NVIBRS - News Video Indexing, Browsing and Retrieval System

Vakkalanka Suresh S. Palanivel B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai-600 036, India.
Email:{suresh,spal,yegna}@cs.iitm.ernet.in

Abstract

This paper presents a system for automatic news video indexing, browsing and retrieval. The system employs visual clues to effectively parse the news video into story units, to generate.visual-table-of-contents and indexes for supporting browsing and retrieval. The efficiency of the system has been tested on several video sequences.

1 Introduction

With the never ending advances in digital technology, more and more video data is generated every day. As the accessible video collections grow, efficient schemes for navigating, browsing, and retrieving video data are required. A good survey of techniques for automatic indexing and retrieval of video data can be found in [1, 2]. Among the various video categories, news programs are important storing objects, due to the fact that they concisely cover large number of topics related to society, politics, business, sports, weather, etc. In recent years, several prototype systems have been proposed for automatic or semi-automatic parsing and indexing of news video allowing interactive news navigation, content-based retrieval and news-on-demand (NOD) applications [3, 4, 5, 6, 7].

These systems employ a two stage classification scheme. First, the video is segmented into shots, and each shot is then classified into anchor and non-anchor categories. Most of the work in shot classification can be categorized into two classes of approaches. One is model based, and the other is based on unsupervised clustering. In [7], by using an offline trained audio model, the theme music segment of the given news program is detected. From that, an anchor frame is located, from which a feature block is extracted to build an on-line visual model for the anchor. Limitations of this approach are : 1) This method will fail in case of a news video whose theme music is not present in our database, 2) When the theme music of a news program changes, the database must be updated, 3) When the set of audio model grows, the model matching method is highly compute intensive. In [6] offline trained audio-visual anchorperson models are used to detect the anchorpersons. The drawback in this approach is that, the approach will fail if an anchorperson whose model is not present in the set of models.

With the unsupervised clustering approach, key frames extracted from each shot are clustered using a similarity metric and the anchor key frames may be identified as the ones from the largest clusters. This kind of methods will work only when the visual appearance of the studio scenes within the news video basically remain the same. Also we need apriori knowledge of the number of anchorpersons in the current video in order to decide how many such large clusters are to be considered.

In this paper we present a system for automatic news video parsing, indexing, browsing and retrieval. This work utilizes our earlier work on Autoassociative neural network based person identification [8], to automatically detect and index anchorpersons in a news video. Our model based approach is superior to the other approaches where offline trained audio-visual models of anchorpersons are used which involve manual collection of training data and provide little flexibility. In our system the anchorperson models are created online without any human supervision and the models once created can be used as off-line models to detect the anchorperson appearance in a different video.

Our system consists of the following modules:

- *news indexer* - a news indexer that automatically parse a news video into story units by using online trained anchorperson models and the domain knowledge of news videos. Extracts key frames from each story unit and constructs the visual-table-of-contents. This module also classifies the video segments into anchor, speech/report, computer graphics, and miscellaneous categories.
- *news browser* - a news viewer that displays visual-table-of-contents generated from the news indexing results. It will show the anchorperson segments and the associated news segments and will allow playback of the news items.

0-7803-8840-2/05/$20.00 ©2005 IEEE

- *news retriever* - a news retrieval module that enable specific news items to be retrieved from the video database. User can choose the news category and from the list of news items display, select the desired news clip for playback.

This paper is organized as follows: Section 2 describes the news indexing module. Experimental results are discussed in Section 3. Section 4 describes the video browsing and retrieval system. Finally conclusions and the scope for future work are discussed in Section 5.

