Browsing News and Talk Video on a Consumer Electronics Platform Using Face Detection

Kadir A. Peker, Ajay Divakaran, Tom Lanning

Mitsubishi Electric Research Laboratories, Cambridge, MA, USA
{peker,ajayd,}@merl.com

Abstract. We present a consumer video browsing system that enables use of multiple alternative summaries in a simple and effective user interface suitable for consumer electronics platforms. We present a news and talk video segmentation and summary generation technique for this platform. We use face detection on consumer video, and use simple face features such as face count, size, and x-location to classify video segments. More specifically, we cluster 1-face segments using face sizes and x-locations. We observe that different scenes such as anchorperson, outdoor correspondent, weather report, etc. form separate clusters. We then apply temporal morphological filtering on the label streams to obtain alternative summary streams for smooth summaries and effective browsing through stories. We also apply our technique to talk show video to generate separate summaries of monologue segments and guest interviews.

1. Introduction

Personal video recorders (PVR) enable digital recording of several days' worth of broadcast video on a hard disk device. Several user and market studies confirm that this technology has the potential to profoundly change the TV viewing habits. Effective browsing and summarization technologies are deemed crucial to realize the full potential of these systems.

News video story segmentation is a well studied subject. A recent TRECVID session on this topic showed that close to product performances could be achieved combining audio, visual, and text features [1,2].

Our focus in this work is driven by the constraints of the consumer electronics platforms, and the requirements and flexibilities of TV viewing application. We have developed audio classification based summarization solutions for sports video in our past work [3]. We now extend our target genres by including face detection technology.

The platform limitations and the generality of broadcast video for PVRs suggest using features that provide maximal application range with the minimum cost. We find faces as the most important visual class that will enable analysis of a wide array of video types, as the humans are mostly the primary subject of video programs. We use the Viola-Jones face detector, which provides high accuracy and high speed [4]. It can also easily accommodate detection of other objects by changing the parameter file used. Thus, the same detection engine can be used to detect several classes of objects. The parameter files on the consumer device can even be updated remotely.

A number of application characteristics define our emphasis points, which may differ from the criteria used in the TRECVID news story segmentation task:

1. Flexibility of user interaction in TV viewing:

Some of the over-segmentation is tolerable – or even desired, for instance when the story consists of a few interviews or reports, and snippets from each is included in the summary.

2. Limited processing power of the CE platform:

The application is run on a consumer electronics platform, as opposed to a general purpose PC. The resources are much more limited. The types of video are very varied as well, and not specialized. The most return on limited investment is desired.

3. Relatively smaller data size:

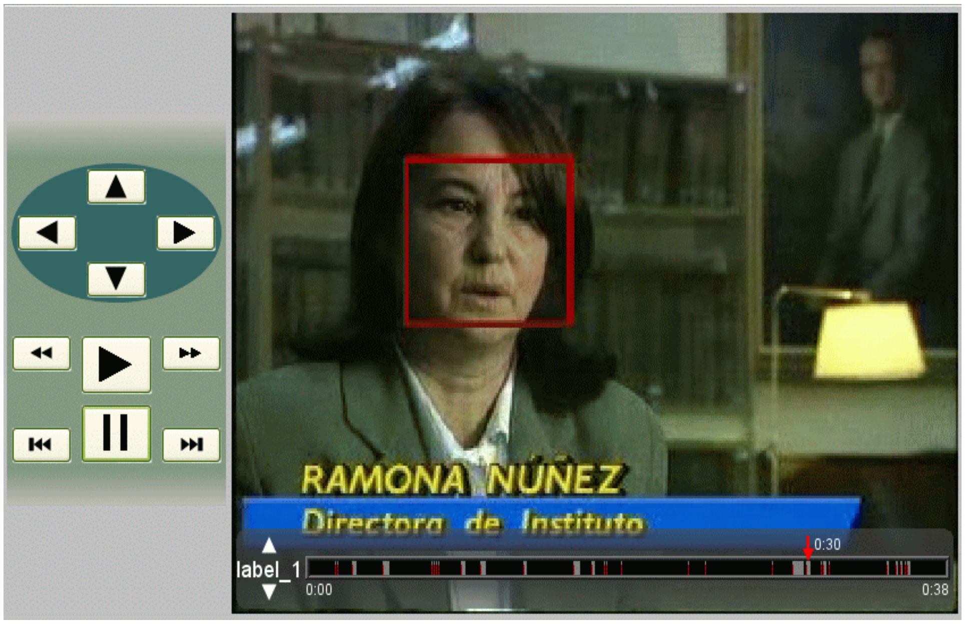


Figure 1. A screenshot of the consumer electronics video browsing interface. The user can browse through alternative summaries using up-down keys, and skip through segments or markers using the left-right keys. A time bar at the bottom shows the included segments, and the current position.

