicians, the most common outliers occur due to the camera flashes. In such cases, the illumination of consecutive frames abruptly changes and results in a false shot boundary. These illumination artifacts are very short in time and cause frequent over detection. To remove such outliers, the frames in each consecutive candidate shots are compared. We compute the average color histogram distributions, K_L and K_R, of the immediate left and right shots of a candidate shot boundary. The visual similarity of these two neighbors is computed using histogram intersection $D(K_L, K_R)$ of K_L and K_R. If $D(K_L, K_R)$ is high, the transition is considered due to the lighting artifact, and the consecutive shots are merged into a single shot. Otherwise, the transition is considered real.

2.3. Determining Transition Type

TRECVID 2003 proposed to classify shot transitions into two kinds of transitions: gradual and abrupt. Examples of gradual transitions are dissolves, fade-ins, fade-outs, and wipes. In gradual transitions, the temporal distance between starting and ending frames is greater than 1. The length of transition, however, may differ for different types of transitions. In abrupt transitions, the visual changes occur between the ending frame of the previous shot and the starting frame of the following shot. Since we estimate the initial boundary from the sub-sampled domain, the transition could take place before the initial boundary, after the initial boundary, or across the initial boundary (Fig. 5).

To identify the type of transition, we consider frames in a neighborhood of size b, on each side of the detected shot boundary P. The average histogram intersections D_L and D_R of each neighbor are calculated as:

$$D_L = \frac{1}{b}\sum_{i=1}^{b} D(P-i)$$
$$D_R = \frac{1}{b}\sum_{i=1}^{b} D(P+i) \quad (4)$$

If both D_L and D_R are very high, the transition is declared as abrupt (Fig. 4). Otherwise, the transition is classified as gradual. This is shown in

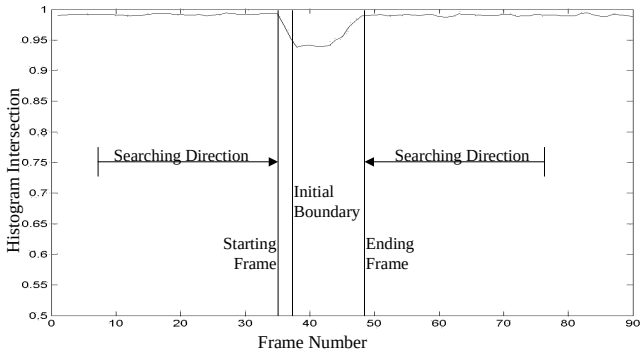

Fig. 8. Locating the starting and ending frames in a gradual transition.

Fig. 5.

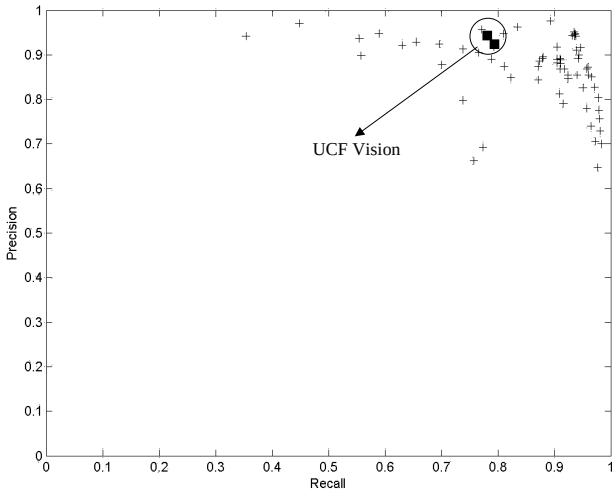

Fig. 9. UCF standing for the abrupt transition detection task. The results are represented by the small squares enclosed in a circle.

2.4. Gradual Transition Boundaries

Once the transition type is determined, the exact transition boundaries needs to be computed next. The determination of boundaries for the abrupt transitions is straightforward, since the transition only takes place in two frames. However, it is important to locate the accurate beginning and ending frames of the gradual transitions as proposed in the task definition.

We assume that the transition length is below some value. If we pick a point that is far away from the initial boundary, and that point is inside the shot instead of the transition, then, the average of the histogram intersection of the neighborhood around that point is high. As we move this point towards the initial boundary, the average of the histogram intersection of the neighborhood around that point will start dropping down at the places where the transition starts and ends. If the point comes from the left side of the initial boundary, it is the starting frame of the transition. If it comes from the right side of the initial boundary, it is the ending frame of the transition. This is illustrated in Fig. 8 .

