IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Incorporating Domain Knowledge with Video and Voice Data Analysis in News Broadcasts

Kim Shearer [a] Chitra Dorai [b]
Svetha Venkatesh [c]

IDIAP–RR 00-14

August 2000

published in
Multimedia Data Mining workshop, part of IEEE KDD-2000

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP
[b] IBM T. J. Watson Research Center
[c] School of Computing

IDIAP Research Report 00-14

Incorporating Domain Knowledge with Video and Voice Data Analysis in News Broadcasts

Kim Shearer Chitra Dorai Svetha Venkatesh

August 2000

published in

Multimedia Data Mining workshop, part of IEEE KDD-2000

Abstract. This paper addresses the area of video annotation, indexing and retrieval, and shows how a set of tools can be employed, along with domain knowledge, to detect narrative structure in broadcast news. The initial structure is detected using low-level audio visual processing in conjunction with domain knowledge. Higher level processing may then utilize the initial structure detected to direct processing to improve and extend the initial classification.

The structure detected breaks a news broadcast into segments, each of which contains a single topic of discussion. Further the segments are labeled as; anchor person or reporter, footage with a voice over or sound bite. This labeling may be used to provide a summary, for example by presenting a thumbnail for each reporter present in a section of the video. The inclusion of domain knowledge in computation allows more directed application of high level processing, giving much greater efficiency of effort expended. This allows valid deductions to be made about structure and semantics of the contents of a news video stream, as demonstrated by our experiments on CNN news broadcasts.

1 Introduction

Research into image databases and image indexing and retrieval has led to the creation of a number of useful tools for similarity retrieval for images [6, 9, 4, 16]. Application of these tools to video is possible, but the principles embodied in the tools do not yield a useful query system. Previous work on video indexing and retrieval [22, 10, 20, 23, 3, 9] has most commonly relied largely on one aspect of video, be it vision or sound, and has been restricted to low-level or undirected processing. The results of this processing are then used for classification, with the goal of detecting either video events, or some form of structure within video. Detection of events or structure permits a summary of the video to be formed, thus permitting more rapid user browsing by a restriction of the information or segments presented for browsing.

Examples of the form of summaries are the Video Icons of Tonomura and Abe [18, 19], the excellent work by Davis [5] on MediaStreams, general systems such as [14, 8, 17] and the scene transition graphs of Yeo and Yeung [21, 2]. These methods aim at presentation of video content in a condensed manner so that the extreme amount of information available may be scanned by the user in a more efficient manner. The scene transition graphs of Yeo and Yeung go slightly further than most earlier work in that they present a possibility for automated deduction of semantically related structure from a video stream.

In this paper we describe a collection of tools and their application to the detection of narrative structure in a news broadcast. In particular, these tools are used to break the broadcast into segments, each of which contains a single topic of discussion. These segments are classified further by labeling each individual shot as one of

- anchor person or reporter,
- footage with a voice over,
- sound bite,

which gives a clear indication of structure within the video. This work differs from earlier work in that it employs not only low-level processing, but uses results from this processing, along with initial deductions about structure within video, to apply higher level processing in a directed manner. This allows a novel iterative approach to be used, with alternating processing and deduction employing progressively more complex computation as the interpretations become more finely focused. The summary produced from this work can then go further than simply presenting a representative sampling of video, by providing a summary based on the semantics of the content.

The aim of this work is to allow automated annotation of video, which will allow intelligent construction of summaries for large video databases. The particular target area is news broadcast and news magazine footage, such as that kept by major news companies. The annotation created will break the video into segments of homogeneous topic, and further label shots as anchor or report footage. A typical summary that might then be created would be a thumbnail of each anchor person or reporter present in a section of video. The user may then select the reporter who filed a story, rather than having to search for a representative frame which might be contained in the story required. Given the large volume of video data retained for such applications, and the volume captured at each moment, this could result in a large reduction in unproductive human time and lead to a scalable and efficient solution for content management in studios.

2 Components as tools

A number of components may be employed in the analysis of video streams. These components are employed to assess similarity of shots within the video stream, along a number of axes. This similarity within the video stream is then used with a knowledge of shot syntax, and higher level processing, to deduce structure within the news video stream.

