Integrating Text and Face Detection for Finding Informative Poster Frames

Michael Smith[1]
msmith+@cs.cmu.edu
http://www.cs.cmu.edu/~msmith

Shumeet Baluja[2,3]
baluja+@cs.cmu.edu
http://www.cs.cmu.edu/~baluja

Henry A. Rowley[3]
har+@cs.cmu.edu
http://www.cs.cmu.edu/~har

[1]Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2]Justsystem Pittsburgh Research Center, 4616 Henry Street, Pittsburgh, PA, 15213, USA
[3]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Abstract

Digital video is rapidly becoming an important source of information and entertainment, and is used in a host of multimedia applications. With the size of digital video collections growing to many thousands of hours, technology is needed to allow rapid browsing of videos. One way to summarize a video is to select *poster frames* to represent segments of the video. Previous techniques for extracting poster frames were based on scene segmentation, using color histograms or optical flow. To provide more informative poster frames, this work combines algorithms for extracting image content, specifically faces and on-screen text, with existing scene segmentation technology.

Introduction

A useful way of browsing a video segment is through a set of *poster frames*, which statically represent a video segment's contents (Mills, Cohen, & Wong 1992). Poster frames are generally selected from scenes, which are segmented by color histogram or optical flow techniques (Arman *et al.* 1994; Hampapur, Jain, & Weymouth 1995; Zhang, Kankanhalli, & Smoliar 1993). To date, the semantic content of the frames has not been effectively used in poster frame selection. We explore two techniques for content extraction: detection of text captions and human faces in video. A system which can combine these content extraction methods with traditional scene segmentation will produce poster frames which are more informative than those produced by scene segmentation alone.

Captions provide significant information about the content of a scene, such as people's names and affiliations, and locations and descriptions of events. Human faces are among the most interesting objects in video. Speakers and interviewees are often shown in full-frame, frontal shots. Detection of both text and faces has been widely studied, and robust algorithms exist for both problems (Smith & Kanade 1996) (Rowley, Baluja, & Kanade 1996; Sung & Poggio 1994). For this work, we used a local spatial frequency-based system for text detection, and a neural network-based system for face detection, both developed at CMU.

In the next sections, we describe the face and text detection algorithms. Results are given for each algorithm. We then describe how to combine these algorithms for poster frame selection. This paper concludes with a summary of the results and directions for future research.

Face Detection

In recent years, there have been several research efforts to detect faces in arbitrary images. Some of them use matching of hand-coded templates (Yang & Huang 1994), others use geometric models (Burl & Perona 1996), while others use a view-based approach (Vaillant, Monrocq, & Le Cun 1994; Sung & Poggio 1994; Rowley, Baluja, & Kanade 1996), treating the whole face as a pattern to be recognized.

This section briefly presents our face detection system (from (Rowley, Baluja, & Kanade 1996)), which locates frontal, upright faces at any location or scale in grayscale images. It is a view-based technique, implemented using neural networks.

Description of System

Our system operates in two stages. It first applies a set of neural network-based filters to an input image, and then uses an arbitrator to combine the filter outputs. The neural-network based filter (shown in Figure 1) receives as input a small square region of the image, and generates an output ranging from 1 to -1, signifying the presence or absence of a face, respectively. To detect faces anywhere in the input, the filter must be applied at every location in the image. To detect faces larger than the window size, the input image is repeatedly reduced in size (by subsampling), and the filter is applied at each size.

To train the neural network, we need positive and negative example images. Positive examples were collected from face databases at Carnegie Mellon University and Harvard, and random negative examples were created. The network was trained, and then applied to an image of scenery with no faces in it. Any regions which the network incorrectly identified as faces were added into the training set as new negative examples. We continue training by applying the network to more images of scenery.