2.5. Results and Discussions

The data set provided by NIST contains 13 MPEG-1 news videos from CNN, ABC and C-SPAN news networks. Each of these videos is around thirty minutes long. The news videos from CNN and ABC contain both news program and commercials. The videos from C-SPAN only contain news programs.

We have applied our shot-boundary detection method on all 13 news videos. There are three kinds of measurements:

- Precision/Recall for Abrupt Transitions:

$$Precision = \frac{A_{abrupt}}{X_{abrupt}}, \quad Recall = \frac{A_{abrupt}}{Y_{abrupt}} \quad (5)$$

  where A_{abrupt} is the number of matched abrupt transitions, X_{abrupt} is the number of detected abrupt transitions, and Y_{abrupt} is the number of reference abrupt transitions.

- Precision/Recall for Gradual Transitions:

$$Precision = \frac{A_{gradual}}{X_{gradual}}, \quad Recall = \frac{A_{gradual}}{Y_{gradual}} \quad (6)$$

  where $A_{gradual}$ is the number of matched gradual transitions, $X_{gradual}$ is the number of detected gradual transitions, and $Y_{gradual}$ is the number of reference gradual transitions.

- Frame Based Precision/Recall for Gradual Transitions:

$$Precision = \frac{A'_{gradual}}{X'_{gradual}}, \quad Recall = \frac{A'_{gradual}}{Y'_{gradual}} \quad (7)$$

Filename	Type	Cut Recall	Cut Precision	Grad Recall	Grad Precision	Frame Recall	Frame Precision
19990303.121216	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980619_ABC	ABC	0.890	0.934	0.811	0.673	1150/1711	1150/1314
19980224_ABC	ABC	0.820	0.874	0.717	0.648	970/1733	970/1074
19980425_ABC	ABC	0.732	0.896	0.872	0.570	1522/2287	1522/1815
19980222_CNN	CNN	0.737	0.904	0.712	0.365	735/1439	735/1043
19980515_CNN	CNN	0.770	0.923	0.824	0.545	1252/2063	1252/1747
19980531_CNN	CNN	0.757	0.897	0.824	0.494	749/1255	749/937
19980412_ABC	ABC	0.800	0.907	0.861	0.598	1105/2118	1105/1301
2001614.1647460	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980308.1216980	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
20010628.1649460	CSPAN	0.962	0.916	0.000	0.000	0/0	0/0
20010702.1650112	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980203_CNN	CNN	0.732	0.843	0.853	0.626	1478/2221	1478/1697
Mean		2097/2644	2097/2285	887/1090	887/1616	8961/14827	8961/10928
		0.793	0.918	0.814	0.550	0.604	0.820

Fig. 6. Results of run1 for the shot-boundary detection task.

Filename	Type	Cut Recall	Cut Precision	Grad Recall	Grad Precision	Frame Recall	Frame Precision
19990303.121216	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980619_ABC	ABC	0.890	0.940	0.826	0.677	1163/1734	1163/1326
19980224_ABC	ABC	0.820	0.886	0.725	0.641	980/1745	980/1084
19980425_ABC	ABC	0.718	0.898	0.866	0.561	1514/2279	1514/1818
19980222_CNN	CNN	0.728	0.914	0.722	0.361	738/1475	738/1039
19980515_CNN	CNN	0.766	0.923	0.824	0.545	1252/2063	1252/1747
19980531_CNN	CNN	0.740	0.898	0.833	0.491	753/1266	753/964
19980412_ABC	ABC	0.794	0.907	0.861	0.595	1101/2114	1101/1292
2001614.1647460	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980308.1216980	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
20010628.1649460	CSPAN	0.962	0.916	0.000	0.000	0/0	0/0
20010702.1650112	CSPAN	1.000	1.000	0.000	0.000	0/0	0/0
19980203_CNN	CNN	0.675	0.887	0.894	0.607	1544/2320	1544/1774
Mean		2065/2644	2065/2193	898/1090	898/1687	9045/14996	9045/11044
		0.781	0.942	0.824	0.532	0.603	0.819

Fig. 7. Results of run2 for the shot-boundary detection task.