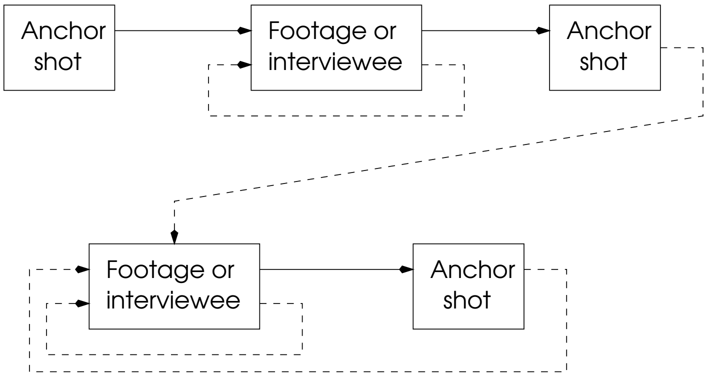


Figure 1: Shot syntax of a broadcast news program.

	111	112	113
112	5886		
113	25759	25559	
114	7839	4681	25544

Table 1: Similarity measure using CCV for the video frames shown in Figure 3.

2.1 Detection of Anchor Segments

The concept of shot syntax was developed to describe the regular structure of camera parameters employed to capture a particular type of semantic content [2, 21]. The clearest example of regular shot syntax is in interviews. In an interview video it is generally the case that the interview will be introduced by the interviewer. There will then usually be either a shot of the interviewer and the interviewee, or a shot of the interviewee alone. Subsequent shots will be of either; interviewer, interviewee, a mid-range shot of the two people involved, or background footage. This repetitive structure is adopted for interviews as it has been found to be the best method of producing this type of program.

If the assumption can be made that such repeated structure will be present within a video stream of a particular program genre, then detection of repetition in shot settings provides a useful first pass for the grouping of shots into meaningful segments. News broadcast does in general adhere to such a structure, as shown in Figure 1. In this figure solid lines indicate required minimum paths through the syntax diagram, with dashed lines denoting optional paths. The regular structure displayed makes it useful to search for repetitions of anchor or reporter segments. That is, shots with one person addressing the camera, and this person presenting a particular segment of the program, therefore, appearing repeatedly. The term anchor shot will be used to refer to this type of shot, whether it is a shot of an actual anchor person, or a shot of a reporter who is the presenter for a particular story. A story presented (or anchored) by a reporter in the field generally represents a self contained sub–syntax of a larger report. Figure 2 shows a possible syntax for such a segment, the field report presented by a reporter is contained within the dashed line box. The shot syntax for this report is clearly similar to the syntax for a general report.

In our news video processing system, the search for anchor shots takes advantage of a property inherent to such shots. Anchor shots are intended to provide continuity for a news broadcast, which means that the intent of such shots is to present a consistent appearance to viewers. Therefore such shots are captured in a consistent location, with mostly consistent shot parameters. This visual consistency makes detection of repetitions of the anchor simple to detect. Reporters in the field also usually present a highly consistent appearance, however, this is less dependable due to outside factors.

Initially colour coherence vectors (CCVs) [11, 8] were used to detect similarity between frames

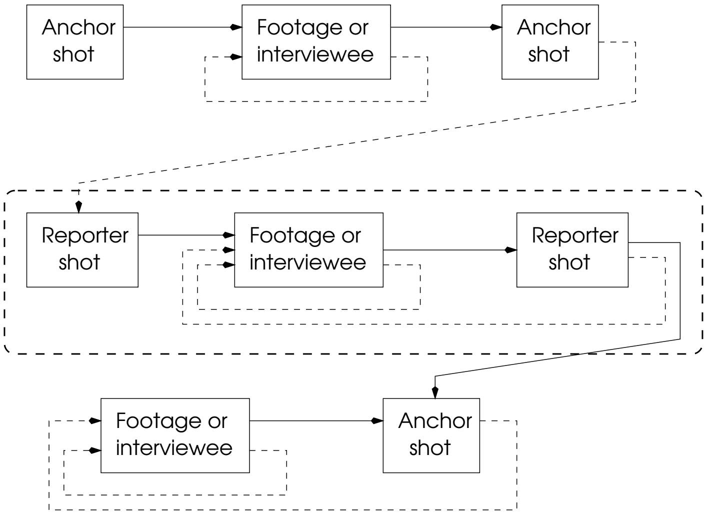


Figure 2: Typical syntax of a news program with a field report.