2 News Indexer

The function of the news indexing module is to extract the temporal, structural, and semantic information of news video programs. It involves detecting the story boundaries and parsing the complete video into story units, constructing visual-table-of-contents, and classifying the video segments into predefined categories. For effective news browsing and retrieval, reliable news story parsing is crucial. Usually, a news program is a simple sequence of news stories (possibly interleaved with commercials), each of which can be further segmented into an introduction by the anchorperson followed by a detailed report. Since an anchorperson hosts a news program, locations of anchorperson segments in the news video provides landmarks for detecting story boundaries. The proposed approach for news parsing consists of two steps: Segmentation of complete video into low and high motion video segments, and identifying the anchorperson segments from the low motion segments.

In general, it can be observed that during the news story introduction by an anchorperson the background around the anchorperson is almost constant, accompanied by a small motion of the anchorperson in the foreground. Whereas during detailed reporting, motion in the background as well as in the objects of interest is high most of the times. We segment the complete video into low and high motion segments by employing a binary pattern matching method on the motion based binary feature vector derived for the complete video. The motion metric is computed from the thresholded difference image between two frames. To reduce the sensitivity of the motion metric to noise, individual frames are smoothed using the 1D-smoothing technique described in [9].

2.1 Definition of the motion metric

Let f_i and f_j represent the i^{th} and j^{th} smoothed frames obtained by using 1D-smoothing technique respectively. The thresholded difference image D between f_i and f_j is given by

$$D(x,y) = \begin{cases} 1, & \text{if } |f_i(x,y) - f_j(x,y)| > \tau \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

where τ is the threshold. Then the amount of motion M between frames f_i and f_j is obtained by

$$M = \frac{1}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} D(x,y) \qquad (2)$$

where W and H represents the width and height of each frame, respectively.

2.2 Video segmentation

To reduce the computational complexity as well as to enhance the sensitivity to motion, the metric is computed between frame pairs that are at a fixed interval $\triangle t$ (20 frames) apart. The motion metric based binary feature vector for the complete video is obtained as

$$B_t = \begin{cases} 1, & \text{if } M_t > \lambda \\ 0, & \text{otherwise} \end{cases} \qquad 0 \leq t < (T/\triangle t) \qquad (3)$$

where λ is the threshold empirically chosen from the data and T is the total number of frames in the video. A binary pattern matching method is employed on this feature vector to segment the complete video into low and high motion segments. Segments of low motion correspond to the sequence pattern "00...0". Patterns with at least three consecutive 0s only are considered as low motion segments.

2.3 Video segment classification

The low motion segments that are obtained during the video segmentation process correspond to anchorperson segments, and some non-anchor person segments like speech/report and graphic objects. By using the visual clues derived from the video segments, these low motion segments are classified into anchorperson and non-anchorperson categories. The non-anchorperson segments are further classified into speech/report, computer graphic, and miscellaneous categories. The decision tree for the video segment classification is show in Figure 1.

2.3.1 Visual feature extraction

The proposed visual feature extraction has four important modules: Face localization, eye location estimation, mouth location estimation and feature vector extraction. The visual features extracted are insensitive to the location, size of the anchor, color and visual content in the studio background.

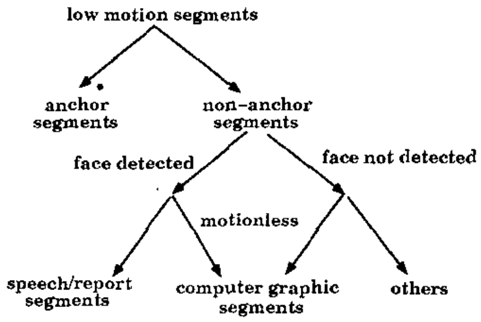


Figure 1: Decision tree for video segment classification.

Face localization

Recent methods for face detection use neural networks [10], skin color segmentation [11] and motion information [12], [13] for tracking faces in video. The method employed in this work is sufficient and simple enough to serve our purpose and can be replaced by any other face detection algorithm. Our approach of face localization contains two major modules 1) skin regions detection, and 2) face region approximation. A Gaussian Mixture Model (GMM) is used to model the skin color in YC_bC_r color space. Skin color patches extracted from various still images and video frames, covering a large range of skin color appearance have been used for training the model. Given an image, we can classify the regions of the image into two classes by finding the likelihood of each pixel to be a skin pixel. Figure 2(a) shows the result of skin color regions detection applied on the original image Figure 2(b).