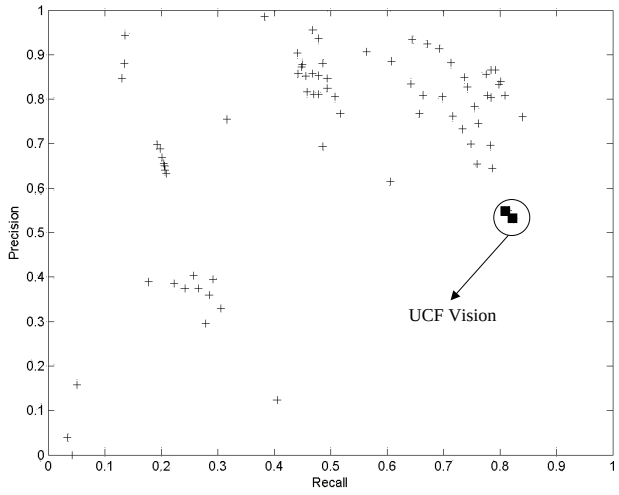

Fig. 10. UCF standing for the gradual transition detection task. The results are represented by the small squares enclosed in a circle.

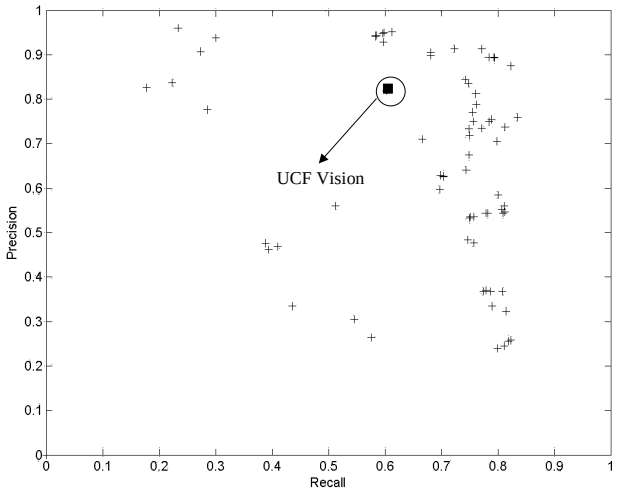

Fig. 11. UCF standing for the abrupt transition detection task (frame-based precison/recall). The results are represented by the small squares enclosed in a circle.

where $A'_{gradual}$ is the total number of frames in the overlapping regions in matched gradual transitions, $X'_{gradual}$ is the total number of frames in the detected gradual transitions, and $Y'_{gradual}$ is the total number of frames in the reference gradual transitions.

For the transitions that are less than 5 frames long, they are compared with either the abrupt transitions or the gradual transitions.

UCF Vision group has submitted two runs for the shot-boundary detection task. The detailed results can be found in Fig. 6 and Fig. 7. Our standings among the participants are shown in Fig. 9, Fig. 10, and Fig. 11. The results demonstrate that using only visual information is sufficient for detecting shot transitions.

3. STORY SEGMENTATION AND CLASSIFICATION

A 'news' story is defined as a segment of a news broadcast with a coherent news focus which contains at least two independent, declarative clauses. Other coherent segments are labelled as 'miscellaneous'. A story can be composed of multiple shots. In the story segmentation task, our method uses only visual cues and is based on the knowledge of the television editing techniques.

3.1. Algorithm Description

It is common to introduce a set of blank frames (blank break) before every story in the television programs. This blank informs the audience that a new story is going to start. It could be inserted into either the audio or the video track of the video. Our algorithm only utilized the visual information:

1. First, we detect all the blank breaks in the video track. These breaks separate the entire video clips into stories. To be visually detectable, the length of the break has to be longer than a certain time. Thus, we remove the short breaks by merging the adjacent stories to prevent over-segmentation.

2. Next, since the news content is the dominant element in the news videos, the length of the news related stories is much longer than the length of other types of stories. In CNN and ABC videos, the news stories usually are longer than 3 minutes, while the other types of stories, like commercials, are less than 2 minutes. Based on this fact, we label the long stories 'news' and all others 'miscellaneous'.

3. Finally, all the adjacent 'miscellaneous' sto-

Fig. 12. Results of story segmentation and classification for 19980421_CNN video. The thumbnails represent the shots in the video. The shots that are enclosed in the bold boxes form the 'news' stories, while the other shots form 'miscellaneous' stories.

Fig. 13. Results of story segmentation and classification for 19980426_CNN video. The thumbnails represent the shots in the video. The shots that are enclosed in the bold boxes form the 'news' stories, while the other shots form 'miscellaneous' stories.