(a) Frame 111

(b) Frame 112

(c) Frame 113

(d) Frame 114

Figure 3: Facial rotation for which CCV performs poorly.

	111	112	113
112	71112		
113	71410	5374	
114	70844	8220	5454

Table 2: Similarity measure using spatial histograms for the video frames shown in Figure 3.

sampled from a video to indicate anchor sections. However, CCVs perform poorly with a number of scenarios that occur frequently in news video. The main problem occurs with faces which dominate the frame, and rotate under studio lighting. In these cases the coherence of the colour regions can change dramatically for a small movement. This situation often occurs in anchor shots, where a reporter glances down at a page of notes, or to the left or right to pass to an interviewee or other reporter.

Simple colour histograms provide a useful indication of similarity, but as expected find too many shots to be similar. Using such a global measure allows too many frames of similar colour to be clustered as similar, and will also find frames within a shot that has a great deal of motion similar. For the task of separating anchor and reporter shots from other shots, it is acceptable that motion in the shot, such as the motion apparent in crowd scenes, cause frames to be found dissimilar. The goal is then that each anchor or reporter shot be found coherent (internally similar) and similar to other shots of the same reporter or anchor.

As a result a different similarity measure was employed in our system, where each frame is broken into 12 subframes, and a colour histogram is computed for each. Each histogram is quantized to 16 bins, and histogram difference Δ is a sum of the differences between values for each bin i. That is

$$\Delta = \sum_{i=0}^{15} |H_1[i] - H_2[i]| \tag{1}$$

The histograms for spatially corresponding subframes are then compared, with the sum of the histogram differences for the subframes representing the distance between frames. The similarity values for the video frames in Figure 3 are given in tables 1 and 2. As can be seen from Table 1, the CCV algorithm finds that frame 111 is far more similar to frame 112 than frame 112 is to frame 113, and also that frame 114 is similar to frames 111 and 112 but not 113. This is due to the changes in colour values for the face and hair of the pictured person in frame 113 as the head tilts slightly. The size of areas containing a particular colour change dramatically with only small head movements. For the same four frames the histogram measure performs much more as expected, easily separating the frames correctly.

In addition to addressing the problem illustrated in Figure 3 the algorithm we employed has another useful property. While each shot of an anchor person or reporter is found to be coherent, most other shots are not. This is due to the sensitivity of the algorithm to overall fluctuations in colour and position of colour. Scenes which might seem likely to be found similar under a colour based measure, such as shots of a crowd, are in fact separated into numerous short pieces. This has the advantage of reducing the number of shots that are detected as repeated shots within a video stream, thus making the task of shot syntax analysis simpler.

There are of course other shots which will be repeated during a broadcast, such as the logo of the news station, advertisements which are repeated and footage used as a preview for stories in later programs. One tool which is often useful in distinguishing these shots from anchor shots is face detection. While face detection is only reliable in constrained applications, it is suitable for this problem. A search for faces in anchor shots will be assisted by the regular presentation of these shots, while advertisements are generally quite erratic and have few static, and therefore detectable, faces.

The face detection part of classification is performed using the CMU face detection software [13]. This is a neural network based face detector, in which neural networks are applied directly to each 20 by 20 pixel location in the image. In order to accommodate scaling transformations the image is presented to the system at actual size, and then repeatedly scaled down by a factor of 1.2 and again presented to the system. Training is accomplished on a set of face images, and non-face images, with false positives in the non-face images being used as negative examples in further training. A number of heuristics are used both to improve accuracy and to improve speed. This system is chosen as representative of the current state of the art in face detection, and its performance is easily sufficient for the given task.

Anchor shots exhibit the following properties which make face detection more reliable:

- the face is turned directly towards the camera,
- the face dominates the shot.

Face detection can therefore be restricted to searching for large faces. The majority of false detections that are artifacts of other parts of the image are small relative to the faces in anchor shots, so size can be used as an effective filter. Searching for only those faces which directly face the camera also simplifies the problem, further reducing the error rate.