The bounding box for the face region is found using the horizontal and vertical projections of the binary image obtained from the skin region detection as shown in Figures 2(c) and 2(d). The first mean crossing points as we move away from peak location on both sides of the peak value in the horizontal projection are taken as left and right boundaries of the face region. The top part of the face is obtained from the vertical projection of the sub image between left and right boundaries. The first mean crossing point as we move away from the peak towards left in the vertical projection is taken as face top location. The height of the face is estimated from the width of the face ($(4/3)\times$width), and a rectangular bounding box for the face can be obtained.

Eye location estimation

Once the approximate face region is found, the next step is to extract the location of eyes. The eye location algorithm is based on our earlier work in [13]. Face region within the bounding box is thresholded to obtain the thresholded face image U, given by

$$U(x,y) = \begin{cases} 255, & \text{if } Y(x,y) < \lambda_1 \text{ and} \\ & C_r(x,y) < \lambda_2 \text{ and} \\ & C_b(x,y) > \lambda_3 \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

where λ_1, λ_2 and λ_3 are the average Y, C_r and C_b values of the pixels in the forehead region, respectively. Morphological closing operation is applied to the thresholded face image, and the centroid of all the blobs are estimated. The relative positions of the centroids with respect to the rectangular bounding box enclosing the face region and the contrast information in the eyebrow region are used to determine the location of the eyes. The method can detect the location of the eyes in the presence of eye glasses as long as the eyes are visible.

Mouth location estimation

For estimating the mouth location, we model the the color distribution of the non lip regions of the face using a Gaussian distribution and it is used to detect the lip region pixels. The non lip region pixels are extracted based on the eye location as shown in Figure 3(a). The distribution of Y, C_r, and C_b values of the non lip and the lip pixels are shown in Figure 3(b). The centroid of the mouth is estimated using the pixels coordinates of the detected lip pixels.

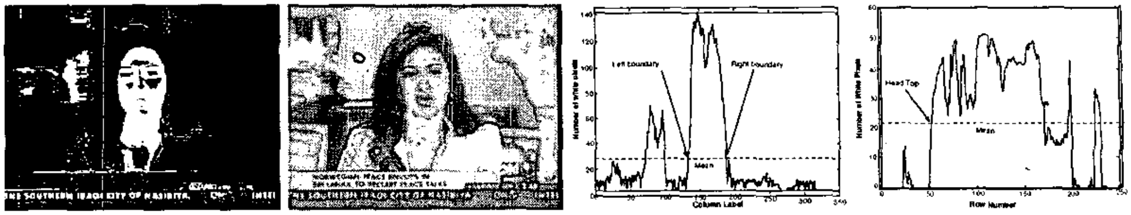
Feature vector extraction

The facial features such as eyes, eyebrows, nose, mouth and face outline in a still image and the shape of the lip contour during speaking plays an important role in recognizing persons by face. A cartoonist extracts information that is unique to a person from these parts to represent in terms of lines and arcs. These lines and arcs correspond to the gradient or local extrema (maxima and minima) in an image. The local maxima and minima can be extracted using gray scale morphological dilation and erosion operations, respectively [14]

An elliptical rigid grid placed over the facial region and a rectangular rigid grid placed over the mouth region, are used for feature extraction as shown in Figures 3(c) and 3(d), respectively. The elliptical grid consists of 73 nodes and the rectangular grid consists of 25 nodes. The position of these nodes are determined relative to the location of the eyes and the mouth center. Gray scale morphological operations are employed at each node to extract the features.