Although the detector described above can detect a high percentage of faces, it produces an unacceptable number of false detections. We use two approaches to reduce the number of false detections:

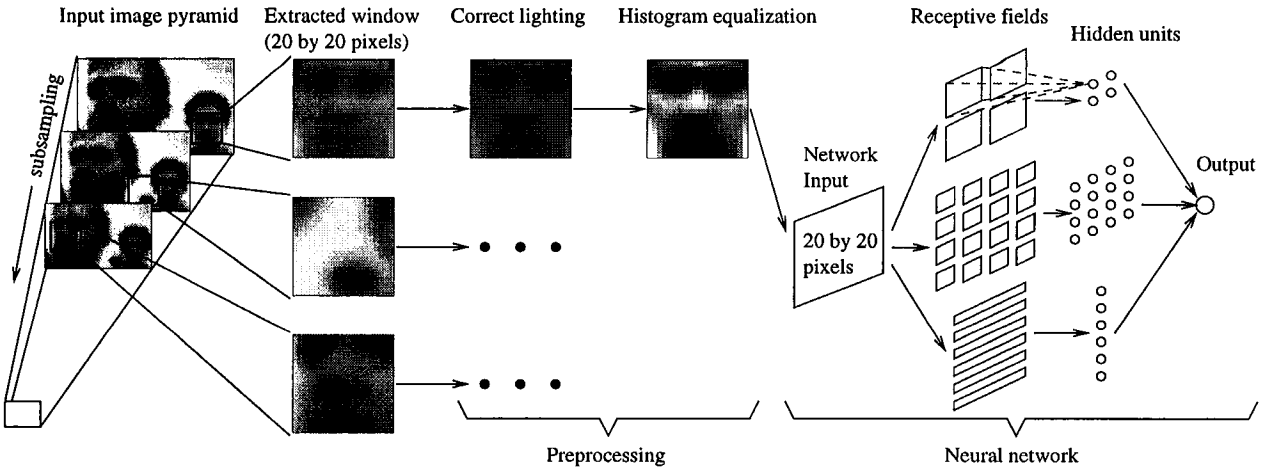

Figure 1: The basic algorithm used for face detection.

Detection Merging: Often real faces are detected at multiple nearby positions and scales, and false detections occur only at a single position. By setting a minimum threshold on the number of detections, many false detections can be eliminated. Because faces rarely overlap, if one detection overlaps with another, the detection with lower confidence can be removed.

Multiple Network Arbitration: During training, identical networks with different random initial weights will develop different biases and make different mistakes. We can exploit this by arbitrating among the outputs of multiple networks, for instance signalling a detection only when two networks agree that there is a face.

Results

The system was tested on a set of 130 images. The images contain 507 faces, and require the network to scan a total of 83,099,211 windows. A single network can detect 470 of the faces (92.7%), with 1768 false detections. Using the detection merging heuristics reduces the false detections to 844, but also removes nine correct detections. Using arbitration among two networks and detection merging, the detection rate is 85.9%, with 63 false detections. Some example output images are shown in Figure 2. Details of the arbitration and detection merging heuristics and more results are presented in (Rowley, Baluja, & Kanade 1996).

Text Detection

Text in the video provides significant information as to the content of a scene. For example, statistics are not usually spoken but are included in the captions for viewer inspection. Names and titles are attached to close-ups of people. A text region is a horizontal rectangular structure containing clusters of sharp edges, due to characters with high contrast color or intensity. By detecting these properties we can extract regions from video frames that contain textual information. Figure 3 illustrates the process. We first apply a 3x3 horizontal differential filter and appropriate binary thresholding to the entire image to extract vertical edge features. Then smoothing is applied to eliminate extraneous fragments and to connect edge elements that may have been detached. Individual regions are identified by cluster detection and bounding rectangles are computed. We then extract clusters with bounding regions that satisfy the constraints shown in Figure 4.

A cluster's bounding region must have a small vertical-to-horizontal aspect ratio as well as satisfying various limits in height and width. The fill factor of the region should be high enough to ensure dense clusters. The cluster size should also be relatively large to avoid small fragments. Finally, we examine the intensity histogram of each region to test for high contrast. This is because certain textures and shapes are similar to text but exhibit low contrast when examined in a bounded region. This method works best with horizontal titles and captions. Table 1 shows the statistics for text detection accuracy on various sets of images. Figure 5 shows several example text detection results.