The target video is several news video programs that a user can reasonably browse in one sitting; as opposed to a professional, dedicated analysis of large volume news broadcasts from numerous sources.

4. The environment does not allow complicated interactions, and the user is not a professional.

The TRECVID results show that anchor person is the most significant feature for story segmentation [1]. In fact, a perfect detection would yield a performance that is better than 6 (out of 8) of the contributors who use several audio, video, and text features together. Face detection is the most common way of detecting anchor person shots [2].

Furthermore, face detection is applicable in analysis, segmentation, and browsing of several other types of video. In this work, we report our work on segmentation and browsing of video that primarily consist of static shots of talking people. News video is the primary example. We also show examples from talk show and documentary/interview programs.

2. PVR video browsing

We developed a user interface that allows selection of several alternative summaries for a video program. The ability to use multiple browsing alternatives, such as viewing the whole program, or the story introductions, or the weather report, etc., or summaries of different lengths, is provided in a clean and intuitive interface. The user can flip through alternative summaries in one dimension, and jump forward or backward through segments in the other dimension. The user can watch the summary without interaction, or skip to next segment at any point in a segment. Each summary can consist of either a set of selected segments or a set of markers in the whole video program. With the markers option, the device plays the whole program if not interrupted. The user can skip to the next marker point, e.g. next story start, using the skip buttons. Figure 1 shows a snapshot of the interface. Note that, this kind of an interface allows easy recovery from segmentation errors, where the user can quickly skip over false alarms, as long as false alarms are not too frequent.

3. Face detection in consumer video

We use the Viola-Jones face detector [4], which is based on boosted rectangular image features. We reduce frames to 360x240 pixels, and run the detection on 1 pixel shifts. The speed is about 15fps at these settings on a Pentium 4 3GHz PC, including decoding and display overheads. About 1 false alarm per 30-60 frames occur with frontal face detector. Using DC images increases the speed dramatically, both through the detector (speed proportional with number of pixels), and through savings in decoding. The minimum detected face size increases in this case, but the target faces in news video is mostly within the range. The detector can be run on only the I-frames, or at a temporally sub-sampled rate appropriate for the processing power.

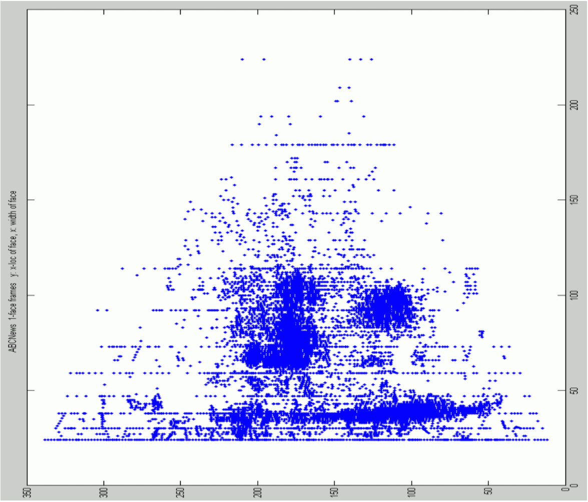
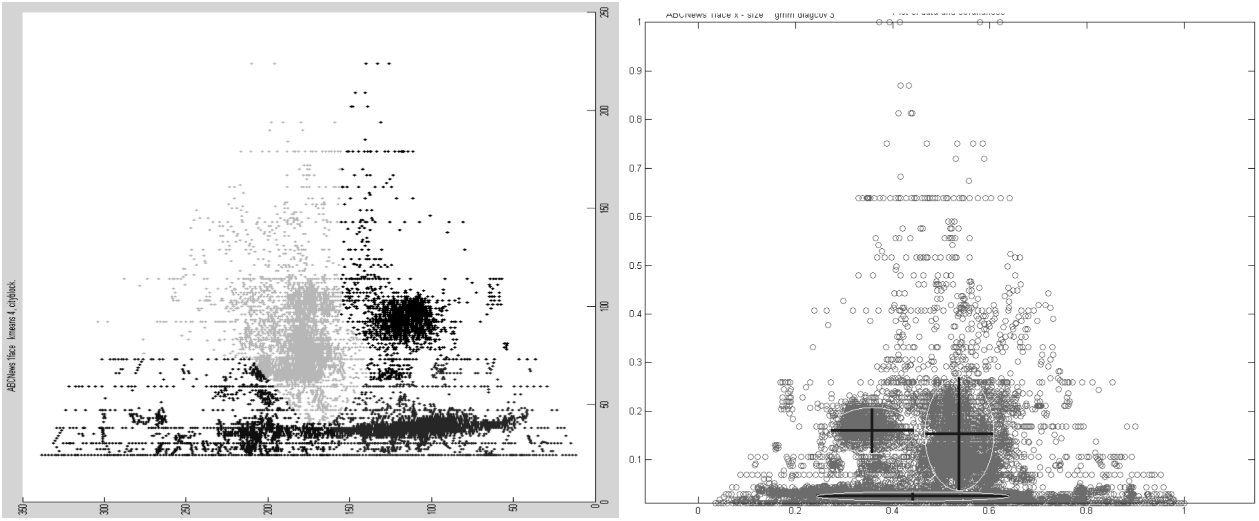