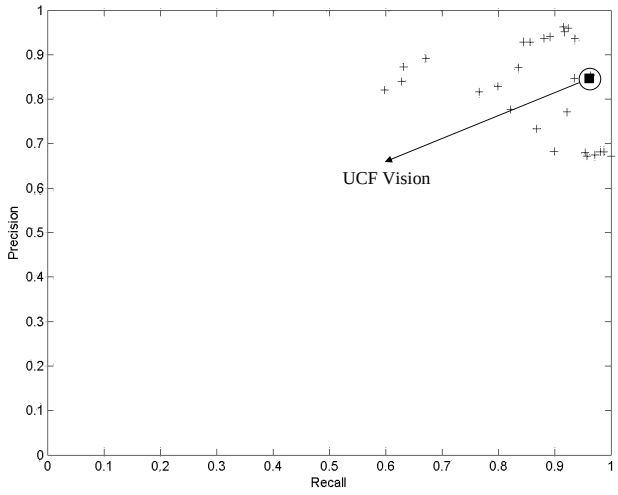

Fig. 14. UCF standing for the story classification task.

ries are merged into a single story.

3.2. Results and Discussions

The data set for the story segmentation and classification task consists of two parts: development data and testing data with their manual shot segmentations. There are 58 ABC videos, 59 CNN videos and 19 C-SPAN videos in the development set, and 117 videos in the testing set. We have applied the story segmentation method on the CNN and ABC videos, since in each of C-SPAN videos, there is only one story.

This task has two stages: (1)Segmentation - the videos are segmented into stories. (2)Classification - the semantic labels 'news' and 'miscellaneous' are applied to the stories accordingly. UCF Vision group has submitted the results for only one run of the story classification task. The submitted stories are compared with the reference data from NIST based on their starting time in terms of seconds. Our standing among the participants for the precision/recall scale is shown in Fig. 14. In Fig. 12 and Fig. 13, we show the results for two testing videos: 19980421_CNN and 19980422_CNN. In the figure, the thumbnails represent the shots in the video. The adjacent shots that are enclosed in the bold boxes form the 'news' stories. The shots that are in between the 'news' stories form the 'miscellaneous' stories.

4. FEATURE EXTRACTION

There are two kinds of features: low level features and high level features. Low level features are the ones that can be computed directly from the given information, for example, color histogram, motion field, audio frequency, and etc. High level features are the semantic labels of a shot, for example, outdoor, sport events, weather news, person, and etc. These feature can be derived from the results of one or multiple low level features. There are 17 semantic labels in TRECVID 2003: outdoor, news subject face, people, building, road, vegetation, animal, female speech, car/truck/bus, aircraft, news subject monologue, non-studio settings, sport events, weather news, zoom in, physical violence, and specific person. We have submitted one run for two features: Non-Studio Settings and Weather News. In Section 4.1, we present the method for detecting shots with non-studio settings, and in Section 4.2, we present our work on the detection of weather news shots.

4.1. Non-Studio Settings

In news related videos, the shots that have studio settings are usually the ones with news anchorperson/s. Therefore, we can transform the problem of finding the shots with non-studio settings to the problem of finding the shots with news anchorperson/s. In this project, we only consider the case that there is only one news anchorperson in each shot. In the news video, the anchorperson shows more frequently than other materials. Knowing this fact, we have the following method for detecting shots that contains news anchorperson:

1. We define the frequency of a shot as the number of times this shot is repeated in the video. We take the middle frame F_i as the key frame of shot i. Therefore, the similarity between two shots i and j can be computed as the histogram intersection between F_i and F_j. For each shot i, we compute its similarities to all

Fig. 15. Results for the shots with News Anchorperson for 19980421_CNN video. The thumbnails show the key-frames of the shots. The last row shows the shots containing the news anchorperson.

