IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Artifacts of the colour coherence vector and an alternative similarity measure

Kim Shearer [a] Svetha Venkatesh [b]

IDIAP–RR 01-02

Febuary 2001

soumis à publication

Institut Dalle Molle d'Intelligence Artificielle Perceptive • CP 592 • Martigny • Valais • Suisse

téléphone +41 – 27 – 721 77 11
télécopieur +41 – 27 – 721 77 12
adr.él. secretariat@idiap.ch
internet http://www.idiap.ch

a IDIAP
b Curtin University

Rapport de recherche de l'IDIAP 01-02

Artifacts of the colour coherence vector and an alternative similarity measure

Kim Shearer Svetha Venkatesh

Febuary 2001

soumis à publication

Résumé. Image similarity measures can be used to capture useful structure in video processing. In this paper one popular variation, the colour coherence vector, is discussed. It is shown to perform poorly for certain tasks and a simpler, but more effective alternative is proposed. This alternative is examined for the initial task of anchor person spotting in news broadcasts, and extended to generic interview detection.

1 Introduction

There are numerous open sub–problems in the video classification field of study. A solution to the video classification problem requires not only information from various low level visual and audio processing routines, but also an intelligent fusion of the information obtained (Cheyer and Julia, 1998; Ma and Manjunath, 1997; Ponceleon *et al.*, 1998). This paper focuses on one sub problem of vision processing for video classification, that of association of video clips containing similar shots. The specific problem presented as an example is shots of anchor people in broadcast news video.

This problem occurs in the detection of narrative structure from video using shot syntax. For this task we are presented with a stream of video containing a number of shots, with some of the shots depicting dynamically changing scenes and others of a mostly static scene. There may be repetition of some static scenes, in the sense that more than one scene will be shot in the same location with the same camera parameters. The desired result is that the sampled frames from each static scene are recognised as belonging to a highly similar scene, and repetitions of a static scene are labelled as similar.

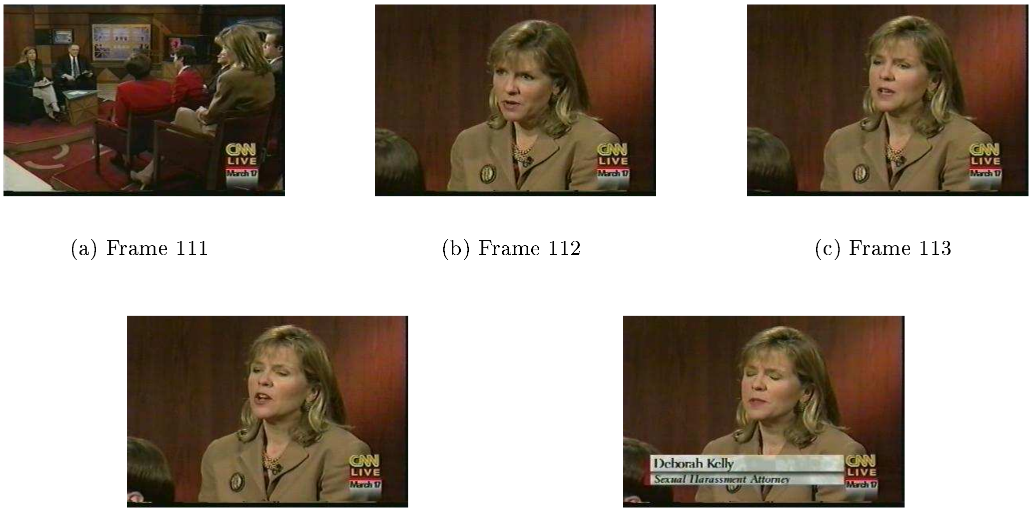
The specific application examined is separation of news video into semantic units. There are shots within general news video, be it post production or raw feed, of footage from events and perhaps shots of historic relevance. There will also be shots of anchor people, reporters and perhaps interviewers. Within the narrative structure of news video a segment will be introduced and presented by a reporter, who will be termed the anchor person. The shots of anchor people are constructed to be as visually consistent a possible, to give continuity to a news broadcast. Shots of interviewers and interviewees, are likewise generally consistent as far as background and camera parameters, that is similar zoom and angle, to avoid distracting the viewer from the content (Reisz and Millar, 1968). Given that the shots of individuals speaking, and possibly small groups conversing, will be kept as similar as possible, it should be possible to recognise these using fairly simple computer vision measures. Further processing can then be applied to repeated shots to determine more detailed annotation. For example, face detection can be applied to search for a face which dominates the frame, which would indicate a likely anchor shot. The face and voice of the people who appear in anchor sections of a news broadcast can then be used to index sections of video, representing candidates for the anchor, interviewer and interviewees. Further, face recognition from a database of know anchor people or reporters could be applied to add depth to the annotation where possible.