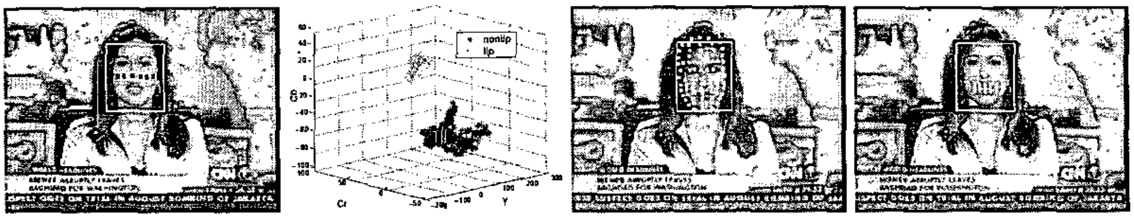
Definition of morphological operations

Let $\mathcal{Z}$ denote the set of integers. Given an image I: $\mathcal{D} \subseteq \mathcal{Z}^2 \longrightarrow \mathcal{Z}$ and a structuring function G: $\mathcal{G} \subseteq \mathcal{Z}^2 \longrightarrow \mathcal{Z}$,



(a) (b) (c) (d)

Figure 2: Face localization: (a) Binary image obtained from the skin color region detection, (b) corresponding original image, (c) horizontal projection of the binary skin color region detected image and (d) vertical projection of the region within left and right boundaries of binary image



(a) (b) (c) (d)

Figure 3: Feature Extraction: (a) Non-lip regions, (b) Distribution of lip and non-lip pixels, (c) facial feature extraction, and (d) lip feature extraction.

the erosion of the image I by the structuring function G is denoted by $(I \ominus G_\sigma)$, and is defined as

$$(I \ominus G_\sigma)(i,j) = \min_{x,y}\{I(i+x,j+y) - G(x,y)\} \quad (5)$$

and the dilation of the image I by the structuring function G is denoted by $(I \oplus G_\sigma)$, and is defined as

$$(I \oplus G_\sigma)(i,j) = \max_{x,y}\{I(i+x,j+y) + G(x,y)\} \quad (6)$$

where $-M_a \leq x, y \leq M_b$, with $1 \leq i \leq W$, $1 \leq j \leq H$. The size of the structuring function is decided by the parameters M_a and M_b, and is given by $(M_a + M_b + 1) \times (M_a + M_b + 1)$. For a flat structuring function $G_\sigma(x,y) = 0$, then the expressions for erosion and dilation reduces to

$$(I \ominus G_\sigma)(i,j) = \min_{x,y}\{I(i+x,j+y)\} \quad (7)$$

and

$$(I \oplus G_\sigma)(i,j) = \max_{x,y}\{I(i+x,j+y)\} \quad (8)$$

respectively.

The erosion operation (7) is employed at each node of the elliptical grid to obtain a 73 dimensional feature vector and the dilation operation (8) is employed at each node of the rectangular grid to obtain a 25 dimensional feature vector. Each feature vector is normalized to $[-1, 1]$. The normalized feature vector is less sensitive to variation in the image brightness.

Autoassociative neural network based person modeling

Autoassociative neural network (AANN) models are the feedforward neural networks performing an identity mapping of the input space, and are used to capture the distribution of the input data [15]. The five layer autoassociative neural network model as shown in Figure 4, is used to capture the distribution of the feature vectors. The structures of the AANN models used in our study are $73L$ $90N$ $30N$ $90N$ $73L$ and $25L$ $40N$ $10N$ $40N$ $25L$, where L denotes a linear unit, and N denotes a nonlinear unit. The standard back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector. AANN based person identification is based on our earlier work [8], where the method is successfully tested on more than 50 subjects.