Shots that repeat with a suitable shot syntax and have a consistently visible face are highly likely to be anchor shots. The assumption of temporal consistency can be used to further reduce error from face recognition by discarding faces that move rapidly or erratically. This will tend to discard footage of people addressing a crowd, but include field reporters. Reporters in the field will be less static than anchors in the studio, but all field reports in the data set tested were detected as dominant faces. Temporal consistency can also be applied to the colour histogram work by using an average histogram for each group of frames which are considered similar, to represent the matching attribute set. This limits the spread of a single group by preventing a chain of frames with small error from each other remaining part of a single group even though the error diverges further and further from a previous group.

Once these two steps of visual processing have been completed a first pass is performed to determine structure from shot syntax. This yields a preliminary label for each shot as either an anchor shot, or a non-anchor shot. To label the shots in finer detail the sound associated with the video is processed. This presents a difficult problem, as there is no simple method to ensure clean audio samples. While voice recognition in an environment for which extensive training samples are available, and voice samples are well separated can show good performance, this is not the case for this application.

2.2 Audio Analysis

To label the shots in finer detail, the audio associated with the video is analyzed. Much of the sound from news broadcast will contain noise of various forms, such as background noise for field reports. In addition, there are a number of behaviours presented by anchor people, which aid in keeping the flow of dialogue, that prevent clean segmentation of sound samples. One example is that the anchor person will often begin speaking before a field reporter or piece of footage has stopped, which aids flow but makes it impossible to separate one voice from another. In addition, the anchor will generally start speaking before the cut from one shot to another, or will start speaking just after the cut with sound from the previous segment continuing slightly past the cut. This means that most audio samples will contain multiple voices when segmentation of the audio stream is performed.

Previous work has suggested that four seconds is a suitable segment length for vocal samples to exhibit a consistent attribute profile [7], and this is the length employed in this work. Three methods of segmentation for sound were studied for comparison. Two methods attempt intelligent segmentation, the first using silence as an indicator for segmentation points and the second using cuts in the video. The final method employed was to simply cut the video every four seconds starting at the first frame. For each of the first two methods, sections longer than 4 seconds are cut into four second pieces, and segments shorter than 4 seconds are discarded. Segmentation based on silence detection performs significantly worse than either of the other methods, for reasons mentioned earlier. As there is little to choose between the performance of the two other methods, simple fixed time segmentation is used in our system for simplicity.

Audio classification is performed using formant frequency estimators [12, 15] and other low-level attributes as in [1], and k-means clustering. The most suitable number of clusters is chosen by minimizing total error, within a reasonable range. Thus at the end of audio processing, each four second audio segment is assigned an audio cluster label.

3 Fusion of component results

The three initial pieces of low-level processing are combined to determine the initial classification of shots as either anchor shot, voice over or sound bite using the following rules:

- Anchor shots will be repeated shots with a sequence of not more than 4 shots between, and a time between anchor shots of not more than 8 times the length of the anchor shot. They will also have a prominent face detected.
- Other shots will be initially classified as footage.
- Footage shots with vocal clustering similar to an anchor shot in the same grouping will be determined as voice over.
- Footage shots with vocal clustering dissimilar from any anchor shot in the initial grouping will be labeled as sound bite.

The first rule is also used to break the video stream into segments, with each segment containing a single story topic.

In practice the grouping of shots based on identification of anchor and reporter shots and duration between these shots detects 100% of the structure in the news video. The test set for this work contains two videos of approximately 50 minutes in length each, and includes a number of CNN news and magazine style programs. The structure detected represents a slight over segmentation, in that some reports have the anchor shot which introduces the segments, and the anchor shot concluding the segment discarded. This is due to the segment being anchored by a reporter, and thus exhibiting the shot syntax expected within the report (Figure 2), with the introductory and concluding segments being no more than a tie–in to the news program. It is deemed reasonable that these shots be discarded. The important feature of the segmentation is that no segment contains more than one topic, which could result in hiding of information from the user.

Table 3 gives a summary of the results from classification using the initial low-level processing and shot syntax. As can be seen, detection of shot syntax allows accurate classification of most of the video. The values in the accuracy column of Table 3 are calculated from the equation

$$\text{Accuracy} = \frac{Actual - F_{neg}}{Actual + F_{pos}} \tag{2}$$

where *Actual* is the correct number of samples for the shot type, and F_{neg} and F_{pos} are the number of false negatives and false positives for the classification. The majority of the misclassifications are due to too few sound samples being available for accurate audio classification of a shot. The false negatives for the anchor shots are due partly to the lead and trailing shots of a long report being dropped as discussed earlier, and also to one group discussion having two presenters. The anchor shots for this section are detected as similar, but have no single dominant face. Further processing discussed in later sections in this paper could be used to improve detection to include this case.