If it is possible to identify the anchor person shots, and perhaps various interviewer shots from the footage, then this information can be used to examine the shot syntax of the production, and further refine classification (Shearer *et al.*, 2000). This would allow a more detailed description, limiting the list of candidates for the anchor role, and possibly separating the interviewers from interviewees.

2 Methods for solution

Image processing literature abounds with attributes and measures for assessing image similarity. These measures vary from simple colour histograms and other colour measures (Corridoni *et al.*, 1999; Pass *et al.*, 1996; Lienhart *et al.*, 1999; Bolle *et al.*, 1997; Huang *et al.*, 1997), to measures of texture and moments (Pedersini *et al.*, 1996; Ma and Manjunath, 1997; Flickner *et al.*, 1995; Stricker and Dimai, 1996), and further more complicated measures (Idris and Panchanathan, 1997). Consideration of the problem statement indicates that the colour of images should be an important cue to recognising repetition of static shot settings. Colour should be highly consistent, and often distinctive, within each repeated studio shot in a news broadcast, and between such shots. Spatial distribution of colour will also be consistent across such shots, providing key information for shots which are otherwise similar in colour content.

The colour histogram is a simple and easily computed measure of image similarity, which has difficulty with spatially significant data. A simple example of this difficulty is a typical interview sequence. The sequence may begin with a full face shot of the interviewer in a studio. The next shot



(a) Frame 111 (b) Frame 112 (c) Frame 113

(d) Frame 114 (e) Frame 115

FIG. 1 – *Facial rotation for which CCV performs poorly*

may be a slightly wider shot of the interviewer and interviewee in the same studio. Although the two shots vary significantly in spatial distribution of colour, the colour histogram for a frame sampled from each of the shots would be similar. There may be a similar amount of facial colour in each shot, even though the colour is central and in one piece in the initial shot, but in two separate pieces, placed at either side of the image in the other.

Various means have been proposed to compensate for this shortcoming in simple colour histograms by incorporating spatial information into a histogram type measure. One measure that is commonly applied to video is the colour coherence vector (CCV) (Lienhart *et al.*, 1999; Pass *et al.*, 1996). The CCV includes spatial information in a histogram measure by taking into account the coherence of the regions of each colour. This is done by assigning each pixel in an image to a colour bin (usually 64 or 128 bins) as in the usual colour histogram approach, and then performing connected component analysis on the regions created by the discrete colours. Each connected component is classified as either coherent if it is larger in area than a predetermined threshold, or incoherent if it is smaller. The number of pixels in each colour bin that belong to coherent objects is summed, as is the number belonging to incoherent objects. Each colour bin is then represented by these two sums: the coherent pixel count and the incoherent pixel count. This measure is effective at differentiating between large (coherent) coloured objects and scattered smaller (incoherent) objects of a consistent colour. Thus an image of a field of daisies might be differentiated from an image of a large sunflower, although both may contain a similar amount of yellow and green. The spatial information incorporated into the CCV measure is non-specific with respect to the location within an image of colour regions.