2.3.2 Anchor segment classification

At the top level in the video segment classification the low motion segments are classified into anchor and non-anchor segments. The anchor segment classification has the following steps:

1. Consider the low motion segments in the descending order of duration.
2. Apply face detection algorithm for the first few frames of the current low motion segment. If a face is detected

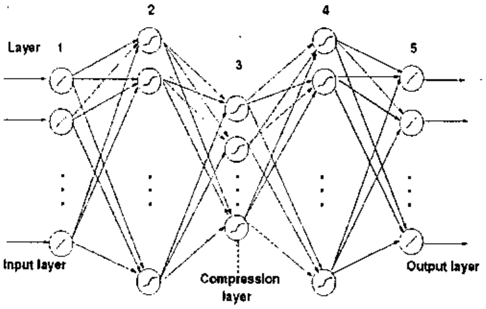


Figure 4: Structure of a 5 layer autoassociative neural network model.

extract visual features from the first few frames (50 in our case), else declare the low motion segment as non-anchorperson segment and goto Step 6.

3. Test against the existing anchorperson model pairs (facial and mouth). If any model pair gives confidence score (sum of the individual model confidence scores in a pair) above certain threshold α, declare the low motion segment as anchorperson segment and goto Step 6.
4. Test against the model pairs if any, created from the current video as described in Step 5. If any model pair gives a confidence score above the threshold α then goto Step 6.
5. Extract visual features for the entire segment and train the model pair (facial and mouth models).
6. Repeat steps 2 to 5 for the next low motion segment.
7. For each visual model pair created using segments of the current video, find the temporal distance between the first and last occurrence of the low motion segments corresponding to that model pair. Since the anchor appears many times through out the news video, if the distance is greater than half of the total duration of the news video, the model pair is declared as an anchorperson model pair. All the low motion segments corresponding to this model pair are declared as anchor segments.

2.3.3 Non-anchor segment classification

The non-anchor segments that are obtained during the anchor segment classification can be categorized into segments where a face has been detected and segments where no face has been detected. The segments where a face has been detected may correspond to: Speech/report segments, and computer graphic (CG) segments containing a referential portrait picture. A distinctive property of CG segments is that they must remain stable for a certain amount of time so people can read and understand. Nevertheless, there are occasional movements, for example a financial report showing the stock market figures. So, the CG segments are classified by total duration of motionless frames. In our case, the duration was set to one second.

2.4 Visual-table-of-contents construction

Visual-table-of-contents construction or video abstraction is the process of creating a presentation of visual information about the structure of a video, which should be much smaller than the original video. Key frames play an important role in the video abstraction process. Key frames are still images, extracted from original video data, that best represent the content of a story in an abstract manner. Since motion is the major indicator for content change, dominant motion components resulting from camera operations and large moving objects are the most important source of information. So, in an effective approach to key frame extraction, the number of key frames needed to represent a segment of video should be based on temporal variations of video content in the segment.

In our approach to key frame extraction, the binary motion vector derived during the structure analysis process, as defined in Section 2.2 is reused to extract the key frames. As described earlier, the low activity video segments correspond to binary pattern "00..0" and the high activity video segments correspond to the binary pattern "11...1" or "01" or "10". From each of the low motion regions, the middle frame is taken as the key frame and from each of the high activity regions, key frames are extracted at the interval Δt (20 frames).

3 Experimental Results

The proposed method has been evaluated on more than 10 hours of news video data recorded at 25 frames per second and frame size 320 × 240 from 4 news channels: BBC World, CNN, NDTV 24 × 7 and ETV. To evaluate the performance of the video segment classification, we use the standard *precision* and *recall* criteria, shown in the following:

$$\text{precision} = \frac{\text{number of hits}}{\text{number of hits} + \text{number of false alarms}} \quad (9)$$

$$\text{recall} = \frac{\text{number of hits}}{\text{number of hits} + \text{number of misses}} \quad (10)$$

A precision of 98.75% and a recall of 97.5% for anchorperson indexing, a precision of 71% and a recall of 74.5% for speech/report classification, and a precision of 55% and a recall of 98% for computer graphic classification, have been achieved. Major reasons for false classification and misclassification were:

- Eyes and lips could not be detected, since the face was not a full face.
- A complete still image of an object was misclassified as CG segment.
- CG segments with portrait pictures and are not stable for a certain amount of time were misclassified as speech/report segments.
- Missed anchorperson segments are classified as speech/report segments.
- low motion segments less than 2 seconds in duration were not considered for classification.