4 Directed application of high level processes

Given this initial segmentation of the shots within the video stream into structured blocks, further processing may now be considered. The main additional processing is a more detailed face detection pass applied to the shots classified as footage. This allows *interview* shots to be more accurately detected.

Allowing a greater range of sizes for a face increases both the time required, and the error rate for face detection. However, when footage is taken in the field it is less likely that an interview shot will show a dominant face front on. In this case greater care must be taken in assessing the results from

	Total number	False positives	False negatives	Accuracy
Anchor shots	44	4	6	79%
Voice Over	54	2	8	82%
Sound Bite	28	4	4	75%

Table 3: Classification results.

	Total	False positives	False negatives
Sound bite	4	8	0
Interview	24	0	8

Table 4: Interview shot detection.

the face detection algorithm. Results are examined closely for consistency of location and size of faces that are detected. Erratic size and or location can be sufficient to discard a face from consideration. Any shot which presents a single consistent face for the majority of the shot is labelled as a reporter.

The result of this further classification applied to the sound bite shots is given in Table 4. These results indicate that the detection of faces in these shots is still less than perfect, however, two thirds of the interview shots were detected. Given this level of recognition further classification can be performed as determined by shot syntax.

Further processing could be employed to specifically search for faces that are not perpendicular to the camera, which could add to the accuracy of this second step. In particular shots which are likely to be part of an interview segment, and which have no dominant face, could be tested for two faces. This would help detect the interviewer and interviewee shots, which would add further weight to the classification of such shots. This is intended as future work.

5 Results

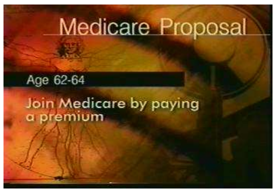
Figure 4 shows the thumbnails for the shots from one segment of detected structure. The caption for each thumbnail gives the visual similarity group computed using segmented colour histograms, the number of faces detected using the CMU face detection software [13], and the similarity group from aural clustering for the shot.

The topic of the segment is a report on the public view of the Medicare bill recently introduced in the USA. There is an anchor shot (Figure 4(a)), followed by a shot of only one sample which coincides with a fade (Figure 4(b)). This shot would be discarded from consideration. There is then a shot of explanatory text (Figure 4(c)), which is correctly identified as a voice over. The next shot (Figure 4(d)) is of Bill Clinton addressing a group of reporters, this is identified as a voice over due to incorrect vocal clustering. No face was detected due to the mobility of the speaker around the stage. Figure 4(e) shows another anchor shot, which is correctly identified. Figures 4(f) and 4(g) are of "people on the street", interviewed about their views on the topic. They are correctly identified as separate pieces of footage, and labelled as sound bites. In both cases the camera parameters are too irregular to expect face detection. The final figure, Figure 4(h) is the closing anchor shot, and is identified as such.

As can be seen, the clip of Bill Clinton (Figure 4(d)) is classified as a voice over, rather than a separate piece of footage. This is in part due to the brevity of the shot, and in part due to the noise and length of pause in the spoken voice. Improved audio processing would perhaps reduce this

(a) Visual group 116, Faces 1, Aural group 1.

(b) Visual group 117, Faces 1, No aural group.



(c) Visual group 118, Faces 0, Aural group 1.

(d) Visual group 119, Faces 0, Aural group 1.

(e) Visual group 116, Faces 1, Aural group 1.

(f) Visual group 120, Faces 0, Aural group 2.

(g) Visual group 121, Faces 0, Aural group 3.

(h) Visual group 116, Faces 1, Aural group 1.

Figure 4: Structure in an example news program.