The problem faced in analysis of news video is that spatial coherence as measured by a CCV can be greatly effected by slight rotations of objects in the video. In this application slight changes to lighting on the face of a person can greatly alter colour coherence due to the changes in reflection. Such changes can occur due to small movements of the face. An example of this can be seen in Figure 1 and Table 1. Here there are five frames extracted from a news broadcast at half second intervals. The first frame (Figure 1(a)) shows a wide shot of the studio, pictured are two anchor people and a discussion panel. The remaining four frames are of the same member of the panel addressing the discussion topic. Table 1 shows the similarity values returned from the CCV algorithm. A typical threshold for this implementation of the CCV algorithm is 10000, as can be seen from Table 1 this

	111	112	113	114
112	5886			
113	25759	25559		
114	7839	4681	25544	
115	4326	7570	25721	8303

TAB. 1 – *Similarity table for frames 111–115 using CCV*

	111	112	113	114
112	71112			
113	71410	5374		
114	70844	8220	5454	
115	71430	16644	14718	15260

TAB. 2 – *Similarity table for frames 111–115 using SCH*

gives poor performance. The essential requirement for this application is that the frames representing a repeated static shot be easily separable from frames representing other shots. In this case the frames of the panel member are a subset from a static shot, in that the camera is stationary on a single person. The initial frame in Figure 1(a) is clearly different, and should be classified separately. A classification based on the CCV measure as shown in Table 1 could not perform this separation as the similarity measure between frame 113 and other frames is far greater than the measure between frame 111 and other frames. This is due to the tilt around the horizontal axis of the head of the panel member. Reflections on the skin textures cause a significant alteration in the brightness of facial regions, causing a change in region sizes.

Simple histograms do not discriminate well in the application proposed in this paper, and the CCV approach for combining spatial data with colour data does not perform as required. An alternative approach is required which incorporates spatial data with colour data in a more suitable manner. The algorithm which is employed splits each frame or image into 12 equal regions, with four divisions along the horizontal axis and three along the vertical axis (Figure 2). A colour histogram is calculated for each of there regions, with each histogram having 16 bins. Frame similarity is then calculated as the sum of the histogram differences for each pair of corresponding regions between two frames. That is for two frames in a video stream f and g, there are 12 regions for each frame: $f_i : 0 \leq i < 12$ and $g_i : 0 \leq i < 12$. Given a simple histogram difference function δ, the frame similarity measure is

$$FD = \sum_{i=0}^{11} \delta(f_i, g_i) \tag{1}$$

The function δ is simple difference of magnitude summed over the 16 bands of the histogram, so for

FIG. 2 – *Regions used for local histogram comparison*

two histograms h and k

$$\delta = \sum_{j=0}^{15} \mid h_j - k_j \mid \tag{2}$$

This method is used to add information on placement of colour within the frame or image. Each of the 12 regions within the frame or image must match the colour of a similarly placed region in the frame with which it is compared. This allows minor local variations within the image at small cost, but not large local variations that may not alter a global colour histogram.

The initial test of this algorithm uses only simple intensity histograms, with intensity being the sum of red, green and blue colour components. Table 2 shows the similarity measure values obtained for the frames in Figure 1. The frames of the panel member are clearly separable from the initial frame, with a typical threshold being 30 000, with frame difference measure of less than the threshold being similar. This threshold is less than half the least difference between a frame of the panel member and the initial frame, yet almost double the greatest difference between two frames of the panel member. In addition to this, the three comparisons with largest magnitude correspond to the difference between the frame of Figure 1(e) and each of the other panel member frames. This frame contains a text overlay not present in the other frames, so the larger difference value is a correct response. For the two 50 minute videos used as test data, all anchor person and reporter shots were located, with no false negatives. The grouping of shots by this similarity measure separated individual reporter from each other, and also located a small number of repeated shots other than anchor and reporter shots. The other shots are generally logos for a station or program, or previews for an interview style program. Both of these are separable from correct anchor and reporter shots using further processing (Shearer *et al.*, 2000).

The approach of splitting each frame into a number of regions before processing was attempted with the CCV algorithm as well, with poor results. If the frames are split into smaller regions, there is a difficulty with the size threshold for coherence using CCVs. If the size threshold is left at the value used for a full sized frame, then some colour regions will be split on a region boundary. This leads to a greatly changed CCV, with a great many artifacts from the region boundaries, and the measure does not perform well. An alternative is to reduce the coherence threshold, so that fewer artifact are formed from the region boundaries, however this alters the coherence measure. This approach was tried with a wide range of thresholds and also various numbers of regions, but no reasonable performance could be achieved.

Initial experiments with news footage show excellent groupings using segmented colour histograms. This grouping along with domain knowledge and face detection allows reliable semantic grouping of the shots in a news broadcast Shearer *et al.* (2000). In fact for the application of anchor person detection the histograms use only greyscale intensity information, yet still easily separate interview and anchor shots.