4 Video Browsing and Retrieval System

The system is developed on the Windows platform using Microsoft DirectX SDK and MFC in VC++. The news browser allows the users to randomly browse through the available video documents in the database. On selection of a video document it displays a tree view of the story units in the current video. On selection of a new story it will show the key frames for the current story unit and allows the playback of the news items. The retrieval system allows specific news items to be retrieved from the video database. User can issue a textual query by specifying the video category or he/she can choose the news category from a list, and select the desired news clip for playback from the list of news items retrieved.

5 Conclusions

We have presented a system for automatic news video parsing, indexing, browsing and retrieval. The system provides fairly accurate process for the parsing and classification of news items from the news program and the facility to browse and retrieve such news items from a database. The proposed AANN model based approach to automatically detect and index anchorpersons in a news video is superior to the methods where off-line trained audio-visual models of anchorpersons are used which involve manual collection of training data and provide little flexibility. In the proposed approach the anchorperson models are created online without any human supervision and the models once created can be used as off-line models to detect the anchorperson appearance in a different video. To further enhance this news video system we can integrate close caption or transcript text information into the system so that we will get more semantic info. which will make the indexing and retrieval more powerful.

References

[1] R. Brunelli, O. Mich, and C. Modena, "A survey on the automatic indexing of video data," *Journal of Visual Communication and Image Representation*, vol. 10, no. 2, pp. 78–112, 1999.

[2] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, pp. 12–36, Nov. 2000.

[3] H. D. Wactlar, T. Kanade, and M. A. Smith, "Intelligent access to digital video: Informedia project," *IEEE Comput. Mag.*, vol. 29, pp. 45–52, May 1996.

[4] M. Bertini, A. D. Bimbo, and P. Pala, "Content-based indexing and retrieval of TV news," *Pattern Recognition Letters*, vol. 22, pp. 503–516, Apr 2001.

[5] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," *IEEE Trans. Circuits, Systems, Video Technology*, vol. 12, pp. 765–776, Sept. 2002.

[6] A. Albiol, L. Torres, and E. J. Delp, "The indexing of persons in news sequences using audio-visual data," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, (Hong Kong), Apr 6-10, 2003.

[7] Z. Liu and Q. Huang, "Adaptive anchor detection using online trained audio/visual model," in *Proc. SPIE Conf. Storage and Retrieval for Media Database*, (San Jose, CA), pp. 156-167, Jan. 2000.

[8] S. Palanivel, B. S. Venkatesh, and B. Yegnanarayana, "Real time face recognition system using autoassociative neural network models," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, (Hong Kong), pp. 833-836, Apr 6-10, 2003.

[9] P. K. Kumar, S. Das, and B. Yegnanarayana, "One-dimensional processing of images," in *Int. Conf. Multimedia Processing and Systems*, (Chennai, India), pp. 181-185, Aug. 13-15, 2000.

[10] H. A. Rowley, S. Baluj, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, May 1998.

[11] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavlet packet analysis," *IEEE Trans. Multimedia*, vol. 1, pp. 264–277, Sept. 1999.

[12] B. Li and R. Cehellappa, "A generic approach to simultaneous tracking and verification in video," *IEEE Trans. Image Processing*, vol. 11, pp. 530–544, May 2002.

[13] S. Palanivel, B. S. Venkatesh, and B. Yegnanarayana, "Real time face authentication system using autoassociative neural network models," in *Int. Conf. Multimedia and Expo*, (Baltimore, MD, USA), pp. 257-260, July 6-9, 2003.

[14] P. Jackway and M. Deriche, "Scale-space properties of the multiscale morphological dilation-erosion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 38–51, Jan. 1996.

[15] B. Yegnanarayana and S. Kishore, "AANN: an alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, Jan 2002.