Combination of Face and Text Detection

A simple way to combine face and text detection is to create poster frames from the scenes which contain both. If this does not result in enough poster frames, we can select frames having just text or just faces. Finally, in the absence of these features, we can fall back on picking the first frame of a set of scenes which are found by scene segmentation.

For specialized applications, such as news broadcast video, we can incorporate more domain-specific knowledge. For example, in the CNN footage shown in Figure 6A, sports results are constantly shown at the bottom of the screen. Some frames (such as the top frame in Figure 6B) contain logos like "Headline News", again giving little information. For this limited domain, the most important text usually appears within a small region of the screen (where people's names or affiliations often appear, for example see Figure 7C), so we can disregard text located outside of this region. In the future, heuristics which rely on

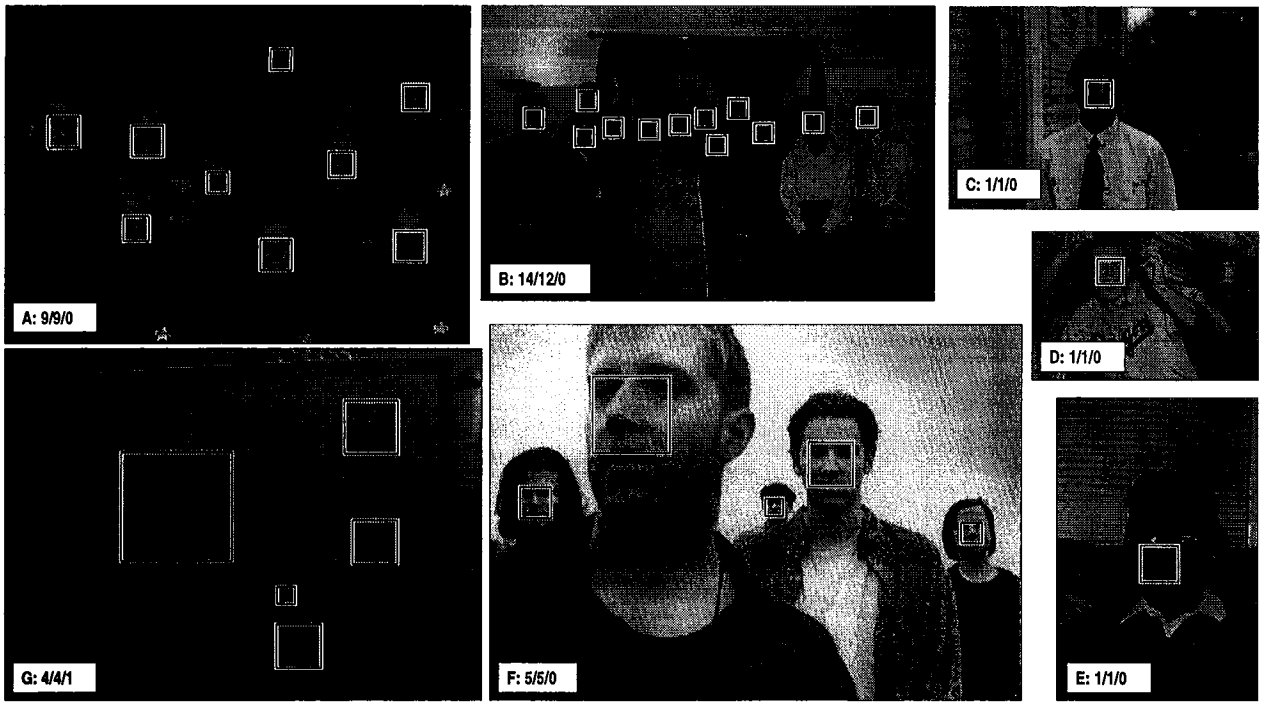

Figure 2: Output from the face detector. For each image, three numbers are shown: the number of faces in the image, the number of faces detected correctly, and the number of false detections. Although the system was not trained on hand-drawn faces, it detects them in G. One false detection is also present in G. Two faces are missed in B (one because of a large tilt, the other because of occlusion).