2.1 Extensions to another domain

This method has been further applied to detection of interview shots within the context of sport coverage, as part of the EC project ASSAVID. In this project the goal is to provide rich annotation for unknown sports video. The video will contain all forms of video, from programs as broadcast (PasB), to raw feeds, such as those from an isolated camera feed. One initial goal is to separate video showing actual sporting events taking place from other video, such as interviews and introductory segments. In this context segmented histograms as used for new footage performs poorly, providing similar discrimination to the CCV algorithm. This is partly due to the less constrained nature of the footage, with a number of interviews being conducted on site, and therefore having a similar colour profile to the actual event. In order to improve performance to a level suitable for this application two enhancements were made to the histogram method. The first, and obvious, enhancement was to extend the histograms from greyscale to colour. For this purpose RBG colour is used in a simple three dimensional colour space, with four bins per axis. While this slows the algorithm slightly, it still performs classification in approximately one tenth real time. The second enhancement is to aid in separating shots that are similar. In the initial version, a sampled frame f_s is considered similar to a previous shot $s = f_i \dots f_j$ if the similarity measure $\delta(f_s,f_n)$, where $i \leq n \leq j$, is less than a threshold. This tends to permit a large spread of shots across similar frames. The value of the average histogram $h_{av} = average(h(f_i),\dots,h(f_j))$, is permitted to drift considerably under this method. In order to control this under circumstances where there are highly similar shots that we wish to distinguish, a new similarity measure is introduced such that once a new category of frames is introduced, it is represented by the average of all frames classified as part of that category. The similarity measure used is the comparison of the frame f_s against the average of all frames in a category, thus $s = \delta(f_s,average(h(f_i),\dots,h(f_j)))$. If the new frame is found to be similar, the average histogram is updated so long as the set of frames is contiguous. This restricts the spread of the category.

This second enhancement has been applied to both the segmented histograms and the CCV algorithm. The CCV algorithm gives little improvement from this change, showing mainly a shifting of the shot grouping, rather than an increased clarity in classification.

The segmented colour histograms however, show a dramatic improvement in performance. The results of their application to a segment of sporting video is shown in the latter part of the results section.

3 Results

The five frames in Figure 1 are part of the video shown in Figure 3, which presents one key frame for each shot of the segment of the video stream. Figure 1(a) is towards the end of the shot represented by 3(e), and figures 1(b) to 1(e) are near the beginning of the shot represented by 3(f). The classification produced by the region histograms matches the actual shots exactly, and produces the labels given in the captions. As can be seen, figures 3(a), 3(d) and 3(p) are classified as similar, giving the anchor shots for the report as a whole. Figures 3(i), 3(k) and 3(o) are also classified as similar, representing an anchor for a sub-report. In fact figures 3(e) and 3(g) are also the same anchor couple, however, the view differs sufficiently that separate grouping is reasonable. Shots 3(e) and 3(g) are found to be similar to each other, as hoped. The other similar shots, figures 3(h) and 3(j) are also grouped together. This classification is entirely automatic, and produces results as expected from manual inspection.

If colour coherence vectors (CCV) are employed to classify this video segment, the results are far less suitable. The initial shot of Greta Van Susteren is broken into three sections, classified as sections 1, 2 and 1. As the central section is only one frame in length, the error is simply corrected by absorbing this frame into the whole. The second and third shots are correctly recognised as different from the initial clip and each other, and similar within their extent. The fourth clip is a shot of Greta Van Susteren in the studio, similar to the initial shot. All but one frame of this shot is correctly recognised, and once again the incorrect frame could be absorbed by a simple algorithm. However, the following shot, which is a wide shot of the studio, is also absorbed into the previous shot of Greta Van Susteren. This is clearly incorrect from a semantic view point, although examination of the CCV values revels that the colouring and the coherence within colours is very similar between the frames in these shots. While both of these shots are of anchor people, there is a clear difference between the shots that is not detected using the CCV algorithm. The following shot displays a further problem with CCVs, in that a shot that seems visually highly consistent is not classified as such. The object in the shot, the face of Deborah Kelly, rotates slightly between each frame. This causes large changes in the coherence of the images, as the size and connectivity of areas of specific colours changes rapidly. As a result of these changes the clip is broken into a sequence of alternate classifications, with short blocks (2 frames to N frames) classified as similar to each other but differing from the rest, interspersed with single frames classified as similar to the initial shot. This is not an error from which recovery can be reliably performed.The most significant problem is that the remainder of the clip is classified into one piece, which is determined to be similar to the initial clip.