	395	396	397	398	399
394	0.747	0.720	0.054	0.102	0.142
395		0.340	0.958	0.887	0.907
396			0.931	0.860	0.881
397				0.167	0.228
398					0.005

Table 5: Vocal (dis)similarity for shots in Figure 5.

difficulty. However, it must be assumed that many of the voices which occur in these shots will be unseen. While some people are regularly included in news bulletins (Bill Clinton as President), many others will be involved in news for only a brief period, corresponding to the time of a particular event and story. Moreover, the "people on the street" interviewed are intended to be random choices. This makes the task of separating such voices from each other more difficult. A further difficulty observed is that the anchor people will have numerous samples of their voice present, and any agglomerative classification method should associate these. The smaller groups of other voices, often with only a small number of samples, and the samples containing multiple voices, make it difficult to distinguish between outliers and separate samples.

An example where voice classification does work well is shown in Figure 5 and Table 5. This sequence of shots shows a male anchor person, Lou Waters, presenting a story on harassment, with two people interviewed (Figures 5(b) and 5(c)) and a commentary over a still (Figure 5(e)). Table 5 presents value of the distance measure used in audio similarity detection for the six shots. The values for the comparison of the two interviewees to the anchor person are clearly separable from those for the comparison of anchor person shots, with a range of $[0.72 - 0.958]$ compared to a range of $[0.005 - 0.228]$ for the similar shots. The voices of the two interviewees are quite similar, and could reasonable be clustered together, their dissimilarity value of 0.34 is classified by the system as similar.

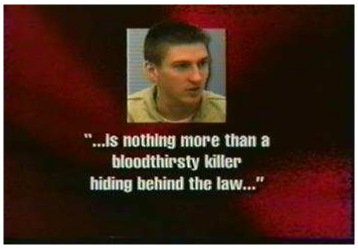
Figure 5 also provides a further example of the frame similarity algorithm, with the shots in Figure 5(f) containing an extra image, but still being found similar to the earlier anchor shots. In addition

(a) Shot 394 – Visual group 122, Faces 1, Aural group 1.

(b) Shot 395 – Visual group 335, Faces 1, Aural group 2.

(c) Shot 396 – Visual group 336, Faces 1, Aural group 3.

(d) Shot 397 – Visual group 122, Faces 1, Aural group 1.



(e) Shot 398 – Visual group 337, Faces 0, Aural group 1.

(f) Shot 399 – Visual group 122, Faces 1, Aural group 1.

Figure 5: Thumbnails of a news report with male anchor.

to this the two shots of interviewees, although visually quite similar are correctly separated. Figure 5(f) also gives an additional example of the type of head movement which is misclassified by the CCV algorithm.

6 Conclusions

The process employed in this work combines a number of image and aural low–level processes that, in isolation, are unreliable for classification of video. The fusion of the results of these processes, together with knowledge of the shot syntax for a particular domain, leads to a reliable and high level structure labeling of the video. While the resulting classification is less than perfect, all significant structure is recognized, albeit slightly over segmented.

The segmentation produced separates shots into homogeneous story segments, and is able to identify the shots which contain anchor people and reporters. The ability to extract the shots containing reporters and anchors is particularly important, as this provides a powerful key for access to the video content. This gives a suitable starting point from which a summary may be produced without hiding information from the user.

Further processing, such as the proposed refinement of face detection, would allow extraction of more detailed structure. Detection of interviewer and interviewee shots in interview segments would allow not only the presenting reporter to be identified visually as a key, but also the interviewee.

Further visual processing in the form of text detection and recognition is a possible future extension. Improvement to the audio processing is also an avenue for increasing the accuracy of the system, and perhaps allowing further information to be extracted. Given key words recognized from audio, and text recognized from video such as can be seen in Figure 4(c), further fusion of results may be useful for improving recognition of these stages.

The inclusion of shot syntax as a model for structure within news video is a major advantage for detection of shot type. This allows the extension of simple attribute based indexing to deduction of semantic structure within video, and the separation of video into segments of homogeneous semantic content. Extraction of semantic segments and deduction of shot type from a video stream greatly increases the utility of a video warehouse. Currently research is being undertaken to examine how well the shot syntax concept generalizes to other forms of video. Interview and news footage have a very regular shot syntax, but there are other forms of video with regular shot syntax which might be detected using similar techniques, or by application of additional measures. Research is also being undertaken to determine methods for the deduction of shot syntax structure from samples of a particular video form. Such a process could be of great value in multimedia and video data mining.