Figure 2. Frames with 1 one faces. Scatter plot of face x-location vs. face size.

4. Clustering using face x-location and size

We first classify video frames (or larger units, depending on the temporal resolution chosen) based on the number of faces detected, into 1-face, 2-face, and 3-and-up classes. In news video and other static-scene talk video such as talk shows and interviews, most of the segments have 1 face. We further classify 1-face segments based on the scene composition. We found that face size and x-location is an effective feature for discriminating between different types of video scenes in our target video genres. Figure 2 illustrates the natural clustering of 1-face video frames in a broadcast news program, using face x-location and size.

We use k-means clustering for its low complexity and wide availability, with 3-5 clusters. Although the clusters found in this way can be poorly aligned with the natural boundaries observed in the scatter plots (see figure 3), the shortcomings are mostly handled in later stages when we temporally smooth the segments.



(a) k-means clusters of the data in figure 2

(b) Ellipsoids showing the GMM clusters.

Figure 3. Clustering of face scenes using k-means and GMMs.

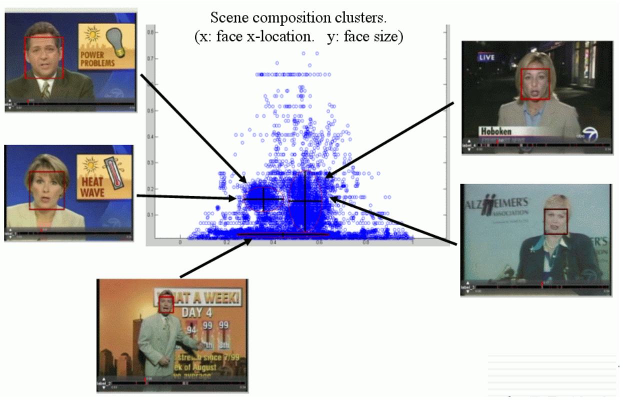


Figure 4. The scenes that each of the clusters illustrated in figure 3 correspond to. The clustering results reflect the editing and camera style of the particular program.

We also experimented with GMMs for clustering which give smoother cluster boundaries and more intuitive looking clusters. However, it is not clear if the advantage is significant enough to have an effect on the final summarization results. Figure 3 illustrates both k-means and GMM clustering of the data in figure 2.

Our experiments with several news video programs and talk show programs indicate that, clustering 1-face frames using face size and x-location gives semantically meaningful classification of video segments into scenes. Figure 4 shows samples from a news video program, where one of the clusters corresponds to anchorperson shots, another cluster to outside correspondents, and another cluster to the weather report.