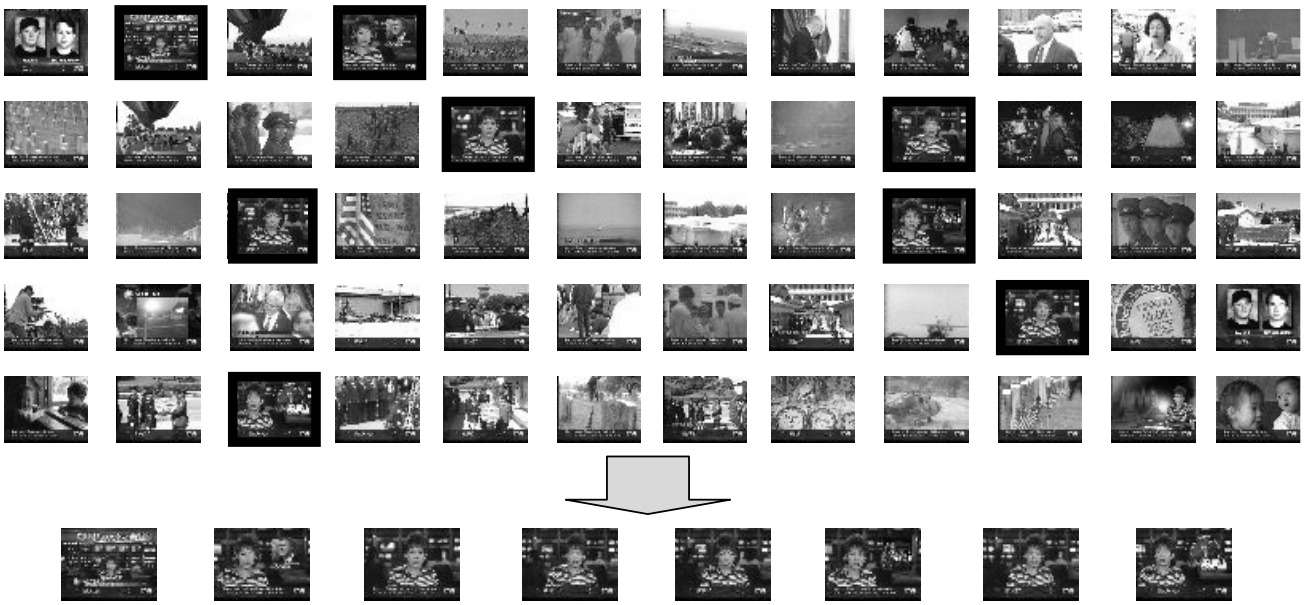

Fig. 16. Results for the shots with News Anchorperson for 19980426_CNN video. The thumbnails show the key-frames of the shots. The last row shows the shots containing the news anchorperson.

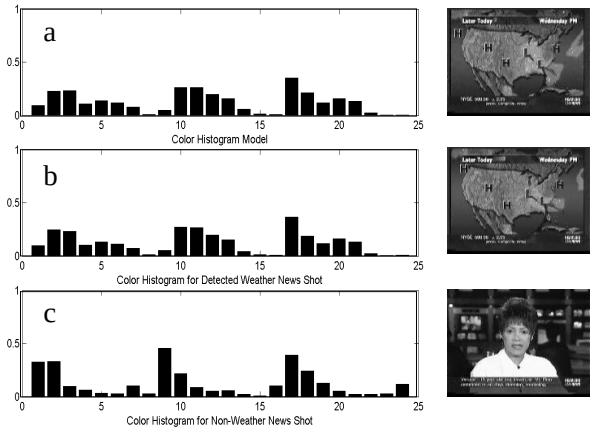

Fig. 17. (a) One of the color distributions in the histogram model and its representative frame. (b) The color histogram of the detected weather news shot and its key-frame. (c) The color histogram of a non-weather news shot and its key-frame.

the other shots. The frequency of shot i then is the total number of the similar shots found (including the shot itself) in the video.

2. The shots with the highest frequencies should contain the news anchorperson. In a typical 30 minutes news program, the maximum number of news anchorperson appeared is three. We select the shots with the top three frequencies as our initial candidate set K for the shots with news anchorperson.

$$K = \{k_1, k_2, k3\} \tag{8}$$

where $k_i,\ i = 1, 2, 3$ are the shots numbers.

3. Next, we expand the candidate set K, such that it has all the shots containing news anchorperson. For every shot j, we find its similarities to the ones in the candidate set K. If their similarity (histogram intersection between frames F_j and F_{k_i}) is large, then shot j is included in the set K.

4. Since we are considering the case that there is only one news anchorperson in each shot, we need to eliminate the outliers. These outliers could be caused by some commercials. We apply Haar face detector [7] on the key frames of the shots in the candidate set and delete those shots containing none more than one faces.

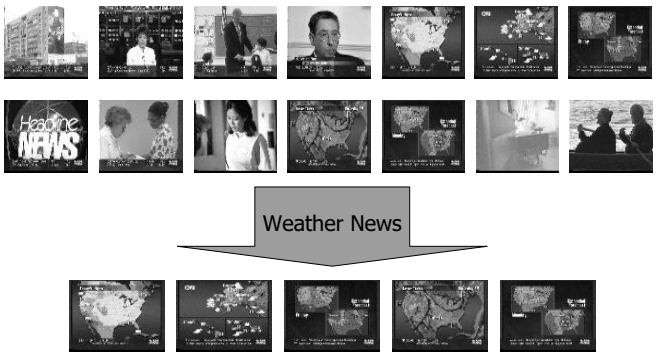

Fig. 18. Results for the detection of the shots with feature 'Weather News'.