(a) Shot 84 – Visual group 80, Faces 1.

(b) Shot 85 – Visual group 81, Faces 1.

(c) Shot 86 – Visual group 82, Faces 1.

(d) Shot 87 – Visual group 80, Faces 1.

(e) Shot 88 – Visual group 83, Faces 0.

(f) Shot 89 – Visual group 84, Faces 1.

(g) Shot 90 – Visual group 83, Faces 0.

(h) Shot 91 – Visual group 85, Faces 1.

(i) Shot 92 – Visual group 86, Faces 0.

(j) Shot 93 – Visual group 85, Faces 1.

(k) Shot 94 – Visual group 86, Faces 0.

(l) Shot 95 – Visual group 87, Faces 0.

(m) Shot 96 – Visual group 88, Faces 0.

(n) Shot 97 – Visual group 89, Faces 1.

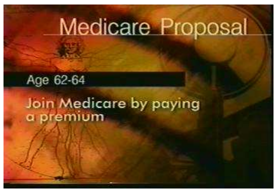
(o) Shot 98 – Visual group 86, Faces 0.

(p) Shot 99 – Visual group 80, Faces 1.

FIG. 3 – *Classification of clip1 by region histogram*

(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.

FIG. 4 – *Classification of clip2 by region histogram*

Person	Frequency
One	3
Two	3
Three	3
Four	2

TAB. 3 – *Frequency of interview sections*

Clip 2 (Figure 4) shows similar performance for the two methods. Each correctly identifies the three sections of anchor person (Natalie Allen) as similar, and each of these sections is classified as a single piece. The difference in performance is that the CCV method classifies many frames of the three men interviewed (Bill Clinton and two anonymous citizens) as similar, which is undesirable, while the SSH method classifies two parts of the second anonymous interview as similar, but no other.

The final sequence shown as results is more difficult to classify, as it comes from the ASSAVID project domain of sports footage. It consists of various interviews and raw segments of a pre-broadcast version of Olympic swimming coverage. The interviews presented are mostly filmed at the poolside, meaning that they have similar background and colour composition. Neither method classifies the coverage completely correctly, but there is a significant difference in performance.

There are 11 separate interview shots in the complete swimming coverage, which are of four personalities. Their frequency is given in table 3. Each repeated interview of the same person is shot in the same location, and each person has a separate location. It is desirable to separate the interviews into classes containing only one person per class. The CCV algorithm has all the interview shots classified into only two groups. The first of these groups contains all three shots of interview three, both of interview four shots, one of the interview two shots and one of the interview one shots. It also contains five other shots, an example frame from each of these shots is given in Figure 6. In addition, one of the interview four shots is merged with one of the interview one shots. The second group contains the remaining two interview one shots, and two of the interview two shots. Once again, an interview one shots is merged, this time with an interview two shot, in addition, the other interview two shot in this group is split into four separate pieces.

On contrast the SSH algorithm performs significantly better. The three interview three shots are correctly grouped together, and are the only shots in their group. Similarly, the interview four shots are placed together as the only two shots in their group. The interview one shots are split into two groups, one containing only the first two shots , and another group containing the final interview one shot and another shot containing a face in a highly similar layout. The interview two shots are also split into two groups, one group contains the first two interview two shots and a further facial shot, while the final shot is in a group alone.

4 Summary and Conclusions

The colour coherence vector (CCV) is a frequently used measure for image similarity, especially in the field of video annotation and indexing. In this paper we have shown that for a particular application within this field, that of interview location, the CCV method is ill suited. This is due to the particular properties of human faces, which have contours that cause large variations in colour coherence for small changes in image appearance. This causes not only miss classification of shots, but frequently also causes fragmentation of shots which present a dominant face. In the particular application of structure detection in news and interview programs this is a major defect.