Table 1: Text Detection Accuracy

Data Set	Number of Images	Text Regions Detected	Text Regions Missed	False Detections
CNN News 1	20	11	1	4
CNN News 2	23	7	0	3
Species Extinction	20	12	1	0

text position will be assisted by the introduction of optical character recognition.

In Figure 6, we present 5 poster frames which summarize 108 seconds of video. Using segmentation only, we select these poster frames from equally spaced scene breaks (see Figure 6B). The first frame is the most informative; it reveals that the story is about "Aetna" and "The Travelers". Poster frame 2 provides no information; the text in the image is obscured by motion blur. Frame 3 contains repeats that the story is about Aetna. Frames 4 and 5 are pictures of buildings, which do not provide additional information.

The system which uses face and text detection reveals more about the story (see Figure 6C). In frames 2 and 3, a location related to the story, Hartford, Connecticut, is given. Frame 4 provides hints about the story's topic, by showing an "insurance analyst" speaking. Finally, in frame 5, we see that the story ends in New York.

In only five frames, neither system reveals the full story, which is about the merger between the insurance companies Aetna and The Travelers, and the effect it will have on their stock prices. Nonetheless, a system based on scene segmentation alone wastes 4 of the 5 poster frames by providing no information beyond what was present in the original frame. Our system wastes only one frame (either 2 or 3), by providing repetitive information.

Figure 7 gives similar results for a 176 second video sequence. Because of the segment length, we allow each algorithm 7 poster frames. Both methods use the same initial frame; the frame indicates that the segment will involve "talks", Northern Ireland, and Great Britain (if one recognizes the flag). Scene segmentation's second frame shows a meeting, the third frame shows an object which is hard to recognize without more context. Frames 4 and 5 show banners related to the story. The last two frames selected by scene segmentation show unidentified people.

The results using text and face detection provide more

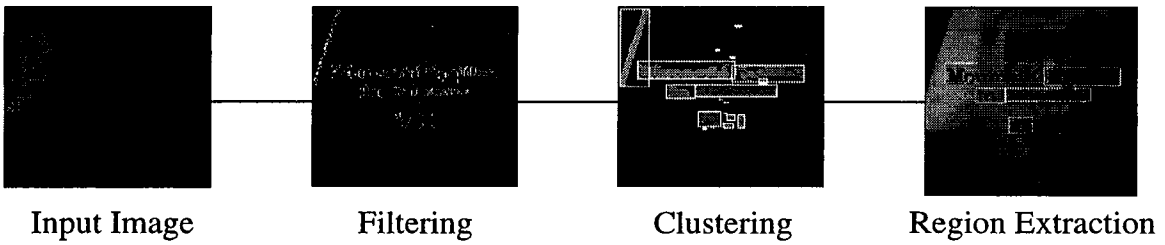

Figure 3: The basic algorithm used for text detection.

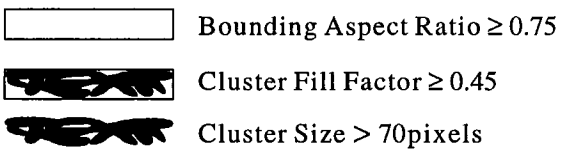

Figure 4: Constraints on the text areas.