5. Finally, the non-studio settings shots are determined by removing the all studio setting shots from the complete set of shots.

Two example results of detection of shots with news anchorperson are shown in Fig. 15 and Fig. 16. In these two figures, the thumbnails are the key-frames of the shots. The ones that are enclosed in the bold boxes are the shots with news anchorperson. The average precision scale for our result on extracting 'Non-Studio Setting' feature is 0.035.

4.2. Weather News

We have concentrated our efforts on finding the shots of weather forecast news. By observing the training videos, we have discovered that there are certain color distributions for the weather forecast shots. Most of the forecast news are blueish. Some of them are yellowish. Therefore, we have built a color model with several distributions from the training videos.

1. From the development data set, we select the shots that are weather forecast news. For each training shot i, we compute the average color histogram D^i_{avg} using the frames from the entire shot. Then, all the averaged histograms are put into k distinctive groups based on the means and variances of RBG channels. Our color model is $T = \{t_1, t_2, ..., t_k\}$, where t_x is the average histogram of the one in group k (Fig. 17(a)).

2. For each incoming testing shot j, we compute the average histogram D^j_{avg}. The similarity between D^j_{avg} and the color model is determined as the maximum of the histogram intersections of D^j_{avg} and $t_i, i = 1...k$. If the similarity is above some color threshold, the testing shot j is declared as 'Weather News' (Fig. 17(b)). Otherwise, the shot is declared as a 'Non-Weather News' shot(Fig. 17(c)).

In our submission for feature 'Weather News', we have 64 detected shots in which 63 shots are correct. The average precision is 0.368. For above two features, we only used the visual information provided in the MPEG-1 videos. The results are very promising.

5. CONCLUSION

The TRECVID 2003 forum motivated multimedia community to improve video processing and analyzing techniques. We have described our methods for shot-boundary detection, story segmentation and classification, and feature extraction.

In shot-boundary detection, our method is a coarse-to-fine approach with a multi-level histogram intersection technique. This method can detect and classify the transitions as 'abrupt' and 'gradual'. The story segmentation and classification method is based on the knowledge of television program editing techniques and uses only the visual cues in the video sequences. We have submitted results for extracting two features: Non-Studio Settings and Weather News. For 'Non-Studio Settings' feature, we first detect all the shots that contain news anchorperson by determining the frequency of each shot. Then, the shots with feature 'Non-Studio Settings' are the complements of the detected news anchorperson shots. We built a color histogram model for detecting shots with 'Weather News' feature.

In all three tasks, our system only incorporates the visual information in the video sequences. The results show that our video analysis methods are efficient and robust.

6. REFERENCES

[1] A. Hampapur, et al., *Virage Video Engine*, Proc. SPIE, Storage and Retrieval for Image and Video Databases, 1997

[2] Hanjalic, A. et al., *Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems*, IEEE Tran. on CSVT., Vol:9 Issue:4, 1999.

[3] http://www-nlpir.nist.gov/projects/tv2003/tv2003.html

[4] http://www.m-w.com

[5] W. Ngo et al., *Motion-Based Video Representation for Scene Change Detection*, IJCV, 2001.

[6] Z. Rasheed, M. Shah, *Scene Detection In Hollywood Movies and TV Shows*, IEEE *Computer Vision and Pattern Recognition Conference*, Madison, Wisconsin, June 16-22 2003.

[7] P. Viola, M. Jones, *Robust Real-Time Object Detection, International Journal of Computer Vision*, 2001

[8] Yeung, M., Yeo, B.-L., and Liu, B., *Extracting Story Units from Long Programs for Video Browsing and Navigation* in *International Conference on Multimedia Computing and Systems*, June 1996.

[9] Yeung, M.M. et al., *Segmetation of Videos by Clustering and Graph Analysis*, CVIU, Vol.71, No:1, 1998.

[10] H.J. Zhang, A. Kankanhalli, S.W. Smoliar, *Automatic Partition of Full Motion Video*, Multimedia System1, pp.10-28, 1993.

[11] D. Zhang, W. Qi, H.J. Zhang, *A New Shot Detection Algorithm*, 2nd IEEE *Pacific-Rim Conf on Multimedia* (PCM2001), pp. 63-70, Beijing, China, October 2001.