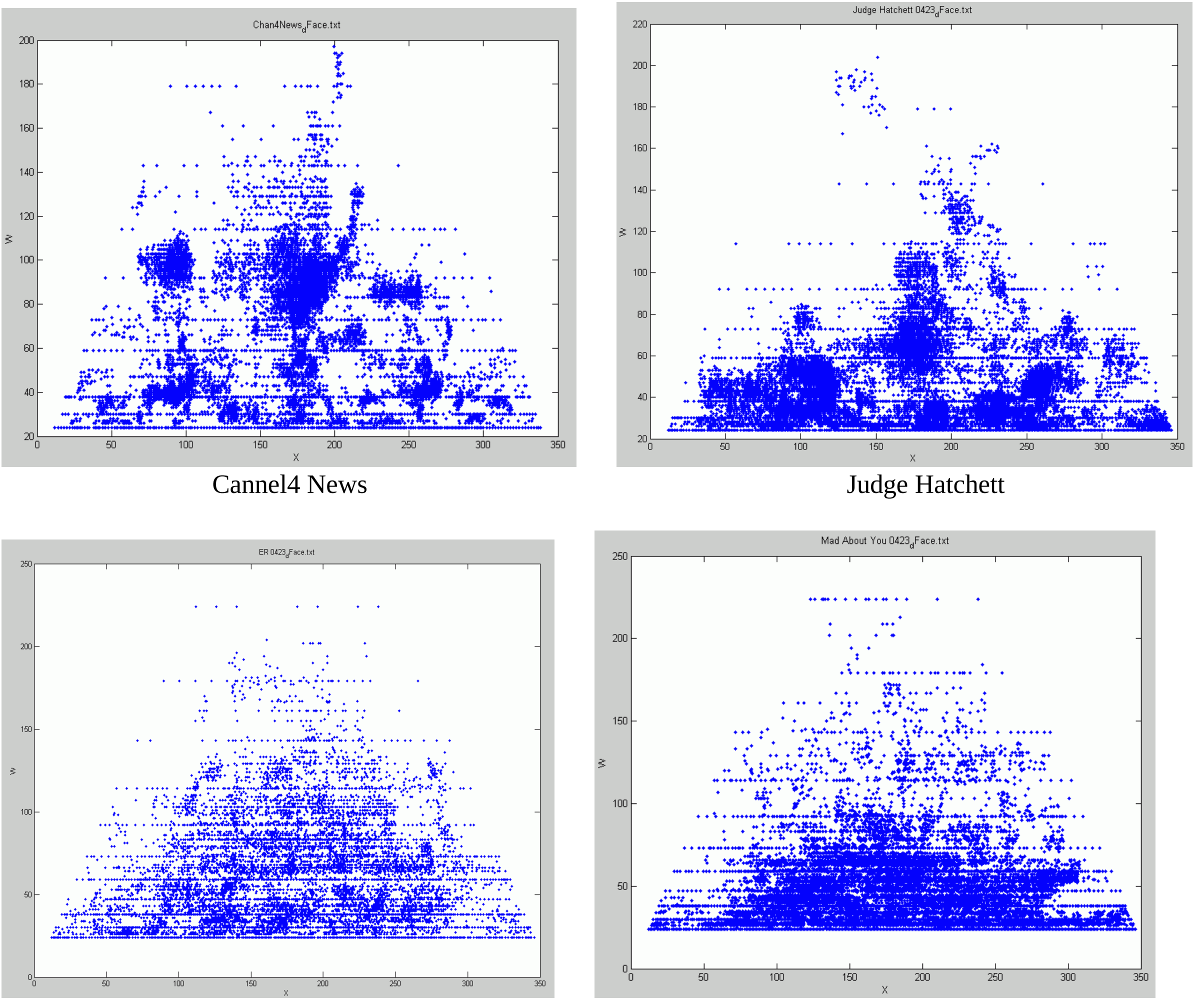
4.1. Limitations of clustering-based scene classification

Note that the face location and size features correspond to different types of scene compositions, hence semantically meaningful scene segmentations, only for static video editing styles such as news, talk shows, interviews, etc. Clustering using these features is not effective for other types of video where the camera angle, position, and focal length changes over a wide range of settings, and the actors move around the scene frequently. It is best suited for video shot in studio settings with a limited set of view compositions (e.g. close-ups of speakers, wide view of scene, etc.). Figure 5 illustrates the natural clustering of data for sample news video data, and a courtroom video (Judge Hatchett), which has above mentioned static camera properties. Two sample drama programs (ER, Mad About You) illustrate that there are no clear clusters of face location and size in these type of dynamic scene content.

The exact discriminative power of face size – x-location clustering depends on the style of the particular program. In some news programs the framing of the anchorperson face is not constant. In some programs the framing is similar for both anchorperson shots and outdoors interview shots. So, for a perfect separation of anchorperson shots and other major talking head shots, we need to use other features such as color histograms. However, even when we use face size – x-location clustering, the resulting segments are meaningful marker points for browsing the video. We have discussed the application parameters for consumer video browsing on consumer electronics devices, and how it differs from other application contexts in more detail in the introduction section.

5. Temporal smoothing

After the clustering, all the video frames (or segments) are labeled with a cluster number, or with ‘2-faces’, or ‘3-or-more’ labels. Face detection errors cause the results to be very fragmented. In some cases a single scene on the border of a cluster falls into multiple clusters, also causing fragments. This raw segmentation is not appropriate for browsing using the described user interface: most of the segments are very short, resulting in jerky playback. Skipping to the next segment, most of the time, will advance the playback only a few seconds or less.