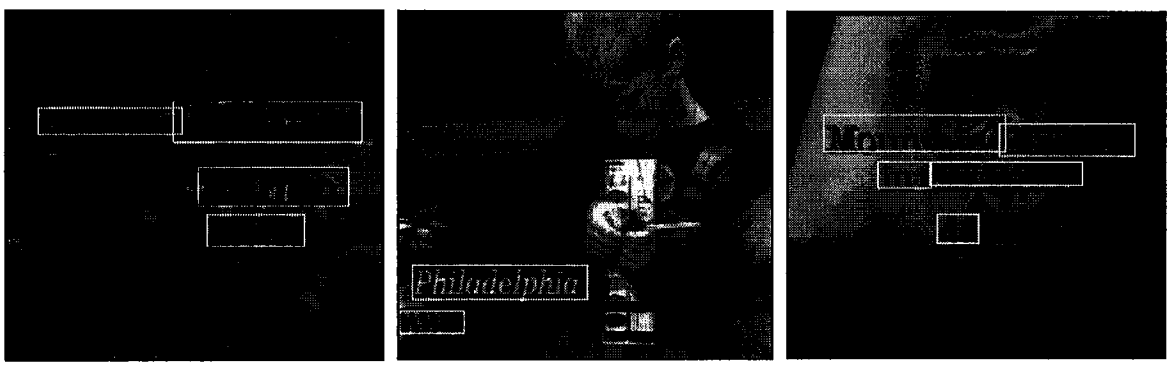

Figure 5: The output of the text detection algorithm.

specific information (Figure 7C). Frame 2 shows the location of the story, in Belfast, Northern Ireland. Frame three is wasted, due to an incorrect detection of a face. Frames 4, 5, and 6 show (and identify) key figures in the story. The last frame identifies the reporter and his location in Belfast.

Conclusions and Future Work

Text and face detection can support other types of analysis of video sequences. For instance, optical character recognition applied to the detected text can be used to support keyword-based searches. Current work by Sato (Sato 1996) explores automatically associating names with faces. The names are extracted from closed caption text, and their co-occurrence with face images are measured. Since names do not always appear in the closed caption text, other sources such as on-screen text can be used. The co-occurrence measurements allow us to recognize people who often appear in news broadcasts. An immediate application of Sato's work is to disambiguate anchor-people whose faces often appear with many names in audio and closed caption text. Repetitive shots in which only the anchor-person appears can therefore be eliminated from video summaries.

This paper has described some initial steps towards using scene content for selection of poster frames. Empirically, we have found that using face and text detection to prioritize frames provides more informative poster frames. Faces are often useful for showing the subject of a news story. Text often reveals information about the subject's name and affiliation, the locations of reported events, or keywords providing context for the story.

References

Arman, F.; Depommier, R.; Hsu, A.; and Chui, M.-Y. 1994. Content-based browsing of video sequences. In *ACM Multime-*