7 Acknowledgments

The authors would like to acknowledge the assistance of CNN for providing news footage that formed the base data set for this work.

References

[1] T. Blum, D. Keisler, J. Wheaton, and E. Wold. Audio databases with content–based retrieval. In M. Maybury, editor, *Intelligent Multimedia Information Retrieval*, chapter 6, pages 113–135. The MIT Press, 1997.

[2] R. Bolle, B.-L. Yeo, and M. M. Yeung. Video query and retrieval. In *Advanced Topics in Artificial Intelligence*, volume 1342 of *Lecture Notes in Artificial Intelligence*, pages 13–24. Springer, December 1997.

[3] A. Cheyer and L. Julia. MVIEWS: multimodal tools for the video analyst. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 55–62. ACM, January 1998.

[4] J. M. Corridoni, A. Del Bimbo, and P. Pala. Image retrieval by color semantics. *Multimedia Systems*, 7(3):175–183, May 1999.

[5] M. Davis. Media streams: An iconic visual language for video annotation. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 196–202. IEEE, April 1993.

[6] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, September 1995.

[7] H. Gish, M. Sui, and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *ICASSP–91*, pages 873–876. IEEE, IEEE, 1991.

[8] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Scene determination based on video and audio features. In *Proceedings IEEE Multimedia 99*, pages 685–690, Firenze, June 1999. IEEE.

[9] W. Y. Ma and B. S. Manjunath. NeTra: A toolbox for navigating large image databases. In *Proceedings of the International Conference on Image Processing*, pages 568–571, 1997.

[10] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura. Video handling with music and speech detection. *IEEE Multimedia*, 5(3):17–25, July 1998.

[11] G. Pass, R. Zabih, and J. Miller. Comparing images using colour coherence vectors. In *Proceedings ACM Multimedia 96*, pages 65–74, Boston, November 1996. ACM.

[12] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Signal Processing Series. Prentice Hall, 1978.

[13] H. A. Rowley, S. Baluja, and T. Kanade. Neural network–based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[14] C. Saraceno and R. Leonardi. Audio as a support to scene change detection and characterization of video sequences. In *Proceedings of ICASSP 97*, pages 2597–2600. IEEE, IEEE Computer Society Press, 1997.

[15] K. Shearer, S. Venkatesh, and C. Dorai. Attribute based discrimination of speaker gender. Technical Report 4, Curtin University of Technology, GPO Box U1987, Perth 6001, Western Australia, November 1999.

[16] D. M. Shotton, A. Rodriguez, N. Guil, and O. Trelles. Analysis and content–based querying of biological microscopy videos. In *Proceedings of the 15th International Conference on Pattern Recognition*. IAPR, IAPR, 2000.

[17] S. Srinivasan, D. Petkovic, and D. Ponceleon. Towards robust features for classifying audio in the CueVideo system. In *Proceedings of ACM Multimedia 99*, pages 393–400. ACM, ACM, 1999.

[18] Y. Tonomura and S. Abe. Content oriented visual interface using video icons for visual database systems. In *IEEE Workshop on Visual Languages*, pages 68–73. IEEE, 1989.

[19] Y. Tonomura, A. Akutsu, K. Otsuji, and T. Sadakata. VideoMAP and VideoSpaceIcon: Tools for anatomizing video content. In *INTERCHI 93 Conference Proceedings*, pages 131–138, 1993.

[20] S. Tsekeridou and I. Pitas. Audio–visual content analysis for content–based video indexing. In *IEEE International Conference on Multimedia Computing and Systems*, pages 667–672. IEEE, IEEE, 1999.

[21] B.-L. Yeo and M. M. Yeung. Classification, simplification and dynamic visualization of scene transition graphs for browsing. In *Storage and Retrieval for Image and Video Databases VI*, pages 60–70. SPIE, December 1998.

[22] M. Yeung, B.-L. Yeo, W. Wolf, and B. Liu. Video browsing using clustering and scene transitions on compressed sequences. In *Proceedings of the SPIE*, volume 2417, pages 399–413. SPIE, 1995.

[23] S. J. Young, M. G. Brown, J. T. Foote, G. J. F. Jones, and K. S. Jones. Acoustic indexing for multimedia retrieval and browsing. In *ICASSP 97*, volume 1, pages 199–202. IEEE, IEEE, 1997.