There is an alternative measure presented in this paper that provides far superior performance for the similarity comparison of images which contain faces. The presented measure not only provides a more useful estimate of image similarity, but is also far simpler and faster to calculate than the CCV measure.

This result suggests that it is necessary to examine such measures in some detail before they

(a) Interview 1.

(b) Interview 2.

(c) Interview 1.

(d) Interview 2.

(e) Interview 4.

(f) Interview 3.

(g) Interview 3.

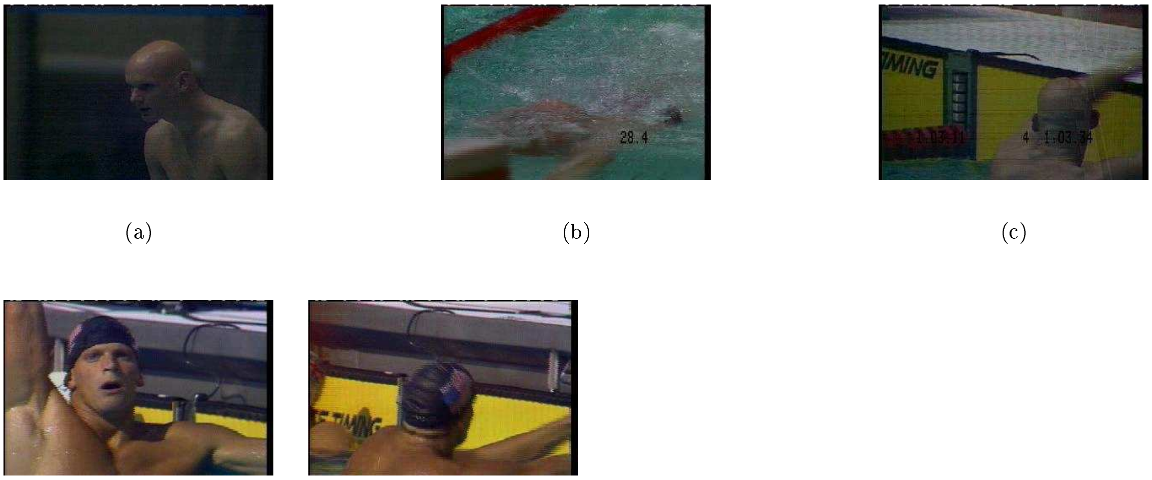
(h) Interview 3.

(i) Interview 4.

(j) Interview 1.

(k) Interview 2.

FIG. 5 – *Olympic swimming examples*



(a) (b) (c)

(d) (e)

FIG. 6 – *Other shots grouped with interviews by the CCV measure*

are adopted for specific applications. This is particularly true for methods that involve a complex computational component. The vast variety of information presented in the multimedia domain makes a general solution to any problem a rare thing indeed. It seems that as with other areas, each measure should be examined for applicability to the proposed area before general adoption.

Références

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

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

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

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

Huang, J., Kumar, S. R., Mitra, M., Zhu, W. J., and Zabih, R. (1997). Image indexing using color correlograms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 762–768.

Idris, F. and Panchanathan, S. (1997). Review of image and video indexing techniques. *Journal of Visual Communication and Image Representation*, **8**, 146–166.

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

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

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

Pedersini, F., Sarti, A., and Tubaro, S. (1996). Combined motion and edge analysis for a layer–based representation of image sequences. In *Proceedings of the IEEE International Conference on Image Processing*, volume I, pages 921–924. IEEE.

Ponceleon, D., Srinivasan, S., Amir, A., Petkovic, D., and Diklic, D. (1998). Key to effective video retrieval: Effective cataloging and browsing. In *Proceedings of the 6th ACM International Conference on Multimedia*, pages 99–107. ACM.

Reisz, K. and Millar, G. (1968). *The Technique of Film Editing*. The Focal Press, second enlarged edition.

Shearer, K., Venkatesh, S., and Dorai, C. (2000). Automatic detection of voice over and sound bite footage in news videos. Technical Report RC 21716 (97799), IBM, IBM T J Watson Research Center, PO Box 218, Yorktown Heights, NY 10598.

Stricker, M. and Dimai, A. (1996). Color indexing with weak spatial constraints. In *Proceedings of the SPIE*, volume 2670, pages 29–40.