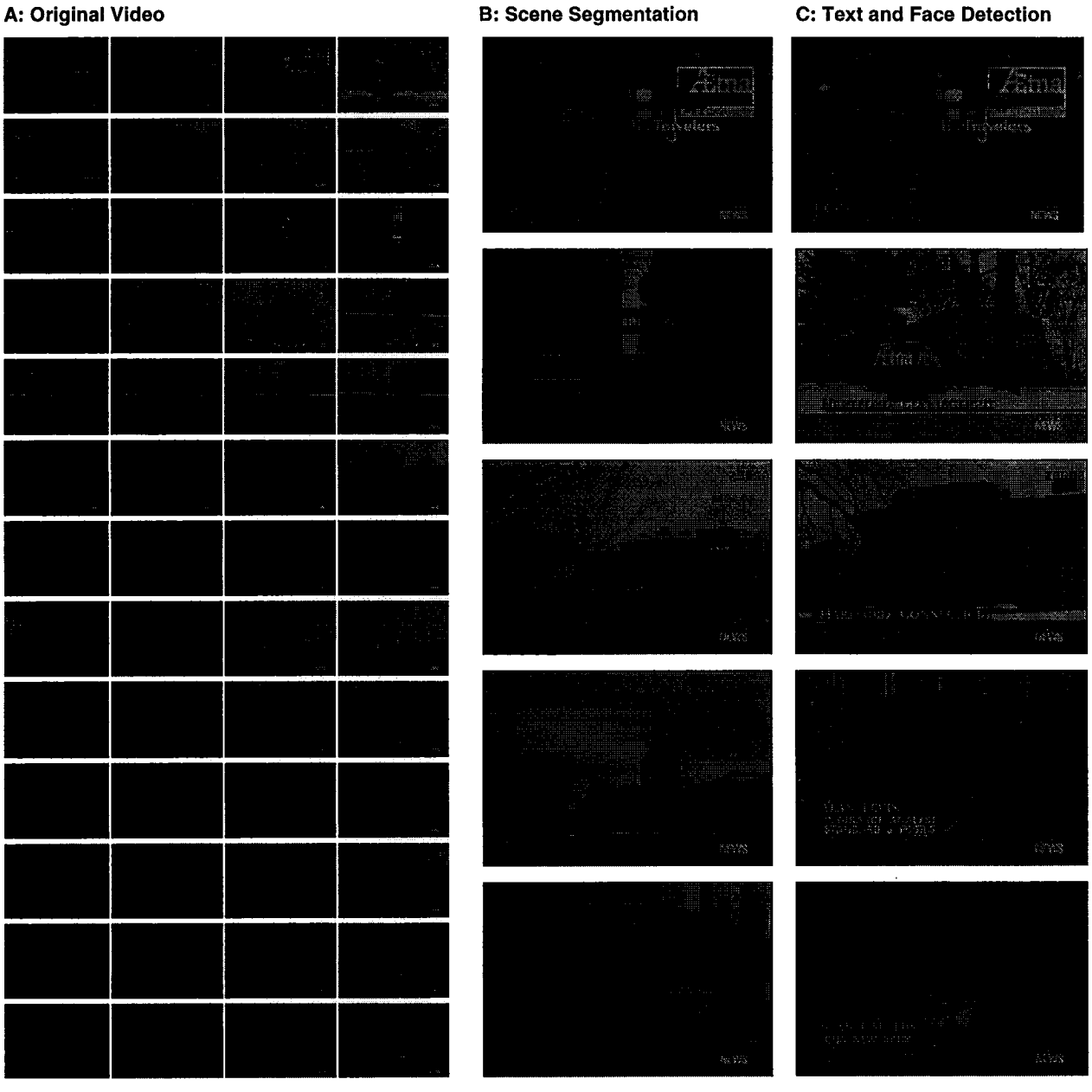

Figure 6: Selection of poster frames in a video sequence: A) Sample frames from the video sequence. B) Poster frames selected by the scene segmentation algorithm. C) Poster frames selected by an algorithm using face and text detection to prioritize frames.