Cannel4 News

Judge Hatchett

ER

Mad About You

Figure 5. Face-X vs Face-Width plots for different programs. Top row (news, courtroom) programs have a limited set of scene compositions with mostly fixed camera shots; the actors have fixed positions in the scene. The second row (ER, Mad About You) programs have dynamic scenes with changing camera positions and lengths, and moving actors. Clustering of face location and size is useful for the first type of programs such as news, talk shows, interviews, etc.

To alleviate this problem, we first correct face detection errors using temporal coherence. We use a running window-based tracking where false detections are removed and gaps in tracks are filled. Tracks shorter than a threshold are later removed.

At the second level, we temporally smooth the segmentation results. We treat each label (e.g. cluster1, cluster2, …, 2-face, etc) as a separate summary. Then we apply morphological smoothing to each of the separate summaries, which removes short gaps as well as short segments below a certain threshold. In our experiments, thresholds of 1 to 3 seconds give reasonable results. Note that, after this process, the labels are no longer mutually exclusive.

6. Browsing news and other talk video

The user can watch each label as a separate summary. One of the clusters usually corresponds to anchorperson segments. Anchorperson segments following another type of segment, in turn, usually indicate story introduction. Thus, in the cluster that corresponds to the anchorperson, the user can watch the whole summary, which goes through the introductions of the stories without the details that usually come from outside footage. Or the user can skip to the next segment at anytime, which is the start of the next story. In one broadcast news program that we have annotated, we were able to detect 11 story introductions and miss only 2.

We experimented with other talk video content with static scenes, such as talk shows and interview programs. We are able to separate the monologue segments from the guest segments. Thus a user can either watch the jokes in the monologue or skip to the guests. We also observed that a good way of finding out the guests at a program is by using the 2-face segments, which usually correspond to the host introducing a guest.

The separate summaries (labels) can also be merged to generate a single, or a small number of summaries. We are currently studying effective ways of merging these summaries. One strategy is discarding the clusters that have high variance. One of the clusters in our experiments had small face size and relatively spread out x-locations. This usually corresponds to the weather report. So, this cluster, although it may have a high variance, is preserved. Outliers in other clusters are also eliminated, leaving more compact cores. The remaining clusters are temporally smoothed, and then merged in a single summary. Markers are inserted at the boundary points where the label changes. This way, even if the playback continues through a full story, the user can still have markers to skip to different segments of the story. The final summary is temporally smoothed again to remove gaps that may result from merging.

7. Conclusion

We presented a consumer video browsing system that accommodates browsing through recorded video programs using multiple alternative summaries. The interface is simple and effective, suitable for the consumer electronics application. We presented a segmentation and summary generation technique for news video and other talk programs with static scenes. We first classify video segments into 1-face, 2-face, and more-faces classes. Then 1-face segments are furthered clustered using face size and x-location. Each video segment is labeled through this process. Each label sequence is individually smoothed to remove gaps and short segments. The user can browse through the video using alternative label streams. Our experiments with several news video programs indicate that the labels usually correspond to different semantic scenes such as anchorperson, outdoor correspondent, weather report, etc. Even when there are multiple semantic classes represented in one label, they are still meaningful for browsing purposes. For example, a label may contain a few interview segments in addition to anchorperson shots. A summary that contains these segments in addition to story introductions is still appropriate for browsing.

Face detection by itself enables effective ways of segmenting consumer video. It is one of the most cost effective visual features for consumer video domain. The results presented here can be improved further by using other visual and audio features, depending on the constraints of the target platform.

10. References

1. T.S. Chua, S.F. Chang, L. Chaisorn, W. Hsu, “Story Boundary Detection in Large Broadcast Video Archives – Techniques, Experience and Trends,” ACM Multimedia Conference, 2004.
2. TREC Video Retrieval Evaluation (2003) http://www-nlpir.nist.gov/projects/tv2003/tv2003.html. Nov 2003.
3. Divakaran, A.; Peker, K.A.; Radharkishnan, R.; Xiong, Z.; Cabasson, R., "Video Summarization Using MPEG-7 Motion Activity and Audio Descriptors", Video Mining, Rosenfeld, A.; Doermann, D.; DeMenthon, D., October 2003 Kluwer Academic Publishers.
4. P. Viola and M. Jones, “Robust real-time object detection,” IEEE Workshop on Statistical and Computational Theories of Vision. 2001.