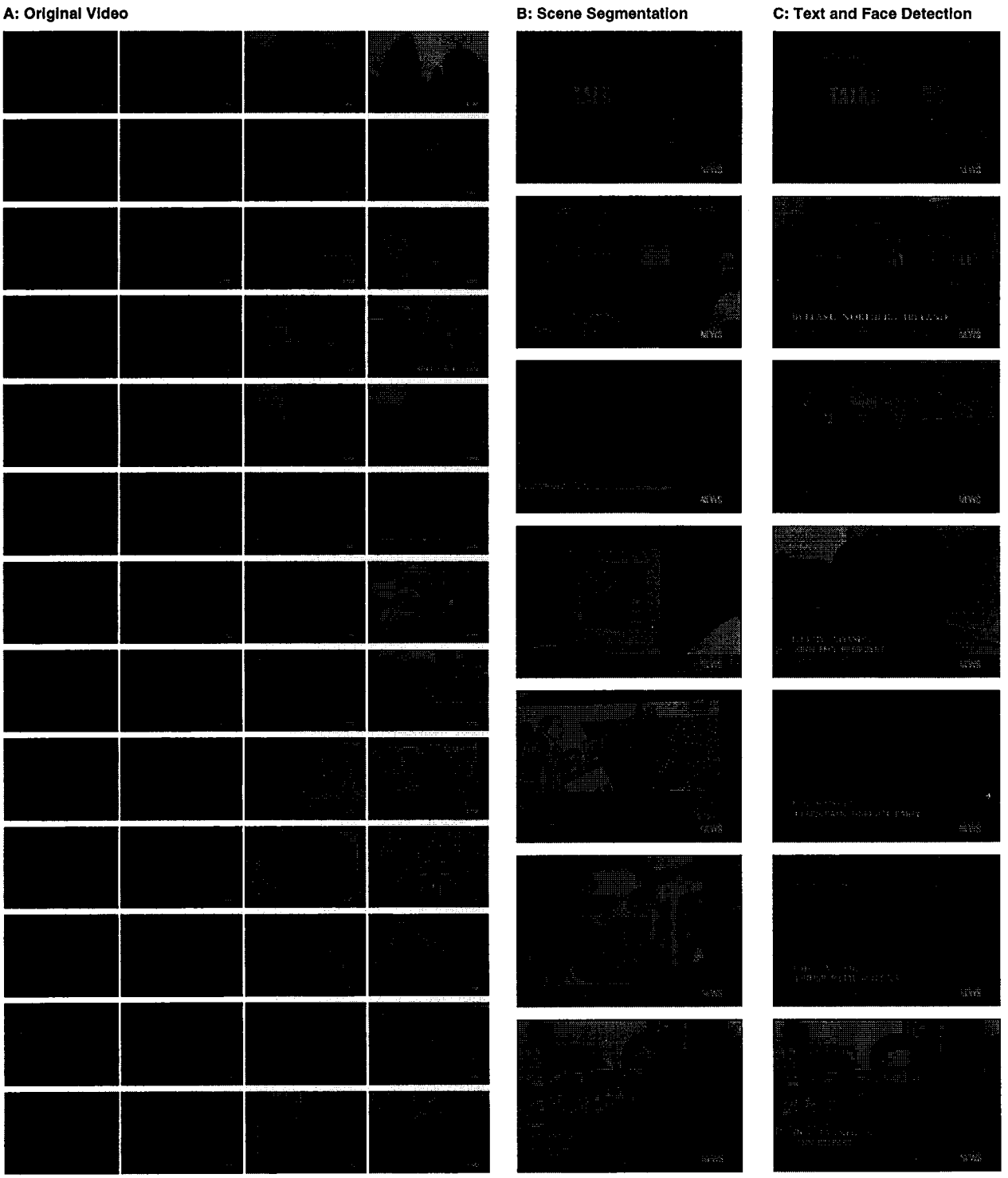

Figure 7: Selection of poster frames in a video sequence: A) Sample frames from the video sequence. B) Poster frames selected by the scene segmentation algorithm. C) Poster frames selected by an algorithm using face and text detection to prioritize frames.

dia, 97–103.

Burl, M. C., and Perona, P. 1996. Recognition of planar object classes. In *Computer Vision and Pattern Recogition.*

Hampapur, A.; Jain, R.; and Weymouth, T. 1995. Production model based digital video segmentation. *Multimedia Tools and Applications* 1:9–46.

Mills, M.; Cohen, J.; and Wong, Y. Y. 1992. A magnifier tool for video data. In *ACM CHI'92 Conference on Human Factors in Computing Systems*, 93–98.

Rowley, H. A.; Baluja, S.; and Kanade, T. 1996. Neural network-based face detection. In *Computer Vision and Pattern Recogition.*

Sato, S. 1996. Personal communication.

Smith, M., and Kanade, T. 1996. Skimming for quick browsing based on audio and image characterization. Technical Report CMU-CS-96-186R, Carnegie Mellon University.

Sung, K.-K., and Poggio, T. 1994. Example-based learning for view-based human face detection. Technical Report A.I. Memo 1521, CBCL Paper 112, MIT.

Vaillant, R.; Monrocq, C.; and Le Cun, Y. 1994. Original approach for the localisation of objects in images. *IEE Proceedings on Vision, Image, and Signal Processing* 141(4).

Yang, G., and Huang, T. S. 1994. Human face detection in a complex background. *Pattern Recognition* 27(1):53–63.

Zhang, H.; Kankanhalli, A.; and Smoliar, S. 1993. Automatic partitioning of full-motion video. *Multimedia Systems* 1:10–28.